UNIVERSITY
OF TWENTE.

European Research Council
Established by the European Commission

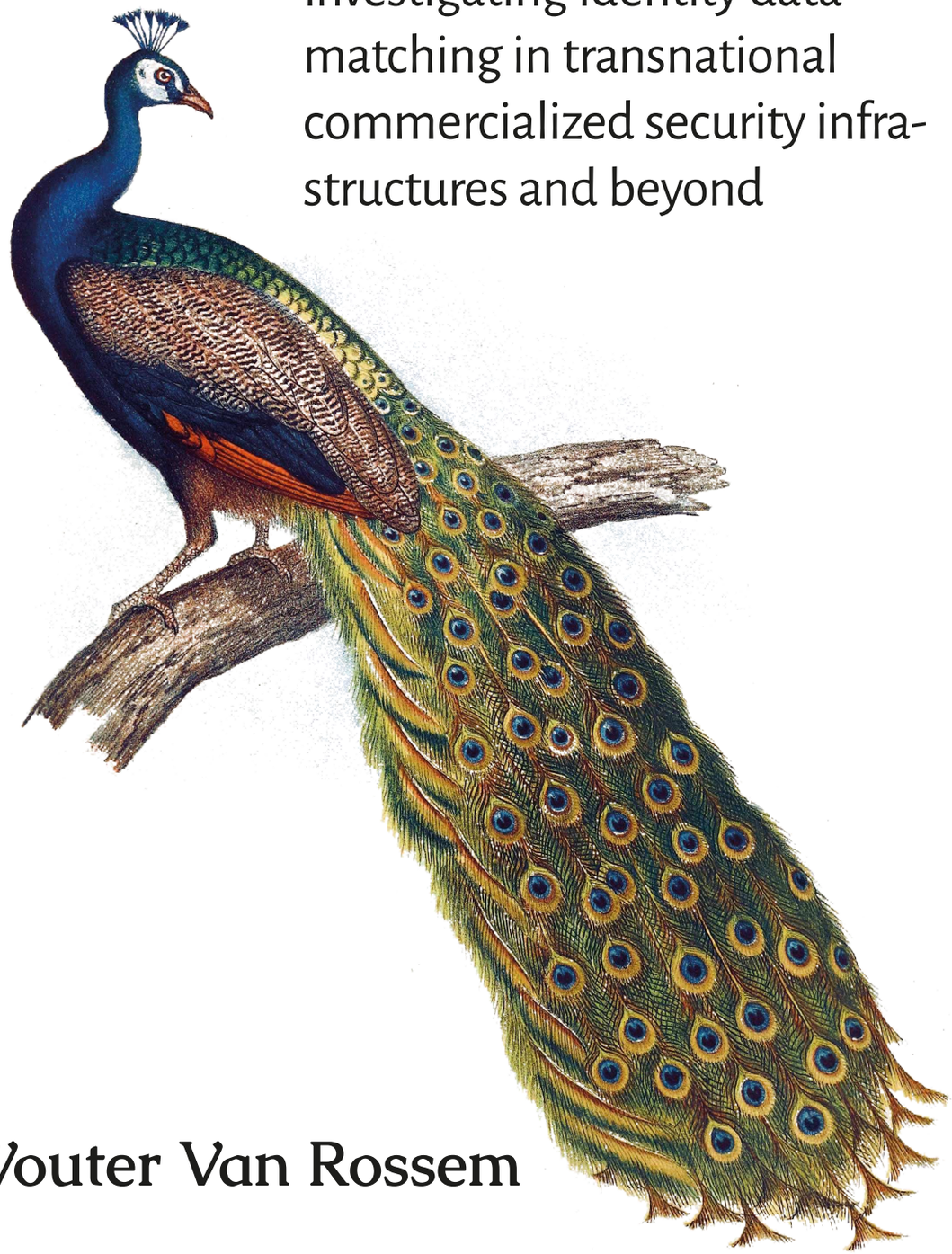
9 789036 561761 >

Wouter Van Rossem

Finding Blind Spots

FINDING BLIND SPOTS

Investigating identity data matching in transnational commercialized security infrastructures and beyond



Wouter Van Rossem

FINDING BLIND SPOTS

INVESTIGATING IDENTITY DATA MATCHING IN TRANSNATIONAL COMMERCIALIZED SECURITY INFRASTRUCTURES AND BEYOND

Wouter Rudi Van Rossem

FINDING BLIND SPOTS

INVESTIGATING IDENTITY DATA MATCHING IN TRANSNATIONAL COMMERCIALIZED SECURITY INFRASTRUCTURES AND BEYOND

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. ir. A. Veldkamp
on account of the decision of the Doctorate Board,
to be publicly defended
on Monday, July 15th, 2024, at 12:45 p.m.

by

Wouter Rudi Van Rossem

born on the 7th of August, 1989
in Brussels, Belgium

This dissertation has been approved by:

Promotor:	prof. dr. S. Kuhlmann
Co-promotors:	prof. dr. A. Pelizza
	prof. dr. ir. M. van Keulen

The investigations in this dissertation were conducted within the context of the "Processing Citizenship: Digital registration of migrants as co-production of citizens, territory and Europe" project, which has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement No 714463.

This dissertation was typeset using (R) Markdown, LaTeX and the `bookdown` R-package. The text is set in Alegreya, a font designed by Juan Pablo del Peral for Huerta Tipográfica.

The cover design was created by Wouter Van Rossem, featuring Figure 224 "Der gemeine Pfau (Pave cristatus)" from Fitzinger's "Bilder-atlas zur Wissenschaftlich-populären Naturgeschichte der Vögel in ihren sämmtlichen Hauptformen" (1864), sourced from https://www.biodiversitylibrary.org/page/33050550."

ISBN (print): 978-90-365-6176-1
ISBN (digital): 978-90-365-6177-8
DOI: 10.3990/1.9789036561778

Graduation Committee

Chair / Secretary:	prof. dr. T. Bondarouk
Promotor:	prof. dr. S. Kuhlmann *University of Twente, BMS, Knowledge, Transformation & Society*
Co-promotors:	prof. dr. A. Pelizza *University of Bologna, Department of Philosophy, and Aarhus University, Department of Digital Design and Information Studies*
	prof. dr. ir. M. van Keulen *University of Twente, EEMCS, Datamanagement & Biometrics*
Committee Members:	prof. dr. M. de Goede *University of Amsterdam, Faculty of Humanities*
	dr. K.E. Konrad *University of Twente, BMS, Knowledge, Transformation & Society*
	prof. dr. A. Rotolo *University of Bologna, Department of Legal Studies*
	prof. dr. R. Torenvlied *University of Twente, BMS, Public Administration*
	prof. dr. R. Williams *University of Edinburgh, Institute for the Study of Science, Technology and Innovation (ISSTI)*

In memory of my mother, Linda Liekendael (1960–2020)

Contents

Acknowledgments

Writing this dissertation has been a demanding yet fulfilling journey, intertwined with various professional and personal experiences. I therefore want to express my immense gratitude to my supervisors, peers, institutions, and loved ones for their support and encouragement, which played a crucial role in the completion of my dissertation.

The unwavering support and mentoring of Annalisa Pelizza has been invaluable in guiding me through the complexities of this dissertation. Thank you for your encouragement and guidance, which helped propel me and the research in the right direction. Your close engagement with the work, along with your always insightful feedback, has had a profound impact on both the quality of the research and my growth as a scholar and writer. I am grateful to Stefan Kuhlmann for his insightful oversight and holistic perspective on my project. Your seasoned guidance, drawing from years of supervising other students, combined with a knowledge of relevant academic fields, has provided fresh perspectives and helped keep the project on track. I am grateful to Maurice van Keulen for his positive energy and sustained interest in the project even as it expanded into other academic fields. It was valuable to have an extra set of eyes to double-check the technical details in the research and provide fresh perspectives on presenting the research and its findings. Thank you as well to the members of the graduation committee for their involvement and engagement with my work during these final stages of the process. While having a team of three supervisors with different backgrounds and perspectives presented some complexities, I am confident that the committee members and readers acknowledge the positive impact of their diverse influences on the dissertation.

The research would not have been possible without the funding from the European Research Council (ERC) which funded the Processing Citizenship project under the European Union's Horizon 2020 research and innovation program (grant agreement No 714463). Annalisa Pelizza deserves thanks for leading the development of this project and bringing together an exceptional team of researchers. I want to give a special shout-out to Chiara Loschi, Lorenzo Olivieri, Paul Trauttmansdorff, and Claudio Coletta for their incredible collaboration and the countless enjoyable moments in and outside of the office, and thought-provoking discussions about the project and much more. Additionally, I want to acknowledge the valuable contributions from other project members at different times, including Andrea Pettrachin, Yoren Lausberg, Chiara Andreoli, Annalisa

Bacchi, Stephan Scheel, Ermioni Frezouli.

I extend my heartfelt thanks to the WCC Group for their collaboration and assistance in facilitating my research project. I wish to express my heartfelt thanks to Roelof Troost for his exceptional support and confidence in my research on their ELISE data matching system. I want to express my gratitude to the identity and security team members, along with other individuals at WCC, for generously sharing their insights and expertise in this particular field with me. I would like to offer my heartfelt thanks to the employees of the Netherlands' Immigration and Naturalization Service (IND) for their kind cooperation in generously sharing their valuable work experiences, which were essential for understanding the relationship between data matching and applicant identification. Thank you to Maarten van Heiningen from WCC for his valuable UX expertise and support with the IND interviews.

I want to thank UTwente STePS/KiTeS colleagues for their support, insightful reflections, and encouragement during our coffee breaks, lunchtime discussions, and presentations: Ewert Aukes, Binod Koirala, Mario Pinzon-Camargo, Arie Rip, Klaasjan Visscher, Peter Stegmaier, Adri Albert de la Bruhèze, Kornelia Konrad, Lissa Roberts. I also want to thank my fellow researchers at the Department of Philosophy and Communication Studies during my stay at the University of Bologna, who offered a pleasant respite in our shared room at the department, which was an especially welcome change after working remotely during the COVID-19 pandemic. I appreciated the moments to take breaks from work, whether it was during lunch, an apperitivo, or *prendere un caffè*: Antonella, Chiara, Eduardo, Filomena, Gioacchino, John, Monica, Matteo, Nicola.

I want to thank Elize Schiweck and Tatiana Mattioli for their help and attentiveness in handling administrative duties and personal support at the University of Twente and University of Bologna respectively. More broadly, I also want to extend my heartfelt thanks to all those involved in the work that forms the backbone of our academic pursuits. This includes the dedicated administrative staff and all the support personnel whose behind-the-scenes efforts ensure the smooth functioning of our academic community.

When it came to facilities, I had a wonderful time exploring the high-tech and green campus of the University of Twente. The experience of living and working on a campus that offers excellent facilities, along with the opportunity to take daily walks in the forest and explore the unexpected beauty of Overijssel through cycling and camping, was a true blessing, especially during the difficult times of the Covid pandemic. Equally inspiring was the time spent in the historic city and university of Bologna, with its gorgeous red buildings, porticoes, and beautiful surrounding hills. My heartfelt thanks to everyone at the University of Bologna and the University of Twente who made it possible for me to spend time at these institutions.

I am happy about the opportunities that the Processing Citizenship program and I had to contribute to larger academic networks. Through these collaborations, we were able to ignite new ideas and establish connections that made our work more meaning-

ful and enjoyable. I feel privileged to have played a role in the STS-MIGTEC network, which has emerged as a significant network for scholars investigating the connection between STS and critical migration, security, and border studies. I want to acknowledge and thank Nina Amelung, Silvan Pollozek, Olga Usachova, Andrea Berger, Aristotle Tympas, Vasilis Argyriou, and all the others who have gone above and beyond to make this network possible. I would also like to express my sincere gratitude to Georgios Glouftsios, Rocco Bellanova, Matthias Leese, Fran Meissner, Nina Dewi Toft Djanegara, Ana Valdivia, and many others who have been involved in this and other academic networks, for their valuable collaborations, contributions, and insightful discussions regarding my work.

Bologna proved to be an excellent environment for us to grow and expand our networks. Here, I am also thankful for having had the privilege of being a part of the organizing committee for the 2023 conference of the Italian Association for Social Studies of Science and Technology in Bologna. Juggling my PhD completion and organizing this event was tough, but the conference turned out to be truly amazing. I wish to extend my heartfelt thanks to Paolo Giardullo, Barbara Saracino, Agnese Cretella, and Simone Arnaldi.

Starting my PhD journey with a software development background, I had to adjust to a very different academic field. However, the WTMC graduate school program provided me with an exceptional education in Science and Technology Studies, and I feel blessed to have been a part of it. I would like to extend my heartfelt appreciation to the exceptional WTMC training coordinators for their immense effort in planning the amazing workshops and schools: Anne Beaulieu, Bernike Pasveer, Andreas Weber. I want to express my gratitude to my fellow WTMCers for the wonderfully fun, reflective, and creative moments we've had together in Soeterbeeck and beyond.

My heartfelt appreciation goes out to my friends and old acquaintances, both near and far, for their support and understanding throughout this lengthy dissertation journey. Despite the challenges and geographical distances, I hold dear the moments we've shared in Bologna, Brussels, Enschede, Maynal, and, of course, even online during the pandemic. This thanks extends to the awesome communities who have made life more enjoyable over the years: the Enschede Swing Out Loud swing dancers, the Haraway book club crew, my fellow Korean language learners in Bologna at the Centro Studi d'Arte Estremo-Orientale, the P-NUT community, and so many others. Special appreciation goes to Gaetan, Ludovic, Isabel, Iza, Guy, Lloyd, Mathijs, Elisa, Kaat, Bart, Kristina, Carla, Fenna, Jef, Iris, Ming, Menno, Ciro and many others.

Graag wil ik ook mijn familie bedanken. Papa, bedankt voor je steun tijdens dit hele proces en voor altijd een thuisbasis te bieden waar we naar terug kunnen keren. Jeroen en Kelly, hartelijk dank; het is altijd leuk om gezellig samen tijd door te brengen, bordspelletjes te spelen, en het is geweldig om te zien hoe jullie twee ketjes, Yoshi en Yuna, zijn opgegroeid gedurende mijn hele proefschrift. Ook bedankt aan Michel, Linda, Rudi, Merce, &co voor de gezellige momenten samen. Het doet me veel verdriet dat mijn mama

er niet meer is om dit moment mee te maken, maar je hebt me enorm gesteund en ik weet zeker dat je enorm trots zou zijn geweest op mij.

Ontelbare dank aan mijn wederhelft Sujin. Bedankt voor je steun tijdens moeilijke periodes en voor je voortdurende geloof in mij en mijn werk. Ons verhaal is onlosmakelijk verwoven met dit proefschrift, en ik koester alle momenten en plekken die we in deze tijd samen hebben mogen beleven en ontdekken. Ik kijk uit naar de vele mooie momenten die ons nog te wachten staan. 한국에 계신 수진의 가족들께 감사의 말씀을 전합니다. 논문을 준비했던 모든 순간 저를 응원해주시고 함께해주셔서 감사합니다.

Chapter 1

Introduction: Understanding identity data matching in transnational security infrastructures

> The message is that there are no "knowns." There are things we know that we know. There are known unknowns. That is to say there are things that we now know we don't know. But there are also unknown unknowns. There are things we don't know we don't know. So when we do the best we can and we pull all this information together, and we then say well that's basically what we see as the situation, that is really only the known knowns and the known unknowns. And each year, we discover a few more of those unknown unknowns. (Rumsfeld, 2002)

1.1 Information gaps and blind spots in identification

Former US Secretary of Defense Donald Rumsfeld made this observation in a 2002 press conference, which has since captivated academic and lay audiences alike. In his observation, he distinguished between what he called "known knowns" (i.e., facts that authorities are confident they know) and "known unknowns" (i.e., facts that authorities are aware they do not yet know). However, he also pointed out that there are "unknown unknowns" or peculiar blind spots that authorities don't know about and don't even realize they don't know. In the context of this thesis, Rumsfeld's idea of "known unknowns" and "unknown unknowns" offers an insightful lens through which to introduce the challenges of identifying people in border security and migration control. Identifying and tracking individuals across national borders is rarely straightforward, as data are only sometimes complete or readily available.[1] This raises the question of how authorities can effectively identify individuals despite incomplete data sets, aliases, or even false identities, as well as

[1] The word data is often treated as a mass noun, and hence, something that cannot be counted or divided (e.g., "the data is available"). In contrast, this dissertation uses data in its countable plural noun form ("data are"). I follow the convention of using this form to highlight that data are multiple and "arise from and are used

how authorities can acknowledge and address the incompleteness of information about something they are not aware of.

Rumsfeld's concept of "known unknowns" can be applied to situations where an individual's data are present in a database but is not directly linked to their identity data in other systems. These "known unknowns" or "blind spots" can hinder the ability of authorities to identify people fully because of legal, organizational, or technical challenges. One example of a "known unknown" in identifying people at borders could be an individual with multiple identity data in different databases or systems that have not been linked. For instance, international watch lists contain information on individuals suspected or known to be involved in criminal activities. Individuals on these lists are often listed with multiple known aliases to address the challenge of linking all their identity data together to establish that they may be the same person. In other words, authorities already recognize these individuals, but there remain uncertainties regarding their identification. As such, the ambiguity of personal identity data creates a "known unknown" regarding individuals' identities, which can have implications for security and law enforcement purposes.

"Unknown unknowns" can apply to information that is not only unknown but also unidentifiable through traditional means. Technology can play a role in detecting—or should we say enacting?—such blind spots by enabling the analysis and correlation of large amounts of identity data. For instance, advanced algorithms and machine learning techniques can detect previously unknown connections and patterns in the data. In their book "Algorithmic Reason," Aradau and Blanke (2022) offer an intriguing case in which two journalists were potentially flagged as persons of interest by a United States security agency using algorithms that detected anomalies in regular data patterns. In the previous case of "known unknowns," individuals may be placed on a watchlist, and it is known that these persons may use different names or identity documents. Conversely, "unknown unknowns" involve entirely unknown connections and patterns that have yet to be discovered. By connecting various identities and other data and finding patterns, it becomes possible to identify or re-identify someone who is not yet known or of interest to authorities, thus potentially uncovering "unknown unknowns."

The problem of identifying and connecting identity data is not new and has been a challenge for various domains of knowledge. However, in recent decades, data collection, storage, and analysis have been significantly impacted by processes of datafication, resulting in vast amounts of personal information being processed (e.g., Borgman, 2015; Kitchin, 2014). Hence, one critical development in this context has been the growth of *data matching technology* to identify individuals across multiple sources (e.g., Christen, 2012; Harron et al., 2017; Talburt, 2013). By utilizing data matching technology, individuals can be identified even when the information is incomplete or inconsistent, thanks to the comparison and reconciliation of data, such as name, address, and identification numbers. As a result, data matching tools are widely deployed in fields where up-to-date

in varied circumstances worth acknowledging" (Loukissas, 2019, p. 13).

information is vital, such as healthcare, finance, and law enforcement (Talburt, 2013).

In the healthcare sector, data matching technology is used to match patient data across different systems to ensure accurate patient identification and prevent medical errors (e.g., Lee et al., 2016; McCoy et al., 2013; Sauleau et al., 2005). By utilizing patient information like names, birthdates, and social security numbers, medical records can be effectively matched and organized across diverse healthcare systems and databases (Zech et al., 2016). In financial intelligence, data matching is used, among others, to detect fraudulent activities and for sanctions compliance. For instance, SWIFT, a worldwide provider of secure financial messaging services, employs data matching algorithms to aid financial institutions in complying with sanctions regulations by accurately identifying individuals on sanction lists who may use aliases or fraudulent identities to avoid detection (SWIFT, 2018, 2021). In law enforcement, data matching technology is used to analyze data and aid investigations, such as identifying individuals involved in organized crime networks by linking their biographical information across databases (Ferguson, 2017; Steinbock, 2005/2006). For example, data matching techniques are employed to analyze flight passenger data to identify patterns and potential threats by linking and analyzing individuals' travel histories across different flights and airlines (Bellanova and Duez, 2012; Hobbing, 2010).

The previous examples underscore how data matching technology has become crucial in linking and reconciling personal data across multiple sources, given the increasing collection of information, such as electronic medical records, financial transactions, and online purchases. The technology enables organizations to create more comprehensive profiles of individuals, contributing to detecting fraudulent activities, ensuring regulatory compliance, and dealing with the siloed nature of data sources. Nevertheless, despite the increasing use of data matching technology in various sectors, there is still a lack of understanding of how it shapes the meaning of the things it connects, including identifying data as suspicious and shaping relations between organizations whose data are being matched and connected. This research seeks to contribute to a more performative understanding of the role of data matching technology by investigating how it shapes the meaning of data, practices, and the organizations that use it. As a result, it is necessary to begin by recalling the history and applications of matching and linking data.

1.2 Connecting the dots: The development of data matching techniques

The use of computing technology to connect personal identity data has a long past that dates to the early days of punch card technology, even predating database technology, as evident from research such as Dunn (1946)'s "Record linkage" and Newcombe et al. (1959)'s "Automatic linkage of vital records." The term "record linkage" is often used in public health, epidemiology, and demography to describe the practice of matching and linking records pertaining to the same individual across multiple data sets. In the fields

of public health ((e.g., Jutte et al., 2011) and demographics (e.g., Abbott et al., 2015), for example, linking data proved beneficial to states seeking to improve services for their citizens and facilitate research. Through the use of identifiers and shared attributes such as name, address, date of birth, or social security number, states could establish a more detailed profile of individuals by connecting data records (Newcombe and Kennedy, 1962). Difficulties arose when attempting to link personal data because of data quality issues, such as inconsistencies in name spellings or missing information, prompting the development of new technologies and techniques to tackle these challenges.

The emergence of electronic computers and database technology enabled more sophisticated matching algorithms to be developed, leading to increased adoption in other fields (Batini and Scannapieco, 2016; Christen, 2012). As a result, the process of matching data sets and linking records is now referred to by various names, such as data matching, data linking, data merging, data integration, record linkage, deduplication, or entity resolution, depending on the context and application (Christen, 2012). This dissertation will use the term *data matching* as it is a more general term referring to identifying records in data sets that refer to the same real-world persons (or other entities) and reconciling duplicates or inconsistencies between data sets.

Another concept related to data matching is *schema matching* (Bellahsene et al., 2011; Kementsietsidis, 2009). Schema matching addresses the challenge of integrating data from multiple sources that have different schema structures or data models. Schema matching's importance stems from its capacity to facilitate data integration across disparate data sources, which is frequently required for effective data matching. For example, identifying records referring to the same person may be challenging without knowledge of the underlying data models, as the same person may be represented differently across different data sets. Therefore, in fields where data are fragmented and dispersed across multiple sources, data matching and schema matching are crucial components of successful data management and integration. While the term "schema matching" will not be used in the dissertation, the question of how to investigate correspondences and differences between different data models (i.e., schemas) that underpin the data will be explored in greater depth.

Over the years, various data matching methods and techniques for classifying matches have been devised (Batini and Scannapieco, 2016; Christen, 2012; Fellegi and Sunter, 1969; Winkler, 2014). The following standard data matching methods can be distinguished based on the literature. One of the most basic techniques for identifying matching records is *deterministic matching*, which employs predefined rules or criteria. For example, when two records have the same first name, last name, and date of birth, they are considered a match. Another approach is *probabilistic matching*, which uses statistical algorithms to calculate the probability that two records are a match based on the similarity of their categories of data. If, as in the previous example, two data records have similar but not identical names or dates of birth, the records may still be considered a match based on the probability calculation. Another approach is *rule-based*

matching, which can combine deterministic and probabilistic methods to find matches and incorporate expert knowledge or domain-specific rules to increase accuracy. Finally, matching techniques based on *machine learning* are gaining popularity. Such methods employ algorithms that can learn from data to improve accuracy and more easily adapt to new data sources.

While data matching may seem like a technical process, its increasing use and impact on society and individuals mean that it has significant consequences that should not be overlooked. With the growth of the internet and the digitalization of many aspects of contemporary life, data matching has become even more ubiquitous, with many actors using these techniques to link data from different sources and gain insights into individuals, their behavior, and preferences (Clarke, 1994; Gandy, 1989; Zuboff, 2015). Furthermore, using data matching algorithms in automated systems can introduce errors and biases, making some people disproportionately the target of surveillance and control (e.g., Aradau and Blanke, 2021; Benjamin, 2019; Eubanks, 2018). For instance, the German-Lebanese citizen Khalid al-Masri was imprisoned and tortured by the CIA in 2003 after being mistakenly identified as a suspected terrorist with a similar name (Priest, 2005). Data matching technologies play a crucial role in these processes by allowing for the analysis and correlation of vast amounts of data and determining previously unknown connections in identity data.

Data matching has a long history of addressing the challenges posed by fragmented, incomplete, and duplicated information across multiple sources, with the development of various techniques. However, data matching is not just a technical process that can potentially discover previously unknown connections, but can alter the things being connected. These connections can affect the meaning of the data, practices, and the organizations that use it. For instance, one could argue that by matching flight passenger data to terrorist watch lists, the identification of a match alters the original meaning of the passenger data, and changes the role of organizations such as airline carriers (see also, Amoore and de Goede, 2005; Bellanova and Duez, 2012). Over time, passenger data has evolved from simple travel information to a powerful tool that connects data, allowing for the identification of suspicious travel patterns and the detection of individuals who may be considered security risks.

As such, it is crucial to understand how data matching technology shapes the meaning of data and practices, including identifying data as suspicious and shaping relationships between organizations whose data are being matched and connected. This research seeks to contribute to a more performative understanding of the role of data matching technology by investigating how it shapes the meaning of data, practices, and organizations. The choice of exploring data matching in border security and migration control is linked to the overarching *Processing Citizenship* (PC) project, which aims to understand how data infrastructures for processing migrants and refugees co-produce individuals and Europe (PC, 2017/2023).[2] The purpose of investigating the use of matching and link-

[2] The Processing Citizenship project, including this PhD research, was funded by the European Research

ing data in the context of identity data in border security and migration management in this dissertation is thus closely connected to the PC project's aim of exploring how the production, evaluation and circulation of data about third-country nationals are reshaping European governance (see also, Pelizza, 2019; Pelizza and Loschi, 2023). Specifically, this research aims to examine the use of matching and linking data in the context of identity data in border security and migration management. The following section takes a closer look at how data matching is used in this context by exploring a contemporary example of data matching in migration and border control within the European Union.

1.3 Leveraging data matching for border control

> In light of the recent terrorist attacks in Europe and the increase in irregular migration in recent years, action needs to be taken to address this risk of information gaps and blind spots. The measures in this proposal [Interoperability of EU information systems for security, border and migration management] will ensure the various systems can exchange data and share information so that authorized bodies and officers have the information they need to strengthen our borders and better protect Europe. (European Commission, 2017)

> Establishing a common repository of data would overcome the current fragmentation in the EU's architecture of data management for border control and security. This fragmentation is contrary to the data minimization principle, as it results in the same data being stored several times. Where necessary, the common repository would allow for the recognition of connections and provide an overall picture by combining individual data elements stored in different information systems. It would thus address the current knowledge gaps and shed light on blind spots for border guards and police officers. (European Commission, 2016b, p. 18)

> [One of] the four technical components of the proposal [is] a multiple identity detector — this will verify whether the biographical data that is being searched exists in multiple systems, helping to detect multiple identities. It has the dual purpose of ensuring the correct identification of bona fide persons and combating identity fraud. (European Commission, 2017)

Council in the context of the European Framework Program for Research and Innovation Horizon 2020, grant agreement No 714463, principal investigator Annalisa Pelizza.

These quotes reveal how in the European Union (EU) context data matching is regarded as a critical component in addressing identity issues in migration and border control systems, including identifying multiple identities. The quotes above refer to a project linking identity data of different EU information systems for security, border, and migration management.[3] Presently, each of these EU information systems operates independently of its database and serves a distinct purpose, such as managing asylum requests, processing visa applications, or supporting law enforcement activities. The proposal explicitly identifies a potential risk of information gaps and blind spots because data are not connected. It proposes to address this risk by connecting and sharing information from those multiple systems. Furthermore, the European Commission (EC) communication underscores the importance of having the necessary information to strengthen borders and identify potential threats.

The second quote describes the need for "establishing a common repository of data," which would "address the current knowledge gaps and shed light on blind spots for border guards and police officers" (p. 18). Finally, the third quote describes a component for finding multiple identities that refer to the same person. Fragmentation of the EU's data management architecture for border control and security is thus portrayed as causing duplicate data storage and leaving border guards and police officers with knowledge gaps and blind spots. According to this logic, the common repository and multiple identity detector components would enable the recognition of connections and provide a holistic view by combining data elements stored in different information systems.

The EU interoperability initiative introduces new components that emphasize the growing significance of data matching technologies. By allowing the matching of biometric data, visa data, and other identity-related information, the initiative aims to enhance the accuracy and efficiency of EU information systems for mobility and border control. However, the use of these new components is not just limited to improving the functioning of these systems. They will also be pivotal in implementing new, interlinked forms of identification of individuals deemed suspicious based on the links between data sets (Quintel, 2018) based on probabilistic, rule-based, or machine learning-based data matching. Understanding the performative nature of data matching technology is essential, as it shapes the perceptions and treatment of individuals in the context of border security and migration management.

Note that the European Commission is not building these systems by itself; it increasingly relies on global information technology suppliers and integrators (Lemberg-Pedersen et al., 2020; Valdivia et al., 2022). A research gap exists in understanding how data matching technologies, which are increasingly developed by commercial entities for global use (see, for example, Leese, 2018; Lemberg-Pedersen et al., 2020; Valdivia

[3] The interoperability initiative follows recommendations from the "High Level Expert Group on Information Systems and Interoperability" (European Commission, 2016a) and a new legislative mandate (European Union, 2018) for the European Agency for the operational management of large-scale IT systems in the area of freedom, security and justice (eu-LISA). It was split into two proposals due to different legal provisions for regulating (a) borders and visas and (b) police and judicial cooperation, asylum, and migration (European Union, 2019a,b).

et al., 2022; Zureik and Hindle, 2004), operate and influence processes of identification in a sensitive domain. The international and commercial dimensions of identification technology mean that there is a growing need to examine how the private sector is involved in developing and implementing these standardizing technologies. As Pollock and Williams (2009) have noted, criticism of standardized software and focus on how poorly such software adapts to different settings is not a sufficient research perspective. The widespread use of standardized identification software requires comprehending how such software is produced and adapted to operate in various contexts, as well.

1.4 Unpacking the challenge of analyzing data matching in transnational data infrastructures

The previous examples show how identity data matching has developed into an integral part of border and migration control systems to enable the identification and tracking of individuals across different systems and jurisdictions. As such, matching data from national and international sources indicates that identification practices extend beyond the borders of nation-states, highlighting the internationalization of identification. Yet, research on identification has typically concentrated on how states identify people. Data matching technology, however, is illustrative of the global dimensions of identification, which become apparent only when looking beyond the borders of individual nations.

A transnational perspective can thus shift the focus away from the nation-state and onto the various other actors involved in identification. The shift from state authorities creating and implementing identification technology to the state purchasing systems created by commercial organizations can be seen in various programmes involving global information technology companies. For example, the United States Automated Biometric Identification System relied on Cogent/Thales' automated fingerprint identification technology (Thales Group, 2021), while India's Aadhaar biometric ID system utilized Accenture and Daon's technology for combining different biometric modalities (Accenture, 2010). Meanwhile, the upcoming European Entry/Exit System will utilize IDEMIA's biometric matching systems (Accenture, 2012). In this way, identification technology is increasingly becoming a commercial product rather than a creation of the state. Yet, little is known about how these actors developed identity data matching systems and put them to work. This lack of knowledge can be attributed to various factors, including the lack of transparency in developing these systems, limited access to information on their design and operation, and the complexity of the underlying technical and trans-organizational processes.

Data matching technology for identification can also be seen as a component of a broader *data infrastructure* (e.g., Flyverbom and Murray, 2018; Kitchin, 2014).[4] Infrastruc-

[4] Earlier studies have employed alternative terminologies such as "information infrastructure" (e.g., Bowker et al., 2009; Hanseth et al., 1996) or "e-infrastructure" (e.g., Pollock and Williams, 2010) to refer to these assemblages of technological and social components that enable the flow and management of data across different

tures are those things we depend on to make other things work (Edwards et al., 2009; Star and Ruhleder, 1996). Hence, data infrastructure includes the technologies, protocols, regulations, habits, procedures, and agreements to handle and utilize data. In the case of data matching, the technology is an essential component of the infrastructure that enables the sharing, linking, and matching of identity data across various systems and organizations. The development and maintenance of this infrastructure require collaboration and coordination among different actors, including government agencies, private companies, and international organizations. Understanding data matching as part of data infrastructure provides a more comprehensive view of the interconnected systems that enable local, national, and international identification practices.

Considering these challenges, the following section will outline this study's research problem, aims, and objectives, which seek to unpack the complexities of analyzing data matching in transnational data infrastructure.

1.4.1 Research problem, aims, and objectives

This dissertation aims to contribute to a more performative understanding of the role of data matching technology by investigating how it shapes the meaning of data, practices, and organizations in transnational contexts, particularly in the securitization of the European border. Recognizing a gap in research regarding the performative effects of data matching technology in transnational security infrastructures, this study aims to empirically investigate its involvement in infrastructure development, security, and internationalization within the realm of identification. To address this research problem and achieve the overarching research aims, the dissertation outlines specific research objectives:

- To map the theoretical landscape related to internationalization, securitization, and infrastructuring of identification and derive the dissertation's research question for investigating data matching in transnational data infrastructures (Chapter 2).
- To develop a methodological framework for analyzing data matching in transnational infrastructures using methodological strategies to uncover the embedded and less obvious technical details of matching and linking identity data in infrastructures (Chapter 3).
- To introduce a new method and software tool for analyzing the schemas that underpin information systems in population management (Chapter 4).
- To examine the relationship between identity data matching technologies and routine identification practices (Chapter 5).
- To investigate the long-term development of identification systems and building of transnational data infrastructures by identifying contingent moments in their evolution to explore how data matching expertise travels and circulates (Chapter 6).

settings and contexts.

The research aim to map the theoretical landscape aims to help understand how the meanings, practices, and technologies of identification have changed as it has become more international, commercial, linked to security issues, and part of broader infrastructures. The methodological framework and methodological strategy aim to provide an overarching framework for analyzing data matching in transnational infrastructures. The objective of introducing a new method and software tool to analyze the schemas underpinning information systems in population management is to examine the expectations and imaginaries of the schemas underpinning information systems. The objective of examining routine identification practices is to understand the relationship with identity data matching technologies, shedding light on how these technologies shape the utilization and meanings of data. Finally, the objective of investigating the long-term development of identification systems and the building of transnational data infrastructures is to explore how data matching expertise and technologies travel and circulate. The following section provides an overview of the dissertation's structure, outlining the chapters and their focus.

1.5 Structure of the dissertation

The organization of this dissertation involves an initial mapping of theoretical concepts, followed by the development of a methodological framework to direct the analysis, and follows with empirical chapters. Chapter 2 starts by mapping theoretical concepts on identification and matching identity data, drawing on literature related to the internationalization and commercialization of identification, the securitization of identification, and the infrastructuring of identification. This chapter lays the groundwork for the subsequent chapters by discussing various theoretical perspectives on matching identity data and the implications of matching identity data for transnational data infrastructures.

Chapter 3 introduces the methodological framework for analyzing data matching in transnational infrastructures. The framework proposes three methodological strategies, wherein data matching serves as both a research topic and a methodological resource. These three strategies are based on comparing data models, analyzing data practices and tracing sociotechnical change. Comparing data models can reveal information collected by various organizations and systems; data practices can show the searching and matching of identity data within and across organizations; sociotechnical change can shed light on the circulation of data matching knowledge, technologies, and practices over time and across organizations. The chapter explains how these strategies were used in the dissertation's fieldwork at a software company developing data matching technology. The chapter also describes the methods of data collection and the techniques of data analysis used in the dissertation.

Chapter 4 introduces the "Ontology Explorer" (OE) methodology, a semantic approach and an open-source tool to analyze the data models' underlying information

systems. The method draws inspiration from schema matching and is designed to compare data models in different formats used by various systems. This chapter explains how it is applied in the dissertation to reveal less visible assumptions and patterns in information systems design. Unlike other methods, the OE allows for the systematic comparison of non-homogeneous data formats and enables comparisons of data models across information systems run by diverse organizations and authorities. Therefore, the OE makes it possible to observe how identity data properties influence the production and circulation of data and the relations between different authorities' data models.

Chapter 5 examines the relationship between technologies for searching and matching identity data and routine bureaucratic identification practices in migration management. The chapter focuses on how a government migration agency searches and matches applicants' data using a data matching system. The chapter introduces the concept of "re-identification" to refer to the process by which subjects of bureaucratic procedures are re-identified in data infrastructures at various points in those procedures. The chapter demonstrates the implications of data matching in bureaucratic settings in two ways. First, the chapter shifts the usual focus from first registration to re-identification practices across data infrastructures. Secondly, the findings underscore that, while integrating data matching tools for re-identification alleviates data friction, it inadvertently also comes with certain costs.

Chapter 6 looks at the long-term development of identification systems and infrastructures. The chapter proposes two heuristics for detecting contingent moments in the evolution of identification technologies. First, it demonstrates how a data matching system's changing "interpretative flexibility" allows discerning actors' varying problematizations of identification, such as those related to the securitization of identification. Second, the chapter demonstrates how "gateway moments" make it possible to see the compromises necessary when building identification infrastructures and adapting globally honed technologies to new settings. Together, the chapter's findings shed light on the activities of under-the-radar actors, such as commercial software vendors, whose distribution and reuse of systems have long-term implications for identification practices and infrastructures in various contexts.

The dissertation concludes in Chapter 7 with a summary of empirical findings, literature contributions, and reflections on the research process. Contributions include mapping the theoretical landscape of identity data matching, introducing a methodological framework for analyzing data matching in transnational infrastructures, proposing new methods for analyzing data matching and using these to examine the relationship between data matching technologies and bureaucratic practices. The study makes an additional contribution by delving into the long-term evolution of identification systems and infrastructures. Finally, the chapter acknowledges the study's limitations and suggests areas for future research. The dissertation aims to advance our understanding of identity data matching by putting it into the STS and critical data studies agendas. It contends that matching identity data is a multifaceted phenomenon that requires a nuanced and

interdisciplinary approach to understand how it shapes and is shaped by transnational data infrastructures.

Chapter 2

Mapping the theoretical landscape: Unraveling the complexity of matching identity data in transnational security infrastructures

This chapter presents a mapping of theoretical concepts based on literature from four broadly defined areas of research, which will be referred to as (1) "the internationalization of identification," (2) "the commercialization of identification," (3) "the securitization of identification," and (4) "the infrastructuring of identification." By exploring how identification operates in international, commercial, securitized, and infrastructural contexts, we can develop an understanding of the challenges, implications, and potential consequences associated with data matching practices and technologies. This mapping exercise thus lays the groundwork for generating hypotheses and formulating research questions that relate to the role of matching identity data in transnational commercialized security infrastructures.

The theoretical mapping approach provides an overview of core theories and concepts that provide a framework for the research. While this mapping provides a compass to navigate the subsequent chapters, each empirical chapter will delve into supplementary literature to address the specific research questions pertaining to that chapter's focus. Nonetheless, the theories and concepts introduced in this chapter will play an integral role throughout the dissertation, albeit in varying degrees of explicitness. They provide a theoretical underpinning and serve as a thread that connects the various strands of inquiry throughout the study.

To begin, let us clarify what the theoretical mapping in this chapter covers and how it works alongside the methodological framework from Chapter 3. First, theoretical mapping is used in this dissertation to integrate theories and concepts from multiple research fields and topics more or less directly pertinent to identification and identity data matching. Second, theoretical mapping plays a role in identifying gaps in the existing literature and formulating the research questions, as it allows for raising doubts, questions, and

hypotheses informed by literature. Chapter 3 then adds to this theoretical mapping by introducing the methodological framework, a written and visual representation of the specific research, hypotheses, and methods of data collection and analysis that will be used to answer the research questions. Now, let us delve into the first aspect of identification, which is the increasing internationalization, referring to identification practices that extend beyond the confines of the nation-state.

2.1 Internationalization of identification

Identifying people on the move is frequently described as a Janus-faced and multi-sited state-led phenomenon entwined with obligations, rights, and coercive measures that can even occur far from traditional border encounters (e.g., About et al., 2013a; Caplan and Torpey, 2001). The accompanying development of registration and identification systems, such as the creation of civil registers or passports, has, indeed, typically been linked to the state-making of modern nation-states (Breckenridge and Szreter, 2012; Caplan and Torpey, 2001; Torpey, 2018). Because identification consequently enables states to track people's movements and label some as irregular, registration and identification systems and procedures have been argued to facilitate and limit people's freedom of movement across international borders (Broeders and Dijstelbloem, 2016; Squire, 2010; Torpey, 2018).

A term often referred to for the state's capacity to identify its citizens is the notion of "legibility" by Scott (1998). Scott noted how the interaction of premodern states and their population (e.g., for purposes of taxation) went hand in hand with projects of standardization and legibility as attempts to identify its people unambiguously. So, in one example, Scott noted the standardization of diverse cultural naming practices serving specific local purposes in distinct and fixed surnames as "a first and crucial step toward making individual citizens officially legible" (p. 71). Thus, Scott argued that state initiatives like rationalizing surnames enabled states to develop the modern statecraft apparatus, which includes taxpayer lists, conscription lists, censuses, and property deeds, by allowing states to identify the vast majority of their inhabitants unambiguously. Civil registries, identity cards, passports, DNA profiles, and other biometric databases have supplanted fixed surnames. Still, they all serve the same purpose: to make people legible for interactions with states.

Research on the history of identification generally focuses on these state practices of identifying people. For example, Torpey (2018), author of the book "The invention of the passport" originally published in 2000, uses Max Weber's aphorism on the legitimacy of the state's use of violence to argue that identification practices gave nation-states a "monopoly of the legitimate means of movement" (p. 2) at the expense of non-state actors like churches and private organizations. Torpey argued that states needed to issue identification documents to maintain the separation between nationals and foreigners and regulate their mobility.

Extending the scope of the discussion beyond the borders of the nation-state, similar registration and identification techniques have been shown to help colonial authorities gain and maintain power, albeit to varying degrees of success (Breckenridge and Szreter, 2012). From this perspective, the categories of registration and identification systems institutionalized the relationships between the empire's center and periphery. However, as About et al. (2013b) points out, registration and identification systems in such imperial and colonial contexts should not be viewed solely as instruments of coercion. What is interesting for our discussion here is how the historical analyses in those edited volumes from Breckenridge and Szreter (2012) and About et al. (2013a) show how identification technologies were brought into a colonial setting, adapted to local conditions, and eventually circulated back and forth between the global power centers.

An illustration of such international exchanges of identification technology is provided by Sengoopta (2004), who described how significant advancements in fingerprinting resulted from interactions between colonizers and colonized in the British Empire in India. Interestingly, Sengoopta compares the history of fingerprinting to the spread of curry dishes in English cuisine: "Developed in India but not indigenous, British but not evolved in Britain itself … incorporated into British tradition and then gradually retransmitted to the world at large, blur[ring] the simplistic distinction we often make between home and Empire" (p. 6, quoted in Cole, 2005). While an overview of the history of forensic fingerprinting is beyond the scope of this chapter,[1] the main takeaway is that fingerprinting as an identification method has a complicated history, starting with scientists on the European continent, moving through British India's colonial provinces, and finally returning to Victorian England's Scotland Yard. Therefore, the circulation of identification theories, techniques, and practices reveals the international aspects of identification technology, which can be better understood when considering perspectives beyond the conventional nation-state-based approach.

The international nature of identification technologies is especially visible in international policing. International police work goes beyond state practices because of the need to "identify, recognize, and track suspects or offenders across an expanding national and eventually international terrain," as the editors of a volume titled "Documenting individual identity: the development of state practices in the modern world" (Caplan and Torpey, 2001, p.9) write. The essays in the book discuss nineteenth-century transnational innovations in identification, such as Bertillon's anthropometric identification system, which became widely used in France and other countries, and the Vucetich fingerprint identification system, which spread from Argentina to other Latin American countries. Based on such examples, the editors note in the book's introduction how policing has been "the source of repeated efforts to rationalize and standardize practices of identification and the systems for the storage and retrieval of the expanding documentation that this generated" (p. 9). International policing, with its emphasis on identifying and tracking suspects across borders, has contributed to the rationalization and standardization

[1] For a comprehensive overview of the history of fingerprinting in law enforcement, see Cole (2001).

of identification practices and technologies.

However, despite the increasing prevalence of identification systems that link identity data from various national and international databases, not much research has been conducted on the operational deployment and utilization of such systems. This scarce research stands in contrast to the amount of real-world examples where such systems have been implemented. Some literature briefly touches upon this point and provides a few examples. For instance, such interconnected database systems are essential for identity verification in the context of Schengen visa applications (e.g., Glouftsios, 2019) and for detecting potential threats by linking identities across national and international policing and watch listing systems (e.g., de Goede and Sullivan, 2016; Zureik and Salter, 2005). However, these volumes do not dedicate much attention to other sources of data crucial to the internationalization of identification: the development of international standards for identification (Leese, 2018; Torpey, 2018). Further investigation and analysis are thus necessary to understand the operational implementation and implications of the methods, techniques, and tools used to link identity data across databases.

Many chapters in About et al. (2013a) still prioritize state-led standardization projects despite its focus on transnational identification. This is surprising because current efforts to rationalize and standardize identification systems rely heavily on the work of organizations like the European Commission, the United Nations, Interpol, the International Civil Aviation Organization (ICAO), and the International Organization for Standardization (ISO). For example, the ICAO publishes guidelines for machine-readable passports and visas (ISO/IEC, 2008) to ensure "global interoperability." The American National Standards Institute/National Institute of Standards and Technology Identification Technology Laboratory (ANSI/NIST-ITL) is an example of a government agency that publishes de facto standards for exchanging biometric information such as fingerprints and facial images. In short, rather than a fully coherent system for making people on the move legible, there is more likely a fragmented landscape of systems and data models — and where state vision often needs "fixing" (Leese, 2022) to identify people across international systems.

In this fragmented landscape of systems and data models, data matching emerges as a technological fix to address identification challenges. For example, when individuals move across international borders, their identity data may exist in disparate databases and formats maintained by different authorities and organizations. Data matching technologies can be employed to link these fragmented datasets, enabling a more comprehensive picture of an individual's identity across multiple systems. However, when we connect the internationalization of identification with data matching, it raises several hypotheses, questions, and doubts. How does the use of data matching techniques and technologies facilitate the international exchange of identity data? Are there challenges in harmonizing data matching processes and standards across different authorities and organizations? Does the internationalization of identification require the development of new data matching approaches to account for cultural, linguistic, and contextual vari-

ations in identity data? How does the increasing reliance on data matching shape identification outcomes in transnational security contexts? These questions emphasize the central goal of the dissertation, which is to investigate how data matching technology shapes the meaning of data, practices, and organizations in a global context.

This section has discussed the internationalization of identification, which refers to research into identification practices that extend beyond the boundaries of the nation-state. The need to identify people across borders has brought attention to the importance of identification technology and the efforts to standardize it. The internationalization of identification thus raises research questions regarding the types of knowledge and assumptions about people-on-the-move inscribed in data models of national and transnational security infrastructures. As we will see, it is necessary to develop methods to investigate the interconnections between diverse organizations' and authorities' data models. Furthermore, the diversity of these data models raises the question of how organizations that collect information about people-on-the-move search and match for identity data in their systems and how data are matched and linked across different agencies and organizations. The next section will examine how identification technologies travel across various authorities, states, and organizations and the involvement of different actors, including commercial entities, in their development.

2.2 Commercialization of identification

About et al. (2013a) thus also highlight the need for new research into the "increasing takeover of individual identification by corporate bodies for economic purposes" (p. 2). A transnational viewpoint can concomitantly shift the focus away from the nation-state and onto the various non-governmental actors involved in identification. Indeed, as noted by Higgs (2013) in one of the essays in the volume, "the contemporary state's use of identification technologies, including biometrics, is often dependent upon products created by commercial vendors, who have helped to drive their adoption" (p. 165). Overall, it is interesting to see a growing body of literature acknowledge the importance of non-state and commercial actors in identification processes and technologies (see also Pelizza, 2021). This is a topic that has, nevertheless, been explored in a disjointed fashion.

Even though "identification technology is increasingly a commercial product, rather than a creation of the state," as Higgs (2013) put it succinctly (p. 165), social and historical research has not followed suit by giving equal weight to these commercial actors in identification. Higgs observes that historians tend to place more emphasis on state identification developments than on, for example, commercial enterprises, where access to historical records is more restricted. Higgs' essay is able to paint a broad picture of the commercial aspect of identification by looking at technological advancements made by private companies in Britain at the beginning of the 18th century. His examples include banks' use of credit and debit cards to identify customers, supermarkets' use of loyalty

cards to create consumer profiles, and credit reference agencies' determination of a person's creditworthiness. Perhaps unsurprisingly, he concludes his article with the well-documented fact that an increasing number of commercial IT companies actively promote biometric identity systems ((see, for example, Amoore, 2006; Olwig et al., 2019; Trauttmansdorff, 2022; Zureik and Hindle, 2004).

The work of Higgs and other authors shows a gradual shift away from state authorities creating and implementing identification technology to the state purchasing systems created by commercial organizations. In the EU context, for example, the political economy of identity management has indeed been shown to consist of only a small network of companies that design, implement, and maintain identity management systems (Lemberg-Pedersen et al., 2020; Valdivia et al., 2022). These small networks have various risks that complicate the relations between standardizing and rationalizing identification as the "involved companies become indispensable, and are granted roles as unrivaled experts in the systems of border control they themselves have designed." (Lemberg-Pedersen et al., 2020, p. 71). Pelizza (2021) provides another example of how technology facilitating identification encounters in Greece forges transnational associations with the EU Commission, corporate contractors, and the US security regime through the production of hardware that adheres to US biometric standards, while also leading to vendor lock-in due to the use of proprietary algorithms to create biometric templates. While the significance of commercial identification systems is clear, little is known about the processes involved in their procurement, development, implementation, and deployment, and their longer-term implications for data sovereignty.

Beyond that, little is known about the consequences of such international and commercial identification technology on the processes used by states to identify people on the move. A compelling element of Higgs (2013)'s line of reasoning is how he shows not only the impact that commercial innovations have had on the widespread uptake of digital identification technology. Notably, he argues that the uptake and replication of these identification technologies and practices shape how states identify their populace. Therefore, this interdependence highlights the need for research on the evolution and diffusion of identification expertise and technology within and between various public and private institutions.

Linking the commercialization of identification to data matching technology further raises a range of hypotheses, questions, and doubts. Firstly, how does the commercialization of identification impact the development and implementation of data matching technologies used in identification systems? Specifically, how do commercial entities acquire and develop their data matching knowledge and expertise? One possible hypothesis is that commercial data matching systems may play a role in standardizing data matching and identification among different organizations by their underlying mechanisms. Yet, what are the effects of deploying these systems on the standardization of identification processes? Moreover, what are the potential risks and challenges associated with the involvement of commercial entities in the design, development, and operation of data

matching in identification systems? Could the re-use of data matching technologies by certain vendors or entities become a concern, leading to potential restrictions on competition and innovation in the field? Furthermore, how do companies determine which data matching functionalities to prioritize based on customer needs and industry trends? Further research is needed to explore such relationships between data matching and the commercialization of identification, providing deeper insights into their dynamics, implications, and potential consequences on data matching and identification practices.

In summary, investigating the commercialization of identification provides insights into the proliferation of commercially and internationally available identification systems, which introduces new challenges and complexities. This commercialization is characterized by a shift from state-driven development and implementation of identification technology to state adoption of systems developed by commercial organizations. By delving into the expanded role of commercial actors, we can uncover the creation of transnational associations and the potential for vendor lock-in, shedding light on the intricate relationships and power dynamics that underlie the commercialization of identification. Moreover, the commercialization of identification raises considerations regarding the development, implementation, and use of data matching technologies. It prompts us to question how knowledge and technology for matching identity data circulate and travel across organizations. Understanding this circulation would provide insights into the implications of the commercialization of identification. And yet identification is not only commercialized, but is entangled with established security regimes. The following section thus delves into the securitization of identification to explore how security concerns may both influence and be influenced by the design and utilization of identification systems.

2.3 Securitization of identification

The identification of people at the border has become increasingly linked by public and private actors to issues of terrorism and international security, as noted in the preceding chapter's examples from the ongoing EU interoperability project. According to (Critical) Security Studies (CSS) scholars, such construction of identification as a security concern supplants traditional understandings of security and should not be taken for granted. Although security has long been associated with state protection (national security), recent scholarship in CSS has examined the phenomenon's growing influence in new contexts, such as migration, health, and food. As a result, scholars have called into question "the precise definition of what it means to be secure, the causes of insecurity, and who or what the concept of security should apply to" (Peoples and Vaughan-Williams, 2021, p. 2). In this way, CSS has reformulated the security issue with what is commonly referred to as "broadening" and "deepening" the security-related areas, problems, and actors (Burgess, 2010; Peoples and Vaughan-Williams, 2021).

On the one hand, security can be characterized as having been "broadened" from its

traditional emphasis on more militaristic dimensions to include other areas and threats, such as migration and border security, or health (as, for example, the Covid-19 pandemic has shown). On the other hand, what is to be secured (also known as the *referent object* of security) can be characterized as having been "deepened" to include other actors such as human individuals, cultural/identity groups, multinational corporations, institutions, or even ecosystems. Scholars in this field, for example, argue that the securitization of migration in Europe has resulted in policies that criminalize and stigmatize migrants while ignoring the root causes of displacement (e.g., Huysmans, 2000). Similarly, we can see how identification has come to be securitized by looking at the techniques and procedures used for identification.

A crucial question for security scholars is how a specific societal phenomenon, like human migration, health, or the environment, is framed as a threat or security concern. The existence of security threats that need to be dealt with by security authorities and experts may be taken for granted in "traditional" or "essentialist" understandings of security.[2] What if, however, the appearance of security threats operates in reverse? How do societies label specific issues as security threats, and what societal and political consequences does this have? This alternative view is generally how constructivist theories of security share a concern for security as a meaning-making (intersubjective) process in which the interaction and communication between actors construct meanings of the world and shape actors' identities and power relations (Balzacq, 2009; Peoples and Vaughan-Williams, 2021). This view becomes particularly relevant in the context of identification and data matching, where the determination of what constitutes a security threat and the practices surrounding it can have significant implications. How are specific issues related to identification and data matching labeled as security threats, and what societal and political consequences do they have?

Constructivist approaches to security can provide insights into the securitization of identification and data matching, by investigating the framing of identification and data matching as security concerns as a social and discursive process, shaped by the interactions and communication between actors.[3] For instance, "securitization theory" strongly emphasizes how language, discourse, and power all play a part in how security is constructed and the value of critical analysis in challenging the prevailing security narratives and fostering new perspectives on security (Balzacq, 2005; Buzan et al., 1998; McDonald, 2008; Taureck, 2006). It is important to note that the theory arose out of a concern with the "intellectual and political dangers in simply tacking the word *security* onto an ever wider range of issues" (Buzan et al., 1998, p. 1, emphasis in original). The core ideas

[2] "Traditional" or "state-centric" accounts of security, on the other hand, see international relations as occurring only between sovereign nation-states, with states seeking security in a global environment with no universal rules (Peoples and Vaughan-Williams, 2021; Walt, 2017). According to this theory, states' inherent competition with one another would compel them to defend their sovereignty against external threats, in this case, the economic and identity threats posed by migrants.

[3] In the fields of security and international relations, some of the earliest sources for the constructivist approach are Adler (1997), Buzan et al. (1998), Huysmans (1998), and Wendt (1992/ed).

of securitization theory can be attributed to scholars such as Buzan, Wæver, and others (Buzan et al., 1998; Wæver et al., 1993) who argue that speech acts that treat certain issues as security threats can lead to the expansion of security concerns into other "sectors." This process involves framing certain issues or problems as security threats and elevating them to the status of urgent matters requiring extraordinary action.[4]

Securitization theory thus clarifies the implications of simply adding "security" to identification. When identification undergoes securitization, it is presented as a pressing threat, justifying the use of extraordinary measures to address associated security concerns. Chapter 1 provided some examples of how securitization of identification may be seen in the context of the European Union's interoperability framework. In those quotations from policy documents, EU authorities framed the risks of unlinked data as threats to the internal security of the EU as these "unknown unknowns" or "blind spots" are presented as risks that can be exploited by individuals seeking to evade authorities. The construction of identification as a critical security concern supports what could be considered exceptional measures to address these risks through the EU's interoperability framework. For example, the integration of multiple databases and information systems, including those related to visas, asylum, and criminal records that were previously unconnected. Of course, such securitization can also fail if it receives little political or societal backing, as it may be seen as unjustified or out of proportion to the actual security threat.

Securitization theory has been criticized for its focus on the role of the state and its security apparatus in the construction of security issues (Aradau, 2004; Bertrand, 2018; McDonald, 2008; Peoples and Vaughan-Williams, 2021). Such a viewpoint emphasizes those who can speak, excluding those who have historically and structurally been marginalized. This results in a narrow understanding of security that fails to consider the complexity and diversity of security methods and devices. Therefore, besides the rhetorical structuring of security issues and the performativity of security speech, scholars set about to understand how security is enacted through security devices and in the everyday practices of security professionals (Balzacq et al., 2010; Côté-Boucher et al., 2014; Davidshofer et al., 2017). Hence, it is essential to examine the technologies, procedures, and protocols used to implement security measures, as well as how these measures are experienced and contested by those subjected to them. Scholars have provided a more comprehensive understanding of how security is enacted and experienced in practice by focusing on the material and embodied dimensions of security (Amicelle et al., 2015).

A helpful example of a focus on such security practices is given by Bigo (2014), who observed at least three different meanings actors attributed to border control: "solid," something to be defended and connected to sovereignty; "liquid," referring to how borders can be controlled through filtering and identifying populations; and "cloudy," only discernible through the use of digital databases and analytics. In all three of Bigo's ex-

[4] Furthermore, securitization theory enables one to think about the inverse process: how issues might be de-securitized and reintroduced into regular politics.

amples, the actors' practices and uses of security technology affected the meanings they gave to border control. On the one hand, the practices and technologies can contribute to the securitization of borders and mobility. On the other hand, the technologies created and used are shaped by the meanings attached to borders and border control. By studying the methods and tools used for identification and data matching, we can thus gain insights into the securitization of migration and border control.

Understanding the construction and enactment of security through identification and data matching technologies can be enhanced by examining it through the lens of materialist security studies (e.g., Aradau, 2012; Aradau and Huysmans, 2014; Bellanova and de Goede, 2022; Hoijtink and Leese, 2019; Walters, 2014). Materialist security studies seek to account for the development and use of devices in security practices and the agency of these devices. According to Pelizza (2021), one compelling feature of materialist security studies is their potential for blurring boundaries between states, private organizations, and commercial entities. Devices like surveillance cameras or biometric scanners are often developed and deployed by various actors, including governments, private security firms, and technology companies. As a result, examining security devices can disentangle different actors' roles and interests in shaping security practices.

Devices are not just passive tools in security practices; they also have performativity and can actualize security (e.g., Amicelle et al., 2015; Bellanova and Glouftsios, 2022; Davidshofer et al., 2017). For instance, watch list matching systems used in border security are an example of how identification and data matching technologies shape security practices. Such matching systems effectively process extensive datasets to generate matches and identify potential security threats at borders and airports (see, for example, Glouftsios and Leese, 2023). Additionally, biometric authentication systems utilized exemplify devices' performativity in security practices by producing and enacting security classification (e.g., Kloppenburg and van der Ploeg, 2020). A materialist perspective on security studies provides a framework for understanding the role of devices in shaping security practices. Focusing on the materiality of security helps illuminate the relationships between actors, devices, and security outcomes.

Similarly, the literature on Science, Technology, and Society (STS) has extensively explored identification in security settings, such as in the context of identifying possible threats in warfare through the lens of sociotechnologies (Follis, 2017; Suchman et al., 2017). For instance, consider data matching technologies utilized in security operations to identify potential threats by cross-referencing information from various databases. These algorithmic systems and technological tools play a pivotal role in shaping our understanding of what constitutes a threat and who is considered a legitimate target (Suchman et al., 2017; Suchman, 2020). However, scholars like Lucy Suchman have investigated how such sociotechnologies can lead to unintended consequences, creating insecurity for individuals who are wrongly identified or targeted. As such, this perspective urges us to carefully consider the potential risks and wide-ranging implications that arise from data matching and other technological means to identify legitimate targets in se-

curity contexts.

The connection between the securitization of identification and data matching thus raises several hypotheses, questions, and doubts that require exploration. Firstly, does the securitization of identification through data matching knowledge target specific groups of people, potentially exacerbating existing inequalities and biases? Secondly, as data matching technology becomes securitized, does the embedded matching knowledge travel between organizations, carrying logic and practices of security with it? Thirdly, do industry trends towards security influence the development and implementation of data matching technology in specific ways? Lastly, how does securitization of data matching technology shape the definition and identification of security threats, and what implication does this have for a broader understanding of security? Further research is necessary to understand these dynamics and the potential consequences for data matching, identification practices, and the broader security landscape.

This section has explored the ways in which critical security studies have highlighted the conversion of societal concerns into security issues and the possible consequences of these transformations. Considering the insights from this literature, it becomes crucial to question the justification for categorizing identification as a security issue and the widespread use of data matching techniques. Earlier theories in this field were often limited by a state-centric perspective and a focus on the discourse and speech acts of dominant actors, such as states and security professionals. Recent scholarship has emphasized the materiality and performativity of security practices and devices, as well as their potential to create insecurity for certain groups. This thesis seeks to investigate how security infrastructures shape and are shaped by the specific practices and devices utilized in matching and linking identity data. By examining these processes, we can better understand how identification practices can become securitized and the potential implications for individuals and society. In the following section, we will explore the construction, organization, and interconnection of identification systems, which ultimately create infrastructures that influence data matching and identification procedures.

2.4 Infrastructuring of identification

Literature on infrastructure has long demonstrated how data infrastructures reflect and shape the values, interests, and power relations of the actors involved in their design and operation, which consequently shape practices (e.g., Bowker and Star, 1999; Edwards et al., 2009; Hanseth et al., 1996; Kitchin and Lauriault, 2018). This has relevance in border security and migration control, where identifying and tracking individuals across different systems and databases has become a crucial component of state surveillance and control. For example, Glouftsios (2021) demonstrated how the often-invisible maintenance of large-scale information systems for border security in the EU has an essential role in sustaining the governance of international mobility. Likewise, Pelizza and Van Rossem (2023), through the notion of "scripts of alterity," make the point that data infrastructures

shape power relations with people and among states. The systems and data infrastructure for identifying people have thus rightly been a frequent concern of studies on border security and migration control (see, for example, Dijstelbloem, 2021; Glouftsios, 2021; Pollozek and Passoth, 2019). This perspective suggests that data matching is crucial for connecting various data sources for a broader data infrastructure, especially in linking biometric databases, watch lists, and migration and border management systems. Connecting various organizations, authorities, and their data into more comprehensive data infrastructures permits a more extensive identification and classification of individuals.

Yet defining what constitutes infrastructure often poses a significant challenge due to its multifaceted and dynamic nature (e.g., Edwards et al., 2007; Larkin, 2013; Karasti et al., 2016; Monteiro et al., 2013; Star and Ruhleder, 1996). Star and Ruhleder (1996), emphasize infrastructure dimensions such as embeddedness, transparency, reach/scope, learned as part of membership, links with conventions of practice, the embodiment of standards, built on an installed base, and becoming visible on breakdown. Recognizing these complexities helps understand how data matching is intertwined with the infrastructuring of identification, shaping the practices and norms of individuals like border guards and security personnel. The embeddedness of data matching technology within infrastructures may also contribute to transparency and seamless integration of multiple identity databases to enable exchange and interoperability of identity data. Applying Star and Ruhleder (1996)'s dimensions to identification infrastructure would highlight the gradual integration of systems within work practices, evolving over time while adhering to established standards. Such infrastructure operates imperceptibly, remaining invisible until significant events occur.

Similarly, Larkin (2013) captures the ontology of infrastructure as both "things and also the relation between things." His and other relational perspectives emphasize that infrastructure extends beyond its tangible components and encompasses the interdependencies and connections that facilitate its functionality. It underscores the significance of comprehending infrastructure as a network of relationships and dependencies, where material elements intertwine with social and organizational factors. In the context of identification and data matching, this means that infrastructure goes beyond mere databases and systems; it includes the web of relationships that enable effective data exchange and matching processes. Moreover, what may be considered infrastructure for one person or organization might not hold the same significance for others (Star, 1999). For identification and data matching, different actors may engage with infrastructures based on their unique needs, interests, and positions within the broader sociopolitical landscape. For instance, an immigration agency may prioritize a centralized and secure database infrastructure for accurate identity verification and efficient visa processing, while a law enforcement body may emphasize data matching technologies for cross-referencing identities across multiple databases.

In the context of data matching and identification in infrastructures, it is necessary to consider the impact of the rise of Big Data and its implications for the generation, circu-

lation, and utilization of data (Borgman, 2015; Dalton and Thatcher, 2014; Kitchin, 2014). These developments have sparked interest in new fields like Critical Data Studies (CDS), which seek to uncover the underlying power structures inherent in data practices (Flyverbom and Murray, 2018; Iliadis and Russo, 2016; Kitchin and Lauriault, 2018). While previous research on infrastructures and social order has taken a relational approach, scholars associated with CDS employ critical frameworks to highlight the power dynamics at play in data processes (Iliadis and Russo, 2016, p. 2). Furthermore, they emphasize that data are never raw but shaped by choices and constraints (see also, Gitelman, 2013). For example, consider a data matching process used for security purposes. Various datasets, such as passport records, visa applications, and biometric information, are combined to verify identities at border crossings. While the process may seem neutral, the choices made during data matching, like the weight given to data points in the matching algorithms, can have significant implications for individuals. CDS, therefore, prompts us to address such questions about the impact of datafication on power relations and the potential risks of data matching in shaping identification outcomes.

The definitions and characteristics discussed bring attention to the dynamic nature of infrastructures as they undergo evolution and adaptation in response to technological advancements, policy changes, and organizational shifts. This raises the question whether infrastructure can be purposefully designed and built. For instance, Edwards et al. (2007) emphasizes that infrastructure does not arise from intentional design but grows in phases. This perspective aligns with Hughes (1983b)'s account of the historical growth of the electrical supply network, where expansion often resulted from addressing critical problems within existing systems. Moreover, as we explore the role of infrastructures in identification systems, it becomes crucial to understand how knowledge and technology for matching identity data circulate and travel across organizations. Edwards et al. emphasize the pivotal phase of constructing gateways in infrastructure development. During this phase, multiple systems compete without a clear winner, and gateways facilitate interoperability, enabling more unified, integrated systems. By exploring how identity data matching technologies traverse organizational boundaries, we can gain insights into the mechanisms contributing to information exchange and interoperability across various organizations.

The connection between the infrastructuring of identification and data matching raises hypotheses, questions, and doubts that warrant further investigation. Firstly, the role of data matching software in facilitating integration and interoperability within identification infrastructures gives rise to hypotheses regarding its impact on the growth and expansion of such infrastructures. For example, do data matching technologies act as crucial components that link diverse databases and systems into broader infrastructures? Accordingly, questions worth considering relate to the implications of data matching for standardization efforts within identification infrastructures. How does data matching interact with existing systems, protocols, regulations, and standards? Does the deployment of data matching software promote or hinder standardization

across different identification systems? Additionally, investigating the relationship between infrastructuring and data matching gives rise to hypotheses regarding the often less visible work and technologies involved in matching identity data. How does the embeddedness of data matching technology within infrastructures influence the visibility of identification processes? How do the practices and technologies of data matching contribute to the construction and maintenance of identification infrastructures, and what implications do these have for the overall functioning of identification systems? Exploring these questions and hypotheses can deepen our understanding of the relationships between infrastructuring and data matching in the context of identification systems.

2.5 Research questions

2.5.1 Main research question

This chapter provided a more precise definition of the research problem and raised doubts, questions, and hypotheses related to data matching in four research areas, (1) the internationalization of identification, (2) the commercialization of identification, (3) the securitization of identification, and (4) the infrastructuring of identification. By analyzing the relevant literature within these areas, we have gained a deeper understanding of the interplay between identification, data matching, and transnational security infrastructures. In addition, the literature review has allowed us to identify key themes within each research area, including the growing use of internationally developed and commercially available identification systems and the increasing securitization of identification processes. These themes have helped to clarify the broader context of identity data matching in the context of transnational commercial security infrastructures.

Through our examination of the internationalization and commercialization of identification, we have identified a gap in research regarding the ways in which the spread of commercially and internationally available identification technologies influences identification processes. Similarly, in our exploration of the securitization of identification, we have identified a gap in research regarding how security concerns shape and are shaped by the design and use of identification systems. This securitization can be seen in the growing emphasis on biometric identification in border control and the linking of identity data to identify security threats. Additionally, our examination of the infrastructuring of identification has highlighted a gap in research regarding the implications of identifying people across data infrastructures. It is within this context that the significance of data matching emerges as a pivotal process in identification and data management that enables the linkage and correlation of identity data across diverse data sets.

The use of data matching technology for identifying individuals across various databases and international borders can raise many questions and uncertainties. When

it comes to the internationalization of identification, it brings up concerns about the heterogeneous data models of different authorities, how disjointed identity data can be linked across borders, and how cultural and contextual variations of identity data are considered. The commercialization of identification through data matching systems also raises concerns about the involvement of commercial entities, the risks of developing proprietary identification knowledge, and the standardization of identification solutions due to software reuse. The securitization of identification through data matching technology raises concerns about transferring security knowledge between organizations, industry trends in technology development, and defining security threats. Additionally, the infrastructuring of identification and data matching raises questions about the role of data matching in standardization efforts, the invisibility of data matching work in identification processes, and infrastructural growth.

Despite its widespread use, further research must explore the dynamics and implications of data matching, particularly in the context of emerging technologies, transnational infrastructures, and evolving security concerns. This dissertation aims to contribute to a deeper understanding of the role and influence of data matching technology on the interpretation of data, operational practices, and organizational structures.

Based on the gaps identified in the literature review, we can now formulate the following main research question:

Main research question How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?

The question focuses on how data matching shapes the meaning of data, practices, and organizations in the domains of migration management and border control. Furthermore, the question highlights the interconnectedness of identity data matching practices and technologies with transnational security infrastructures, underscoring the importance of comprehending the wider social, economic, and political implications of identity matching.

2.5.2 Sub-research question

A range of sub-research questions that delve deeper into the relationship between identification systems and transnational security infrastructures can help to answer our main research question. To this end, the research is structured around three sub-questions that investigate the types of knowledge embedded within data models used by national and transnational security infrastructures (RQ1), the processes and technologies for searching and matching identity data across different organizations (RQ2), and the circulation of knowledge and technology related to matching identity data across organizations (RQ3). In addition, the following sub-questions will address several hypotheses and inquiries regarding the connection between data matching

and the internationalization, commercialization, securitization, and infrastructuring of identification that were previously discussed. The three research questions are as follows:

RQ1 Which types of knowledge and assumptions about people-on-the-move are inscribed in data models of national and transnational security infrastructures? What implications does this have for how organizations can search and match identity data?

RQ2 How do organizations that collect information about people-on-the-move search and match for identity data in their systems? How is data about people-on-the-move matched and linked across different agencies and organizations?

RQ3 How do knowledge and technology for matching identity data circulate and travel across organizations?

The first research question (RQ1) seeks to uncover the types of knowledge embedded within data models used by different organizations and authorities — what Pelizza and Van Rossem (2023) have called "scripts of alterity." Hence, this research question draws inspiration from the technical process of schema matching used in data matching, to focus on data models' design, and the assumptions about people-on-the-move that can be inferred from designs. Uncovering differences can be useful to bring to light division of work in an international setting and throw light on how infrastructures shape identities. For instance, identifying records that refer to the same person could be challenging without knowledge of the underlying data models of the data sets being matched because the same person may be represented differently across data sets. By focusing on design, we can compare these assumptions to uncover differences in the division of work in an international setting and shed light on how infrastructures shape identities, thereby illuminating the internationalization of identification through the connections between the information about people-on-the-move gathered by various organizations.

The research for RQ1 contributes to infrastructure studies and the use of digital methods in analyzing data models in population management information systems. Methodologically, there is potential for innovation in rendering visible the data infrastructures for population management through an examination of the semantic data models. By analyzing "thin" data models, researchers can discern discrepancies and omissions between systems, laying the groundwork for critical analyses that illuminate the underlying power structures and biases embedded within these infrastructures. Additionally, examining data models presents an opportunity for experimental forms of participation and deliberation on alternative ways to represent data about populations.

The second research question (RQ2) focuses specifically on the ways in which organizations search for and match identity data related to people-on-the-move. This research will explore the processes and technologies used to link and match data across different agencies and organizations, and examine how identity data are collected, stored, and analyzed. By analyzing how data are matched and linked across different organi-

zations, this research can help us better understand the challenges and opportunities associated with managing data related to people-on-the-move. Furthermore, this question can shed light on how emerging and commercially deployed international identification technology shapes state practices. We can gain insight into the invisible aspects of data matching work within infrastructures, and better understand how these technologies shape and maintain identification systems.

In this way, RQ2 contributes to the current literature on identification by shifting the focus from the initial registration phase to the continuous retrieval and matching of identity data across diverse contexts. The question suggests considering the design and functionality of data matching tools, which play a pivotal role in reducing errors and streamlining iterative identification processes. In addition, the research contributes to a deeper understanding of the practical challenges faced by organizations and institutions in achieving data accuracy and efficiency during real-world identification procedures.

Finally, the third research question (RQ3) explores how knowledge and technology related to matching identity data circulate and travel across organizations. This research will examine the networks of actors and organizations involved in developing and disseminating technologies used for matching identity data and explore how knowledge about such technologies is shared and disseminated among different actors and organizations. Answering this question can aid in our understanding of the broader social and political contexts in which identification technology is developed and used to have long-lasting effects in numerous locations. Through this question, we can investigate how data matching technologies developed in one context can be adopted and adapted in different international settings, possibly affecting identification practices across borders. This research will also spotlight the role of commercial actors in acquiring and disseminating data matching knowledge, which could lead to the concentration of power and expertise in the hands of vendors or organizations. Additionally, exploring the circulation of knowledge and technology related to data matching aims to highlight the influence of securitization on identification practices, such as industry trends in security and how they affect software development.

By investigating the dynamic processes of identification technologies, the research on RQ3 enhances our theoretical understanding of identification systems and infrastructures, shifting away from viewing them as static systems. This approach broadens the range of actors traditionally examined in identification studies to include non-state and commercial entities. The involvement of these additional actors in developing data matching technology is expected to reveal connections between the internationalization, commercialization, securitization, and infrastructuring of identification. Focusing on the sociotechnical aspect of identification systems highlights the role of adaptable and functional identification infrastructures in serving diverse communities and ensuring that the systems remain responsive to their evolving needs.

Chapter 3

Towards a methodological framework for analyzing data matching in transnational infrastructures

This chapter develops a methodological framework that employs data matching within transnational commercialized security infrastructures to investigate the internationalization, commercialization, securitization, and infrastructuring of identification. To achieve this, three distinct methodological strategies are introduced, wherein data matching serves as both a research topic and a methodological resource. These strategies encompass 1) comparing data models, 2) analyzing data practices, and 3) tracing sociotechnical change. These strategies draw inspiration from Bowker (1994) and Bowker and Star (1999)'s concept of "infrastructural inversions," which emerged to tackle the tendency of infrastructure, associated practices, and technologies to fade into the background when functioning seamlessly. Similarly, this dissertation proposes to use three aspects of data matching as inversion strategies that can bring less visible aspects of identification to the forefront. Comparing data models can allow apprehending organizations' underlying assumptions about people in identification. Analyzing data matching practices can afford to comprehend how people are identified within and across organizations. Exploring sociotechnical change of data matching technology can enable tracking the circulation of identification knowledge, technologies, and practices over time and across organizations. The resulting methodological framework, which integrates these three inversion strategies, is then operationalized to address the research questions posed in this dissertation.

3.1 Data matching within data infrastructures as a topic and a resource of research

The methodological framework borrows the classical ethnomethodological understanding that "social structures [can serve as] both a topic and a resource for their inquiries"

(Garfinkel, 1964, p. 250). More recently, and particularly pertinent to this dissertation, this concept has been expanded to encompass digital devices, social media platforms, and data infrastructure as research subjects while also leveraging them as tools to explore other aspects of social life (Marres, 2017; Pelizza, 2016a; Pelizza and Van Rossem, 2023; Rogers, 2013; Weltevrede, 2016). For example, utilizing social media platforms for social inquiries involves using platform features like posts, likes, and tags to gain insights into social dynamics rather than just researching social media use. However, it also always entails grappling with the inherent ambiguity arising from the intertwined nature of the research process and the affordances of the studied medium (Marres, 2017; Weltevrede, 2016). In a similar vein, this chapter will employ data matching practices and technologies as both a research topic and a methodological resource for investigating the internationalization, commercialization, securitization, and infrastructuring of identification.

There are two compelling reasons for adopting data matching as a methodological resource. Firstly, considering data matching as a distinct research topic presents an opportunity to discover how matching and linking identity data is practically achieved. Existing research, as highlighted in the preceding chapters, has yet to comprehensively explore the nuances of how identities in migration and border control are interconnected across diverse systems and databases. Secondly, delving into the "technical minutiae" (Pelizza, 2016a) of data matching serves as a valuable resource for investigating broader shifts in identification. Data matching, characterized by its intricate interplay among various stakeholders, including diverse government agencies with interconnected data and commercial and security companies offering data matching technologies, can serve as a resource to explore transnational, commercialized security infrastructures in our research.

However, studying data matching presents its own challenges, given its propensity to operate inconspicuously within identification. Consider, for instance, the inconspicuous moments where our financial or travel data is automatically cross-referenced with police watch lists. Infrastructure studies have long been aware of such predicaments and have actively sought methods to bring these background elements to the forefront of analysis.

3.2 Infrastructural inversions for matching identity data

Infrastructure studies have provided different methodological strategies to invert the tendency of infrastructures to disappear and to make visible the interconnections between technical minutiae and the politics of knowledge production (e.g., Bowker, 1994; Bowker and Star, 1999; Edwards et al., 2009; Monteiro et al., 2013). Methods proposed by these authors to invert the tendency of infrastructure to disappear include looking at, among others, moments of breakdown (Star, 1999), tensions in the emergence and growth of infrastructure (Hanseth et al., 2006), and material aspects (Ribes, 2019). I propose to develop methodological strategies aimed at inverting transnational commercial-

ized security infrastructures to elucidate the links between the intricate technical aspects of data matching and the dynamics of knowledge generation in identification. As indicated by Bowker and Star (1999):

> Infrastructural inversion means recognizing the depths of interdependence of technical networks and standards, on the one hand, and the real work of politics and knowledge production on the other. It foregrounds these normally invisible Lilliputian threads and further more gives them causal prominence in many areas usually attributed to heroic actors, social movements, or cultural mores. (p. 34)

Expanding upon this quote in the context of data matching emphasizes the significance of acknowledging the less visible interdependencies between technical intricacies, politics, and knowledge production, often overshadowed by more apparent actors and social dynamics. This section will elaborate on three inversion strategies centered around 1) comparing data models, 2) analyzing data practices, and 3) tracing sociotechnical change. Comparative analysis of data models can provide insights into organizations' underlying assumptions about individuals, as evidenced by the distinct data categories employed in the heterogeneous models used in matching data across various data infrastructures. Examining data matching practices across different infrastructures can facilitate a comprehensive understanding of how people are identified within and across organizational boundaries. Lastly, delving into the sociotechnical evolution of data matching technology can offer the means of tracing the diffusion of identification knowledge, technologies, and practices over time, thereby uncovering the long-term infrastructural implications.

3.2.1 First inversion strategy: Comparing data models

While data matching necessitates addressing the technical alignment of categories when integrating data models from diverse origins (Christen, 2012), this aspect can be expanded to serve as a valuable research resource into the dynamics of knowledge production. Data models, designed to represent phenomena, are inherently interconnected with the politics of knowledge production as they become embedded in infrastructures and practices, instituting specific ways of knowing and working (Bowker and Star, 1999; Bloomfield and Vurdubakis, 1997; Hine, 2006; Lampland and Star, 2009; Timmermans and Epstein, 2010). Designing data models and the knowledge they aim to capture is inherently political, determining how information is presented, what aspects are emphasized, and which elements may be marginalized (Bowker and Star, 1999). This is particularly relevant in the context of border and migration management, where data models are used to categorize and identify individuals on the move across borders (Pelizza and Van Rossem, 2023). While data models delineating specific categories of data may seem to offer limited information, they can be considered valuable research topics and methodological resources.

As a research topic, a systematic analysis of data models involving the detection of disparities and omissions holds promise in recovering the knowledge representations. Technical solutions designed to compare data models and automatically identify correspondences between data model concepts — as seen in knowledge engineering, linked data, and natural language processing techniques (Euzenat and Shvaiko, 2007; Kementsietsidis, 2009) — also emphasize the potential for recovering knowledge representations from discursive domains through an exploration of the relationships inherent in data models. In this context, a parallel can be drawn between these knowledge engineering methods, which aim to enhance machine comprehension of data for improved system functioning and information source integration, and data matching's need for establishing connections between data models.

Data models can also be considered a methodological resource as a source of knowledge about their producers (Pelizza and Van Rossem, 2023). Through comparative analysis, such as examining distinct authorities' data models for identifying individuals, we can unveil not only the presence or absence of specific categories within one model as opposed to another but also reveal the inherent possibilities, limitations, and thus the underlying conceptions of diverse authorities regarding individuals in identification (Van Rossem and Pelizza, 2022).[1] A pertinent illustration comes from Bowker and Star (1999)'s exploration of classification systems, such as their work on the standards for categorizing nursing work, unveiling the types of work organizations value. In another instance, Cornford et al. (2013) compared digital data standards initiated by public services in the United Kingdom to represent familial relationships. Their study showcased how data models can be employed to analyze "the kinds of family relationships that are recorded and those that are not recorded or harder to record, any hierarchies, implicit or explicit, for family forms or relationships and the implicit and explicit assumptions that underlie the terms and classifications used" (p. 8). Thus, the first inversion strategy capitalizes on insights gained from data matching's need for data model alignment and extends this as a methodological resource to uncover the implicit assumptions about individuals within various identification systems.

3.2.2 Second inversion strategy: data matching practices

The second inversion strategy focuses on another dimension of data matching: the actual alignment of data within and across databases. Acknowledging the well-established notion that data cannot be regarded as "raw" values effortlessly aligning with their database models, this strategy draws on the notion that actual data results from a combination of

[1] As semiotics has theorized, meaning emerges from comparison. For example, Latour (2005) explains the significance of not defining groups a priori because "whenever some work has to be done to trace or retrace the boundary of a group, other groupings are designated as being empty, archaic, dangerous, obsolete, and so on. It is always by comparison with other competing ties that any tie is emphasized." (p. 32). Similarly, Pelizza (2010) recalls that "the situations where the social is made visible and graspable are those where meaning emerges from comparison" (p. 67).

localized human decisions and technical limitations (Bowker and Star, 1999; Gitelman, 2013). The process of matching data across varied organizations is thus thought to be intricately entwined with challenges stemming from ambiguity and uncertainty in data, which emerge from these diverse local contexts. Consequently, delving into data matching practices within the realm of identification unveils the practical mechanisms and challenges while also providing an opportunity to gain insights into the complex relationships that unfold among diverse actors in the interlinking of their data within transnational commercialized security infrastructures.

This recognition of the dual use of practices has also been well-established in the field of "practice theory," which regards social practices as integral to broader societal investigations (Reckwitz, 2002; Schatzki, 2005; Shove et al., 2012). For example, practice theory scholars have examined routine activities like cooking (Rinkinen et al., 2019) or daily showering habits (Hand et al., 2005) to gain insights into broader societal dynamics, such as food supply chains or patterns of energy consumption. A focus on practices has also been taken to understand the interlinkages between data practices and knowledge production in migration and borders (Cakici et al., 2020; Scheel et al., 2019). For instance, to investigate how practices contribute to the enactment of race (M'charek et al., 2014), the government of mobilities (Glouftsios, 2018), or how both European and non-European populations are enacted in Europe, along with the simultaneous enactment of institutions through data practices and infrastructures (Pelizza, 2019). Analyzing data matching as a practice can offer a similar dual opportunity.

Firstly, it can enable the exploration of everyday practices associated with matching identity data and the dynamics of their evolution and adaptation. By analyzing data matching practices, we can also gain insight into the specifics of how technologies and practice mutually influence each other, revealing their evolving interdependence (Schatzki, 2010; Ruppert and Scheel, 2021; Hui et al., 2016). For instance, we could investigate how practitioners effectively match and link identity in the context of border control despite the heterogeneity between databases. In doing so, we can look at the specific technologies for matching identity data and how their embedded expertise in matching identity data, such as rules for determining name similarity, shapes the mutual interdependence between technology and practice. Such an examination of mutual relations may be crucial in understanding the interplay of data matching technologies, such as flagging irregularities in identity data that may lead to doubts about someone's identity at the border.

Secondly, examining data matching as a practice can provide insights into the broader interconnections between practices and other significant societal phenomena (Shove, 2016; Nicolini, 2016). Just as the study of everyday practices such as driving can unveil interconnected patterns of social activities that illuminate broader societal dynamics, such as energy consumption (Shove et al., 2015), the investigation of data matching practices is anticipated to unveil how these practices are shaped by and contribute to broader shifts in identification. Analyzing routine data matching practices

will thus be used as a resource for investigating how these practices are intertwined with larger sociopolitical and economic forces, playing a role in the internationalization, commercialization, securitization, and infrastructuring of identification within the context of transnational commercialized security infrastructures.

3.2.3 Third inversion strategy: sociotechnical change

The third inversion strategy leverages the evolving developmental trajectory of data matching technology, employing it as both a research topic and a methodological resource. In migration and border control, data matching can be considered as operating as a component of more extensive infrastructures that interconnect various contexts and temporal scales, facilitating the exchange and utilization of information collected across diverse locations and moments. The evolution and use of data matching technologies will thus be interwoven with broader infrastructural developments, characterized by interactions among individuals, states, companies, and technologies, often giving rise to tensions as they strive to achieve distinct objectives (Edwards et al., 2007; Ribes and Finholt, 2009). As previously noted, tensions and changes within such infrastructures tend to become invisible over time. While the first inversion strategy proposes to recover such design choices and tensions through comparative analysis, tracing technological evolution is another approach to investigate the matching and linking of identity data in transnational commercialized security infrastructures. Although researchers often concentrate on the final products of identification systems, adopting a methodological stance that traces their development can be advantageous for two reasons.

Firstly, tracing sociotechnical evolution can unveil technological design choices and elucidate the interactions among users, designers, system builders, and technological artefacts. Designers encode specific expectations into technological artefacts regarding users, systems, and data, which can, for example, be uncovered by focusing on how designers encourage specific interpretations (Woolgar, 1990) or by comparing successful and failed adaptations to actual usage (Akrich and Latour, 1992; Akrich, 1992). For instance, data matching systems may hold certain expectations about how heterogeneous identity data can be matched, observable in technical specifications or instances of success or failure. Overall, the creation of technologies involves long processes with numerous choices and unforeseen outcomes, where diverse social actors often attribute varying meanings to artefacts and utilize technologies in distinct ways (Jackson et al., 2002; Pinch, 2008; MacKenzie and Wajcman, 1985; Pinch and Bijker, 1984). For example, meanings and utilization of data matching technologies may have evolved differently between contexts like border and migration control, healthcare, or finance. Documenting the design choices and trajectories of data matching technology as a research topic can prove valuable for comprehending the interplay between users, designers, and technologies throughout the practical design, development, and real-world utilization of data matching technology, shedding light on embedded expectations and providing insights from both successful and unsuccessful adaptations to actual usage.

Secondly, utilizing sociotechnical evolution as a resource can facilitate exploring how these evolutions are intricately intertwined with broader sociopolitical and economic forces that both shape and are shaped by transnational commercialized security infrastructures. The instances when social groups challenge, alter or stabilize the meanings of technologies (Pinch and Bijker, 1984) can offer insights not only into how technologies adapt but also into the evolving sociotechnical problems and solutions related to linking and matching identity data. This perspective can help understand how the technical intricacies of data matching intersect with the redefinition of sociotechnical challenges and solutions for identity data linking and matching, particularly within the context of growing securitization. Moreover, data matching plays a role in interconnecting previously independent systems into more extensive infrastructures. These moments can bring to light the technical challenges that arise when integrating once incompatible systems, such as requiring gateways for compatibility (Edwards et al., 2007). These junctures offer insights into broader dynamics within identification processes, including the impact of commercially and internationally deployed data matching systems on the standardization and stability of data matching practices.

3.3 A methodological framework for analyzing data matching in transnational infrastructure

3.3.1 Working model for operationalizing the inversion strategies

This section combines the three inversion strategies into the methodological framework, which will be operationalized for the dissertation's investigation into how practices and technologies involved in matching identity data are both shaping and shaped by transnational commercialized security infrastructures. Hence, the operationalization of this framework encompasses three approaches to using data matching, each rooted in the inversion strategies and aligned with the overarching research questions expounded in Chapter 2:

1. Comparative analysis of data models: This strategy involves examining the types of information collected by various organizations and systems, providing insights into organizations' underlying assumptions about people through comparisons (RQ1).
2. Examination of data practices: This strategy centers on routine identification practices, investigating how identity data is matched and linked within and across organizational boundaries (RQ2).
3. Investigation of sociotechnical change: This strategy delves into the sociotechnical transformations within data matching software, exploring the trajectories and dissemination of data matching knowledge, technologies, and practices over time and across various organizational contexts (RQ3).

Table 3.1: Methodological framework.

RQ	Inversion	Pair	Data	Analysis
RQ1	Comparing data models	(data models, categories of data)	Traces of data models	Mixed (The Ontology Explorer)
RQ2	Data practices	(categories of data, data values)	Fieldwork WCC & IND	Deductive & inductive coding
RQ3	Sociotechnical change	(data models, data values)	Fieldwork WCC	Thematic and historical analysis

These three strategies will harness diverse data collection and analysis methods, which will be elaborated upon in the subsequent sections. The accompanying table delineates the relationships between the research questions, inversion strategies, and the techniques employed for data collection and analysis.

The methodological framework can also be visually represented in a three-dimensional graph, as illustrated in Figure 3.1. This graphical representation resembles a relational model, where data are organized into tables with interrelationships. The visualization serves as both a guide for analysis and a communicative tool, conveying relationships between inversion strategies, data matching dimensions, and research questions. The three axes of the graph represent distinct data dimensions: *data models*, *data categories*, and *data values*. To better understand these dimensions, we will first explore each axis in more detail, their characteristics and implications for data matching. After that, subsequent sections will explore the interconnections between these axes and how they correspond to inversion strategies.

First, the axis for *data models* represents the database and other informational models that standardize the kinds of information about people collected by different organizations' IT systems. An example of such a data model could be the schemas specifying data collected about migrants by a government agency. Each system and database may have its own unique data models, while standardized formats can also exist for data exchange and interoperability between different systems. Second, the axis for *data categories* refers to the specific types of information captured about individuals within a data model. For instance, in the example of the asylum agency's data model, data categories may include "name," "place of birth," and "date of birth."

Connections between data models can often be identified, whether explicitly stated or implied. For example, one data model may utilize the data category "surname," while another may use "family name" to represent the same concept. These variations demonstrate attempts to capture the same real-world information within different data models. Third, the axis for *data values* pertains to the actual values stored in databases corresponding to a particular data model and its associated data categories. For instance, a data value for the "place of birth" category could be "Brussels." It is worth noting that similarities between data values can exist across different databases. For instance, another database might contain the data value "Bruxelles" for a similar category. By considering these three axes, we can now explore three combinations of them to illustrate three interconnected aspects of data matching.

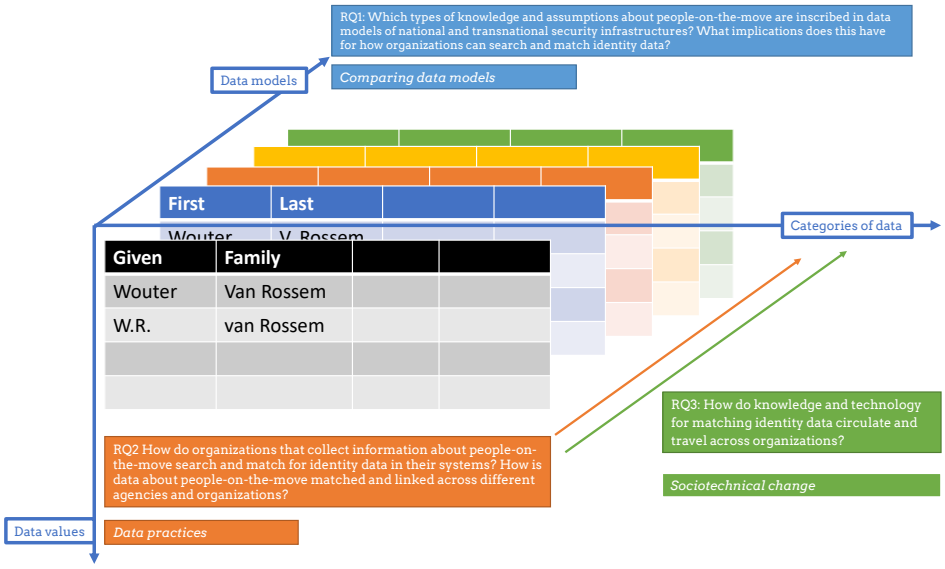


Figure 3.1: The methodological framework is represented graphically in three dimensions in this visualization.

3.3.1.1 Comparing authorities' data models

The relationship between the axis for data models and the axis for data categories represents the similarities and differences between various data models through their corresponding data categories. This relationship forms the operationalization of the first inversion strategy, which involves conducting a comparative analysis of data models. Within this strategy, the focus is on examining the types of information collected, represented as data categories, by various organizations and systems, which are represented by data models. Through these comparisons, we can examine organizations' underlying assumptions about individuals (RQ1).

The operationalization of examining specific authorities' data models aligns with the objectives of Work Package 2 within the Processing Citizenship project. Therefore, it is necessary to contextualize the objective of data model analysis within the broader scope of the PC project, which seeks to investigate how national and European authorities formalize knowledge about individuals in data models for managing migration. Through the PC project's script analysis methodology (see also, Pelizza and Aradau, 2024), comparing data models contributed to producing a typology of "intended migrants," which is the formalized knowledge of migrant identities inscribed in information systems (Pelizza and Van Rossem, 2023).

Data model analysis aimed to detect differences in scale, mainly through compar-

ing data models between EU and Member State authorities. Specifically, the study considers three significant information systems developed by European Commission agencies—Eurodac, the Schengen Information System (SIS), and the Visa Information System (VIS)—as representative of EU authorities. These three systems employ distinct data models to support various policing tasks related to travel, cross-border crime, and irregular migration. For national authorities, the analysis includes the data models of the Hellenic and German Register of Foreigners. The specific data models were thus chosen as part of the broader Processing Citizenship's task plan of examining actual identification practices in "processing alterity" (Pelizza, 2019).

The EU and national systems are characterized by diverse data models tailored to fulfill distinct functions within policing, travel, cross-border crime, and irregular migration management. However, within this array of systems, there are thought to be both disparities and commonalities in terms of data models and data categories. Eurodac, for instance, serves the purpose of aiding in identifying asylum seekers via fingerprinting while concurrently determining the Member State responsible for processing their asylum applications as part of the Dublin System.[2] This system predominantly collects the fingerprints of asylum seekers. On another front, the Schengen Information System (SIS II) contributes to external border control and law enforcement cooperation within the European Union.[3] It stores alerts containing comprehensive information about individuals and objects and directives for actions when encountering these entities. The Visa Information System (VIS) facilitates the exchange of visa-related data, encompassing personal and biometric information, in support of a unified EU visa policy. On the national level, the Hellenic Register of Foreigners is used to identify and register individuals arriving at the border in Greece. This system also collects data beyond mere identification, aiding various tasks encompassing the asylum process and assessing health conditions (Pelizza and Van Rossem, 2021). Similarly, the German Register of Foreigners (GRF) holds a substantial repository of personal information about foreigners in Germany, including residence permit holders, asylum seekers, and recognized refugees (Bundesverwaltungsamt, 2021). Chapter 4 delves further into these EU and national systems.

Two key considerations will need to be addressed in operationalizing the inversion strategy, focusing on comparing data models. Firstly, data models come in various formats, and specific formats might be less accessible due to their confidential nature, particularly within the context of migration and border control. In such instances, alternative forms of data model description, such as legislative documents or graphical

[2] The Dublin System (Regulation No. 604/2013; also known as the Dublin III Regulation) establishes the criteria and mechanisms for determining which EU Member State is responsible for examining an asylum application. The system is currently operational in the Member States of the European Union (plus Norway, Iceland, Switzerland, and Liechtenstein). Law enforcement agencies and Europol also use the system, which has significantly expanded its original scope since its inception (Ajana, 2013). New proposals aim to include more biographic and biometric data, including a facial image (Procedure 2016/0132/COD, a recast of the Eurodac Regulation).

[3] In addition, the SIRENE network can exchange this information between law and border enforcement authorities.

user interface screenshots, will need to be relied upon for analysis. These substitutes should allow for the reconstruction of data models, enabling a comprehensive comparison across diverse authorities. Secondly, developing connection methods is crucial to identifying links between data categories in distinct data models, as these connections may not always be evident. For instance, one authority's data model may feature the data category "family name," while another may employ "surname" to refer to the same concept. Although the terminology differs, both categories pertain to an individual's familial identity. Similarly, one data model may categorize all languages simply as "language," whereas another might differentiate between "native language" and "spoken languages." Chapter 4 will develop and use a method for extracting, analyzing, comparing, and visualizing data models from heterogeneous sources. However, beyond merely identifying these differences, these disparities will be used for gaining insights into authorities' expectations and intentions about individuals.

3.3.1.2 Data matching within and across organizations

The second pair involves the relationship between the axis representing data categories and the axis denoting data values, establishing connections between data categories and their corresponding database entries. Consider, for instance, a database containing data categories such as "first name" and "last name," wherein individuals' records are associated with specific values. However, deviations from these data models' expectations are commonplace. Instances may arise where first and last name values are inadvertently interchanged, or databases contain duplicate entries for the same individual. This pair of axes is connected to examining identification practices, which inherently involve dealing with inconsistent data—a challenge often addressed through data matching technology. By employing the second inversion strategy, our investigation delves into the routine practices of identity identification, scrutinizing how identity data is matched and linked.

Furthermore, the connection between the axis of data models and the axis of data values can be conceptualized as data matching practices across databases and organizations. In this context, data matching involves the procedures related to identifying and potentially linking or merging identity data that is distributed across multiple databases and organizations, where records of identity data for the same individual may coexist. A notable challenge for data matching mechanisms in this context is the inconsistent availability of unique identifiers, necessitating the reliance on, at times, ambiguous personal information. For instance, a woman might use her married surname in one system and her maiden name in another following marriage. Consequently, data matching processes across different organizations must consider variations in identification and registration practices. By employing the second inversion strategy, our research not only investigates the practices and technologies involved in matching data within and between organizations, but also explores differences and shifts in identifying individuals across various data infrastructures.

The operationalization of this aspect of the empirical framework, using the second inversion strategy to examine data matching practices, will be rooted in an empirical investigation of identification procedures within a government migration agency and its utilization of data matching software. Specifically, the research will focus on the applicant identification processes at the Netherlands' Immigration and Naturalization Service (IND) and its interactions within broader inter-organizational networks. This investigation will particularly emphasize the role of the ELISE data matching software in these practices. The IND is responsible for processing residency and nationality applications, and it incorporates data matching software into its identification procedures for searching and matching applicants' identity data within the back-office system and managing data anomalies like duplicate records. The analysis of data practices will rely on data gathered through fieldwork, including interviews, documents, and field notes, conducted at both the data matching software provider and the IND agency itself. Further details regarding setting up this fieldwork will be elaborated upon in subsequent sections of this chapter.

The operationalization of the second inversion strategy involves analyzing diverse data matching practices within the agency. This analysis encompasses at least three critical data matching practices identified in technical literature: batch data matching, real-time data matching, and data deduplication (Christen, 2012). In the context of matching identity data, these practices operate at distinct moments. Batch data matching entails an offline, scheduled batch processing of data sets to identify matching identity data. For example, an organization might periodically enhance the data about individuals in their database by matching and integrating data from various sources. In contrast, real-time data matching addresses the need for immediate data retrieval through direct search queries. For instance, police officers may need to query databases using data categories such as "name," "nationality," and "date of birth" to identify approximate matches for these personal details instantly. Finally, data deduplication employs data matching technology to detect and merge multiple records in a database that are considered pertaining to the same individual.

The methodological framework's focus on specific data matching practices is twofold. First, it involves analyzing the practices and technology used to match applicant data within the IND's data infrastructure. This approach provides insights into the intricacies of the practices and technologies used to match identity data within the IND's operational context. Second, the research investigates how individuals are identified across the broader data infrastructure. This dual perspective not only sheds light on the identification mechanisms but also highlights the connections between shifts in identification practices and the integration of commercial systems for matching identity data. By examining these dynamics, the study offers a comprehensive understanding of the interplay between data matching practices, technological solutions, and broader shifts in identification within the context of transnational commercialized security infrastructures.

3.3.1.3 Traveling data matching software and expertise

A perspective spanning all three axes (data models, data categories, data values) can be conceptualized as the expertise ingrained within data matching technologies, which encompass expertise related to matching models, data categories, and values, along with the potential dissemination of data matching knowledge and technology across multiple organizations over time. The underlying assumption is that businesses develop and deploy tools to facilitate data matching efforts in diverse contexts and eras. As a result, expertise in domains such as comparing identity records with typographical errors and name variations may similarly propagate from one organization to another over time through these data matching technologies. Employing the third inversion strategy, this approach investigates the sociotechnical evolutions within data matching software, thereby examining how knowledge, technologies, and practices related to data matching may circulate across different organizational contexts and evolve over time.

The operationalization of this aspect of the empirical framework involves examining the evolutionary trajectory of data matching software technology. In particular, this analysis will focus on the ELISE data matching software, a commercial product developed by the company WCC[4], which is widely deployed in various national and international border and migration control systems. The analysis will draw from fieldwork with WCC, encompassing the exploration of ELISE software development and deployment, as detailed in subsequent sections of this chapter. As explained above, operationalizing this third inversion strategy for data matching technology can offer insights into the dynamics between users, designers, and technologies throughout the practical phases of design, development, and real-world application. Moreover, tracing the sociotechnical evolution of the software is expected to contribute to a deeper understanding of broader shifts in internationalization, commercialization, and securitization within the realm of identification.

To operationalize the inversion strategy effectively, it is necessary to establish the appropriate heuristics to identify the analytically significant moments in the evolution of data matching software. These moments represent times when the interactions and involvement of users, designers, and other actors reveal the dynamics and evolution of the data matching process (Hyysalo et al., 2016). In Chapter 6, two distinct heuristics will be introduced to pinpoint such pivotal moments within the lifecycle of identification technologies. The first heuristic will draw from the Social Construction of Technology framework's concept of "interpretative flexibility," (Pinch and Bijker, 1984) to junctures when social groups challenge, alter, or stabilize the meanings attributed to identification practices and technologies. The second heuristic employs the concept of "gateway problem" (Edwards et al., 2007) from infrastructure studies to pinpoint moments where diverse identification software systems and infrastructures intersect and interact. These proposed heuristics will offer ways to capture and analyze shifts and developments in

[4] Went Computing Consultancy Group BV.

the evolution of data matching technology and how these evolutions are intricately intertwined with broader sociopolitical and economic forces that both shape and are shaped by transnational commercialized security infrastructures.

3.3.2 Contextualizing the research within the Processing Citizenship project framework

The methodological framework is closely linked to Processing Citizenship's goal to examine how identity management systems support the creation of knowledge about non-European populations (Pelizza, 2019; PC, 2017/2023). Consequently, this section delineates the pertinent Work Packages (WPs) within the Processing Citizenship project framework, aiming to emphasize this dissertation's distinct contributions to each of these WPs.

WP1 *To develop a theoretical framework that integrates globalization studies, border studies and surveillance studies of migration with science and technology studies and media geography. Design a coherent methodological approach that combines ethnographic and computational techniques in order to establish information systems mediated registration practices at Hotspots as analytical sites.*

This dissertation aligns with the objectives of WP1 by contributing to developing a methodological approach that combines ethnographic and computational techniques for analyzing registration practices at Hotspots. Specifically, the utilization of the second inversion strategy, which involves comparing data models, is in line with this goal. Chapter 4 introduces a methodology and software tool for comparing data models, facilitating the computational analysis of diverse authorities' assumptions about individuals. The insights derived from this approach serve as a foundational component for critical analyses when integrated with ethnographic observations of the practical utilization of information systems. This differentiation aligns with the approach of Processing Citizenship, which aimed to identify "intended migrants" and compare them with actual migrants using ethnographic analysis.

WP2 *To analyze and compare information systems used to register migrants across diverse Hotspots.*

WP2 encompassed a range of tasks, including the semantic analysis of the ontologies underpinning information systems and identifying the "intended migrant" for relocation or resettlement using ontologies and algorithms. The Ontology Explorer method and tool, introduced in Chapter 4 and Van Rossem and Pelizza (2022), which is based on comparing data models, provided methodological support for these tasks. This methodology offered a comprehensive means of comparing data models employed by different authorities. Furthermore, the analysis of intended migrants was further advanced in the publication Pelizza and Van Rossem (2023), shedding light on the scripts of alterity that delineate the assumptions and limi-

tations of border security frameworks through classification schemas. Additionally, as part of tasks related to WP2, interviews were conducted with IT developers within the EU, and these interviews were carried out during the fieldwork with the supplier of data matching software.

WP3 *To describe identification and registration practices at Hotspots, focusing on the material devices involved, and assess them on the basis of migrants' adaptation or resistance.*
The Ontology Explorer (OE) not only offers a valuable method and tool for comparative analysis of the scripts related to uncovering authorities' assumptions of people to be registered. Moreover, it contributes to WP3's objectives by providing a lens to explore resistance to identification and registration practices. People do not resist in isolation but rather within the context of these scripts. The script analysis made possible by the OE thus serves as an initial analytical step, offering a foundation that can be utilized by fellow researchers within the Processing Citizenship project. While this dissertation does not directly address this WP, the tool was considered a way to support their ethnographic investigations to uncover various forms of resistance enacted by diverse actors.

WP4 *To map architectures of data circulation in EU migration information systems.*
The main contribution here is to the specific research question (RQ7A) "How are relationships between EU and MS enacted through efforts to achieve interoperability?" This dissertation contributes to WP4 by addressing the task of investigating tensions, data frictions, and controversies related to classifications, standards, and semantic interoperability. While the original plan was to delve into semantic interoperability at eu-LISA, the focus shifted towards examining a supplier of identity-matching software for the EU-VIS system. In Chapter 6, the dissertation conceptualizes the integration of WCC ELISE, a data matching software, as a gateway moment. This analysis illustrates the software's role in facilitating semantic interoperability between EU and MS systems. Furthermore, the analysis of the data matching system in EU-VIS highlights the specific relations between EU, MS and commercial actors who needed to configure it while balancing new features and adhering to backward compatibility needs from EU member states' systems.

3.3.3 Fieldwork context, goals and limitations

In this section, more context will be provided regarding the establishment of the fieldwork, encompassing its objectives and constraints. As elucidated earlier, one of the tasks outlined in the Processing Citizenship project was to engage in interviews with IT developers in the EU to identify tensions, data frictions and controversies over classifications, standards and semantic interoperability. At first, we sought to collaborate with eu-LISA on fieldwork related to data quality, but unfortunately, the plan did not come to fruition. This failed attempt led to a recalibration of focus toward scrutinizing a supplier of data matching software employed in both EU and Member State systems. Furthermore, the fieldwork presented the opportunity to investigate data matching practices and trace the

evolutionary trajectory of data matching software.

3.3.3.1 Establishing the fieldwork site: failures and successes

The first, unfortunately unsuccessful, attempt at opening up a potential fieldwork site concentrated on the problems and technological solutions EU institutions face with identity data and data quality in border security and migration management. Establishing contacts with eu-LISA, the European Commission agency managing several EU border management information systems, was possible with the support of the PC project's Principal Investigator. It was possible to participate in the 2018 annual eu-LISA conference and meet with the head officer for research and development to discuss a potential traineeship on-site at the agency's office. After an initial agreement, I applied for a traineeship with a proof-of-concept proposal based on probabilistic databases, an active area of research that uses a database model based on working with uncertain data (van Keulen, 2018). Such a proof of concept was thought to potentially give insight into data uncertainties that the EU and MS encounter. Unfortunately, the proposal had to be abandoned by August 2019 due to issues surrounding the confidentiality of the data that the research would collect. As such, the study faced common barriers to qualitative secrecy research; the proposal could not pass the "gatekeepers" that could permit access to the fieldwork site (de Goede et al., 2020).

The second attempt at opening up a potential fieldwork site was successful by focusing instead on an eu-LISA supplier of identity-matching software. By looking at other technology companies that work with eu-LISA, we pinpointed the company WCC Group, which develops technology for matching identity data in border security and migration management. Indeed, the EU Visa Information System uses WCC produced software, namely the ELISE ID platform, to search and match identity and visa data. In short, ELISE provides data matching for fast querying based on inexact data. The company's proprietary data matching technology uses various fuzzy logic algorithms for data matching that consider that data may be incomplete or inaccurate. In this context, fuzzy logic refers to a type of mathematical logic that computes truth variables using probabilities rather than boolean "true" or "false" values. The difference between these approaches is how they handle data uncertainties and produce results.

For example, consider a scenario where the data matching technology identifies potential duplicate records in a customer database. In a boolean search, the system would match exact values, such as names, email addresses, or phone numbers, and return either "true" (a match) or "false" (no match). In contrast, using fuzzy logic, the technology considers the possibility of minor variations or errors in the data, such as misspellings, nicknames, or incomplete information. It then calculates match probabilities based on the similarity level between records. This approach means that records with slight discrepancies can still be identified as potential matches, with varying degrees of confidence in their accuracy. By employing fuzzy logic algorithms, the company's data matching technology can provide more comprehensive results by accommodating variations and

uncertainties in the data. The company and its software were thus deemed as excellent opportunities to empirically observe the problems and solutions of matching and linking identity data in security settings. Indeed, next to using ELISE in VIS, WCC has various other customers who use the software in border security and migration management settings.

A meeting was set up in December 2019 with WCC at the company's headquarters in Utrecht to discuss a research project that would meet the needs of the Processing Citizenship and the PhD project's research and live up to WCC and its customers' expectations. Sometime after, we proposed a project for on-site research focusing on the deployments of ELISE in the Visa Information System and the Netherlands' immigration and naturalization government agency, which WCC later approved. Unfortunately, at the same time, the Covid-19 virus spread worldwide and became a global pandemic. As a result, the government of The Netherlands announced in March 2020 that people should work from home. We decided against starting remotely and instead pushed back the start date because we believed the research would benefit more from face-to-face interaction.

In May and June of 2020, the government of the Netherlands loosened up some restrictions on working on-site and taking public transportation. Therefore, we decided to start the fieldwork hybrid, with remote access to relevant resources and a few on-site visits to the Utrecht office when the staff members were present. During the summer of 2020, the office environment had a relaxed atmosphere. However, due to COVID-19 restrictions, only a limited number of people were allowed to be present at the office, and many employees were on holiday schedules. As a result, the office was less crowded than usual. In September 2020, the government introduced stricter COVID-related regulations, and subsequent fieldwork had to be conducted online. More information regarding these limitations can be found in later sections and chapters.

3.3.3.2 Fieldwork scope and links with the methodological framework

Aligned with the overarching objectives of Processing Citizenship to explore tensions between the European Union (EU) and Member States (MS), the initial focus of the fieldwork centered on the utilization of the ELISE data matching software within the context of two distinct customers of WCC: one within a national context and the other within a transnational context. This dual focus is also associated with two infrastructural inversion strategies. Firstly, the deployment of ELISE within national settings was scrutinized through its integration with the systems of the Netherlands' Immigration and Naturalization Service (IND). In this instance, the second infrastructural inversion strategy was employed, delving into the data matching practices of IND to examine identity data matching within and across organizations. Secondly, the investigation extended to ELISE's application within transnational settings, observed through its integration into the EU Visa Information System (VIS). This initial focal point expanded to encompass a trajectory analysis of the data matching software. Here, the third inversion strategy was employed, investigating the deployment of VIS in conjunction with the trajectory of the

data matching software, thereby providing insights into broader shifts in identification and the underlying socio-political and economic dynamics.

Within the context of the Immigration and Naturalization Service (IND), the ELISE data matching system plays a role throughout the agency's bureaucratic process of evaluating applications from individuals seeking residence or citizenship in The Netherlands. The agency integrates the ELISE system into its information system to streamline the search for applicant data within its back-office system. Through the fieldwork, this study delved into the data matching practices of diverse organizational actors within the IND as they leverage data matching capabilities in their daily tasks. The technical documents and meetings with WCC staff provided insight into the technical integration of the ELISE data matching system into the IND's information systems.

Additionally, semi-structured interviews were conducted with IND staff, focusing on various aspects of their applicant identification procedures and their utilization of data matching tools. These interviews explored the nuances of formulating search queries, the computation of matches, handling search outcomes, and addressing data-related issues such as managing duplicate records. Consequently, this research had the unique opportunity to unveil a spectrum of re-identification practices, a concept introduced in Chapter 5 to encapsulate the continual use and interlinking of data from diverse sources to validate whether disparate sets of identity data correspond to a singular real-world individual. Thus, adopting data matching practices as an inversion strategy led to recognizing a diverse array of practices associated with matching identities within and across organizations. Furthermore, by juxtaposing the design of the systems with practical use, the fieldwork investigation shed light on numerous forms of data friction that can impede successful re-identification efforts. For instance, the fieldwork uncovered the challenges of unsuccessful re-identification, discernible through duplicate records and the labor-intensive deduplication process.

The second part of the fieldwork pertained to the utilization of the ELISE data matching system within the context of the EU Visa Information System (EU-VIS), which functions as the central hub connecting member state VIS systems. In this framework, the ELISE data matching system serves as the search engine for facilitating connections between the member states' systems and the EU-VIS, enabling the execution of search queries for retrieving visa data. However, the primary objective here extended beyond merely examining the integration of ELISE into the EU-VIS, encompassing a broader exploration of the software's evolutionary trajectory. This research phase was grounded in interviews conducted with WCC staff members. These interviews were designed to unveil insights into the development and deployment of the ELISE data matching system within the EU-VIS context, as well as pivotal historical events that have shaped the evolution of this identity-matching software system. The research further draws from field notes and document analysis, which collectively served to analyze the software's evolutions.

This part of the fieldwork harnessed the inversion strategy of tracing the sociotech-

nical change of this particular data matching software. To accomplish this, the research necessitated the creation of heuristics capable of discerning analytically significant moments during the software's development and deployment. Such moments would shed light on how the circulation of knowledge and technology for matching identity data unfolds and traverses across distinct organizational contexts. First, integrating the ELISE software into the EU-VIS emerged as a juncture facilitating connections among disparate systems, rendering it conceptually akin to a gateway technology. Additionally, the research uncovered the software's substantial transformations over time as it ventured into domains beyond migration and border control. Hence, an alternative heuristic was employed to dissect the shifting interpretative flexibility of the software, coupled with mechanisms that culminated in the present-day data matching solutions. Together, they demonstrate how knowledge and technologies for matching identity data travel and sociopolitical and economic forces that both shape and are shaped by transnational commercialized security infrastructures.

3.4 Methods for data collection

The methodological framework guided the collection and analysis of data matching in European security data infrastructures. Various techniques were used to gather and analyze qualitative data to support the investigation, including document analysis and user interviews. Naturally, these methods informed each other at various points throughout the research.

3.4.1 Data models

Regarding the collection of data models, the research draws on data collected during fieldwork conducted in the context of the Processing Citizenship project at border zones in Europe. In addition, given linguistic constraints and the PC project's task plan organized as a matrix, some documents were collected by other researchers employed as collaborators in the Processing Citizenship project.[5] Overall, the data collection efforts included desk research of European regulations over Eurodac, VIS, and SISII, technical documents made available by European and German authorities, systems screenshots collected at border zones in the Hellenic Republic, interviews and ethnographic observation with IT developers and users in Italy and Greece, and technical documents collected during fieldwork in The Netherlands.

The collection of documents on the analysis of data models encompassed the following:

Regulation (EU) No 603/2013, European Parliament and Council, 26 June 2013 The regulation governing the establishment of Eurodac includes Article 11, which

[5] Over the years, PC researchers have been: A. Bacchi, E. Frezouli, Y. Lausberg, C. Loschi, L. Olivieri, A. Pelizza, A. Pettrachin, S. Scheel, P. Trauttmansdorff.

addresses the "Recording of Data." This article outlines the specific data that must be recorded in the Central System.

Regulation (EU) No 2018/1862, European Parliament and Council, 28 Nov. 2018 The regulation pertaining to the establishment, operation, and utilization of the Schengen Information System (SIS) includes Article 20, which addresses "Categories of Data." This article outlines the specific categories of data provided by each Member State and contained within the SIS.

Regulation (EC) No 767/2008, European Parliament and Council, 9 July 2008 The regulation governing the Visa Information System (VIS) and data exchange among Member States concerning short-stay visas contains two pertinent articles. Article 5, titled "Categories of Data," delineates the specific categories of data that must be recorded in the VIS. Concurrently, Article 9, titled "Data upon Lodging the Application," elucidates the categories of personal data recorded of applicants when lodging their visa applications.

Hellenic Register of Foreigners screenshots Screenshots of the Hellenic Register of Foreigners (also known as ALKYONI system), used in the Hellenic Republic for managing asylum applications, were captured by fellow members of the Processing Citizenship project during their fieldwork. These screenshots offer a detailed view of the system's interface and functionalities, particularly in relation to the registration and identification of asylum seekers in Greece.

XAusländer Version 1.14.0 The XAusländer standard is a protocol based on XML (eXtensible Markup Language) that facilitates the electronic exchange of identity data among various authorities within Germany's immigration administration. This standard is accessible through the website of the "Koordinierungsstelle für IT-Standards" (Coordination Office for IT Standards). Section "2.2 Das Informationsmodell" (2.2 The Information Model) within the standard elaborates on the specifics of the information model for individuals.

3.4.2 Documents

During fieldwork, I gathered a range of documents pertinent to the ELISE data matching systems. The collection of documents comprised a diverse range of materials, including technical design specifications, product brochures outlining the software's features and advantages, sales presentations delivered to customers, and other materials. Access to these materials was facilitated through my contacts within the WCC ID Team, who provided me with access to these documents. Furthermore, when I visited WCC's headquarters in person, I received further clarification regarding the technical complexities I had previously encountered while reviewing those materials.

This collection of documents can be categorized into two main groups. Firstly, some documents pertained to the broader technical intricacies of the ELISE ID platform, enabling me to familiarize myself with the software suite and its design by WCC. Secondly, documents linked to the platform's specific implementations at the IND allowed me to

gain insights into how WCC practically configures the system for its customers. Notably, the collection of documents encompassed archival materials, such as past presentations and meeting minutes, that shed light on the system's deployment and configuration process at the IND. For instance, one set of meeting minutes provided detailed insights into discussions concerning the configuration of the matching criteria. Another noteworthy document was a presentation that covered an upgrade of the ELISE system and introduced new features implemented at that time.

Hence, the diverse range of materials underscores the multifaceted nature of documents and the documentation practices, highlighting that they serve a purpose beyond mere information recording, making them valuable topics and resources for research (e.g., Shankar et al., 2017). Documents exhibit an active role within social contexts, functioning to account for and coordinate various workplace activities. In this regard, documents are closely intertwined with the processes that engender them (Hull, 2012; Riles, 2006). Consequently, the research adopted a two-fold approach towards these documents. On the one hand, they were examined to gain insights into the technical workings and features of the software system and its configuration. On the other hand, the documents were harnessed to comprehend larger trends in the evolution and utilization of identification technologies.

The documents collected during the research process played a dual role in contributing to the investigation, aligning with both the first and second inversion strategies within the methodological framework. Firstly, these documents provided insights into data matching practices through the technical documents outlining the data matching mechanics of ELISE, shedding light on the designed functioning of the data matching system. This designed functioning was then juxtaposed with the real-world utilization of different users within the IND, as discerned from interviews, which will be elaborated upon in the subsequent section. Chapter 5 explores contrasts between the design of the matching system and its practical implementation, centered on three key aspects: the formulation of search queries, the computation of search results, and the interpretation of search outcomes. This comparative analysis ultimately revealed distinct frictions between the intended design and actual usage in data matching practices, thereby shedding light on impediments that may obstruct the IND's identification processes.

Secondly, the collected documents were used for the second inversion strategy to trace the sociotechnical evolution of the data matching software. In this context, a focus is on documenting practices, investigating alterations within document versions or the incorporation of supplementary annotations (see also, for example, Bowker and Star, 1999; Shankar et al., 2017; Sweeney, 2008). For example, the documents associated with the deployment of ELISE at the IND encompassed materials detailing package updates and records of discussions about configuration changes. Chapter 6 effectively utilizes the documents to delve into the specifics of this update, revealing new data matching features originated from diverse contexts, including an international name matching competition and a feature tailored for the EU-VIS. By analyzing these documents, the

chapter effectively highlights the process of disseminating and exchanging data matching knowledge among various organizations.

The collection of documents also brought about particular challenges and unexpected insights. Notably, the absence or inaccessibility of documents emerged as a significant but less evident aspect of the documentation process. During fieldwork, it became apparent that certain records related to the deployment of ELISE in the EU Visa Information System would be inaccessible. This lack of access resulted from the specific dynamics among actors involved in the development of the EU-VIS systems. In this arrangement, WCC was the technology supplier and collaborated with Accenture, the technology integrator. Consequently, the pertinent documents were under the purview of the EU agencies and the technology integrator. Moreover, even for WCC, the technology integrator effectively acted as a gatekeeper, hindering access to such documents. Despite not having access to the technical details of the integration of ELISE in the EU-VIS, this instance proved insightful in understanding the dynamics of WCC's role as a technology supplier and its dependencies on technology integrators and consultancies within the larger sociopolitical and economic contexts, as elucidated in greater detail in Chapter 6.

The accompanying table provides an overview of the documents that were consulted. The findings obtained from the documents played a role in shaping the structure and content of the interviews, as will be elucidated in the following section.

3.4.3 Interviews

The fieldwork encompassed a series of semi-structured interviews designed to align with the methodological framework's second and third inversion strategies. These interviews can be categorized into two participant groups, each serving distinct research objectives. The first group was dedicated to exploring data matching practices. This interview group included a diverse range of IND personnel, encompassing individuals engaged in technical development associated with INDiGO and those responsible for identity lookup and matching tasks using the ELISE search and match engine as an integral part of their tasks. The second group of interviews was centered around unraveling the sociotechnical evolution of the data matching software, involving interviews with WCC employees responsible for various aspects of the company's identity matching system, including design, implementation, and (pre-)sales efforts.

The interview protocol for IND staff was designed to delve into specific aspects of searching and matching applicant data within the IND organization. Interviews commenced with establishing interviewees' roles to contextualize their experiences and tailor subsequent questions. Three primary factors influencing data searching and matching were addressed: query formulation, match computation, and result handling. For query formulation, questions delved into categories used, preferred data elements, and wildcard utilization. The second factor explored match result expectations and engine understanding. The third factor probed result processing, encompassing quality perception, match ranking, and score interpretation. Additionally, duplicates and dedupli-

Table 3.2: Summary of main documents consulted.

Document	Author	Year	Description
Functioneel ontwerp systeemfunctiebeschrijvingen	IND	2009	The functional design documents specifying the search functions for matching in INDiGO and searching the BVV system.
Analyses and technical documents for ELISE implementation in INDiGO	WCC	2009	Technical documents detailing the analyses, functionalities, deployment and maintenance of ELISE in the INDiGO system.
Verslag en memos INDiGO ELISE zoekmodel optimalisatie	WCC & IND	2010	Reports and memos on the findings and decisions taken between WCC and the IND regarding updates to the way match scores are calculated.
WCC whitepapers and marketing material	WCC	Undated	These documents include marketing or sales brochures by WCC. They were useful to compare the features of ELISE highlighted by WCC with the knowledge of these features by actual users.
Protocol Identificatie en Labeling	Directorate migratieketen	2020	This document describes the standardized method for identification and registration in the migration chain and the different roles of the different parties involved. In the annex it reports in some detail the databases and search interfaces.

cation criteria were discussed. Further details about the interview protocol and analysis of responses are provided in Chapter 5.

The interview approach was guided by the hypothesis that differences in data matching practices could be best discerned by contrasting how personnel from various IND departments employed these tools for searching and matching in their distinct identification tasks. This hypothesis emerged from initial insights from document analysis and discussions with the WCC ID Team. Consequently, efforts were made to recruit interview participants from different organizational units within the IND (a list of these units is available in Table 3.3). While it was not feasible to interview users from every unit, insights into the utilization of search and match tools in various departments did emerge during the interviews. This insight was made possible by particular participants with extensive experience, having worked across different organizational units within the IND or in roles involving collaboration with multiple departments. Five interviews with IND staff were conducted, each lasting about an hour, where participants detailed their experiences with identifying applicants and utilizing search and match tools.[6] This approach yielded valuable results, and Chapter 5 leverages the diverse data matching practices revealed in these interviews to develop an analytical tool for interpreting these practices.

The second group of interviews, conducted concurrently with and following the IND interviews, centered on delving into the sociotechnical evolution of the ELISE data matching software to unveil broader shifts within the realm of identification. These interviews primarily involved discussions with staff members from WCC with experience within the identity domain. The interviews were designed to uncover significant milestones in the history of their data matching software system. Given the diversity in participants' roles, projects, and experiences within the company, the interviews followed a more flexible format than those conducted with IND employees, allowing for tailored discussions based on each individual's background and expertise. For instance, when interviewing individuals involved in projects like the EU-VIS, the questions were honed to delve deeper into these areas. In cases where less information was available about the interviewee, the conversation typically began by exploring their connections with current and potential customers in the security and identity market, gradually steering the discussion towards sociotechnical changes. Seven interviews were conducted in this group, each spanning approximately an hour.[7] The participants

[6] Repeated efforts were made to expand the sample size of participants from the IND for interviews. Initial attempts began through the WCC's primary contact with the IND, who provided contact details for several individuals. However, only a subset of these individuals agreed to participate. Subsequently, efforts were made to leverage these participants to connect with their colleagues, aiming for a respondent-driven sampling approach. However, the process encountered challenges and delays, compounded by the online nature of communication and the necessity of remote calls. Over several months, these endeavors ultimately proved more complex than anticipated. Additionally, the need to produce a report for WCC within a specific timeframe further constrained the possibilities for expanding the participant pool. Despite these challenges, the sample size provided valuable insights and rich details for analysis.

[7] Considering the relatively modest size of the company and the ID Team, the sample size, to the best of my knowledge, comprehensively encompassed most individuals with relevant backgrounds or experiences related

Table 3.3: Summary of IND units and interview participation.

Unit	Included	Description
Documentaire Informatievoorziening (DIV)	Yes	Management, archiving, registration and making available of documents within the organisation.
Directie Regie Vreemdelingenketen (DRV)	Yes	This unit is responsible for the registration and startup of procedures. These applications can arrive via paper mail or though digital platform.
Titels en Identiteit (T&I)	Yes	T&I handles among others the sharing of information about the right of residence of aliens to partners in the migrationchain, correcting titles, checking automatic messaging between the partners, processing deduplication of clients in the IND databases.
Keteninformatie & Controle (K&C)	No	Collects information from various government sources for the various IND processes.
Handhavingsinfor (HIK)	No	Collects and registers signals of fraud, abuse and human trafficking and smuggling.
Beslisteam	No	Decides on the residence application of the foreign national. They will consult information in the INDiGO system to make an assessment individual circumstances and draw up a binding opinion for the decision.

comprised individuals in various roles, including consultancy, pre-sales, solutions management, software development, and user experience design.

Participants from both interview groups consented to record the sessions using Processing Citizenship's informed consent form. The form allowed participants to specify how the research would use their provided data while guaranteeing anonymity and confidentiality. The recording procedure followed a protocol to enhance anonymity, including only audio recordings with a distorted voice. Additionally, manual transcription of interviews ensured additional confidentiality by preventing confidential information from being leaked via automated transcription platforms. Table 3.4 provides an overview of the interviews.

The number of interviews conducted in this study was lower than initially expected and fell short of the number initially foreseen by the Processing Citizenship project. Several factors contributed to this outcome. Firstly, due to the sensitive nature of their border control and security work, it was challenging to gain access to additional customers beyond the software technology supplier and one of their customers, despite the initial intention to include them. Clearance and background checks required for individuals involved in these areas posed difficulties for the researcher in expanding the participant pool. Additionally, the COVID-19 pandemic further complicated the situation by limiting networking opportunities and hindering the ability to find additional interview participants, particularly in the case of IND interviews. As a result, the study had to adapt to conducting online interviews, which, although enabling data collection to continue, introduced constraints such as reduced rapport-building and limitations in gathering nuanced insights compared to face-to-face interactions.[8] The study made the most of available opportunities despite these challenges; it provided valuable insights within its defined scope, shedding light on the perspectives of the software technology supplier and their customer.

3.4.4 Events

The operationalization of the second inversion strategy, which involves examining sociotechnical change within the realm of identification technologies, also encompasses data collection through participation in relevant events. As recognized in Social Construction of Technology, the development of technologies involves a wide array of relevant social groups. These social groups can extend beyond developers and users to include researchers, journalists, politicians, civil organizations, and other stakeholders. Therefore, attending events such as industry conferences, academic symposiums, and

to the field or the company and its products.

[8] In the context of the IND interviews, the research could have benefitted from observing the practical use of the search and match tools. During these interviews, the research aimed to utilize the WCC company's secure Microsoft Teams installation for online meetings, which offered the possibility of screen sharing for such observations. However, constraints emerged due to installing and using the Microsoft Teams application on their company-issued laptops, which ultimately necessitated conducting phone interviews.

Table 3.4: Summary of interviews.

Date	Org	Channel	Language	Description
2020, July 30	WCC	MS Teams	English	ID solutions manager.
2020, Aug 05	IND	Phone	Dutch	Involved in technical projects at IND.
2020, Nov 02	IND	MS Teams	Dutch	Senior employee at IND. Previously worked in team responsible for deduplication.
2020, Nov 10	IND	MS Teams	Dutch	Senior employee at the unit responsible for registering foreigners for the first time and starting their procedures (DRV).
2021, Jan 29	IND	Phone	Dutch	Senior employee at IND at Titels & Identiteit.
2021, May 21	WCC	MS Teams	Dutch	Recent employee at WCC, but whose has a background in identity matching in financial sector.
2021, May 28	WCC	MS Teams	Dutch	Senior Software Developer/team lead.
2021, May 31	WCC	MS Teams	Dutch	VP Identity Solutions.
2021, June 03	WCC	MS Teams	Dutch	Senior UX Designer.
2021, July 01	WCC	MS Teams	English	Pre-sales.

other gatherings involving diverse stakeholders—such as industry representatives, academics, Member State authorities, and EU agencies—helped gain additional insights into the sociotechnical evolution of identification technologies. These attended events were characterized by discussions revolving around related topics such as data matching, data quality, identification, and data interoperability within EU identification systems. Additionally, some of these events were directly related to the fieldwork, involving the participation of WCC. A summary and description of these events, whether attended in person or online, can be found in Table 3.5.

Participating in such events was perceived as a means to explore how diverse social groups define challenges concerning developing identification technologies within border and migration control. The gatherings were seen as opportunities to unveil how varying interpretations of these challenges may lead to conflicts and the differing interpretations of how these conflicts can be resolved through technological means. To illustrate this, consider an observation made during a 2018 eu-LISA industry conference. During the Q&A session following a presentation on technical solutions for achieving interoperability among EU systems, an audience member raised a pertinent concern about handling the potentially significant volume of false positives resulting from integrating disparate databases and matching data containing outdated information. In response, one of the panelists acknowledged the need to address this issue through a one-off intensive effort to resolve these challenges. This example illustrates differing interpretations of the problem and technical solutions; in this case, it questions the perceived smooth integration of databases and highlights the costs, such as the labor-intensive manual efforts required to integrate diverse identification systems.

Moreover, these events were seen as opportunities to not only observe how various actors and social groups define and interpret problems and solutions, but also to observe how these definitions and interpretations are disseminated and circulated. For example, examples given by professionals may involve data quality and matching concerns related to security and counter-terrorism, which could illuminate the securitization process within data matching and identification practices. A particularly insightful illustration emerged during one of my online sessions, which I integrated into the opening of Chapter 5. The presented case underscores how security companies frame the importance of technologies for searching and matching data stored in diverse databases, especially within counter-terrorism initiatives. The company recounted the instance of authorities adding one of the "Boston bombers" to police watch list databases before the attack but with inconsistent and invalid transliterations of his name. The professional posited that discrepant information across databases poses hurdles for authorities' investigations, which could result in potential blind spots for authorities (the problem), which could be solved by utilizing their data matching technology. Subsequently, I discovered this example was recurrently cited across various companies in similar contexts. In this manner, these events also offer a glimpse into the circulation and transfer of the problems and solutions of data matching knowledge across different organizations.

Table 3.5: Summary of attended events.

Date and location	Event name	Event description
15 May 2018, Brussels, Belgium	European Migration Network (EMN) 10 Year Anniversary: Understanding Migration in the EU: past, present, future	The 10 year anniversary of EMN was celebrated in this event and provided an opportunity to understand how a diverse range of European and international organisations understand the past, present and future of migration in the EU.
15 June 2018, Oxford, United Kingdom	Deconstructing Biometric Refugee Registration	Different perspectives on biometric registration of refugees and the registration systems used, such as by UNHCR.
17 October 2018, Tallinn, Estonia	eu-LISA annual conference 2018: EU Borders —Getting Smarter Through Technology	Presentations from European Commission Directorate General and Agencies on the IT architecture for interoperability and issues of data quality.
18 October 2018, Tallinn, Estonia	eu-LISA industry roundtable 2018: Technologies to Facilitate Land Border Crossing	Presentations from industry on technologies for the Entry-Exit System.
31 January 2019, Brussels, Belgium	CPDP 2019: Data Protection and Democracy	Round tables with practioners, Members of the European Parliament, and EU agencies on effects of data quality and checks and balances in the Justic and Home Affairs.
01 May 2020, Online	Festival of Identity Webinar: How to Make Law Enforcement Systems Interoperable and Available on Any Device	Webinar by WCC on how to link identities across systems that have different data structures.
26 October 2020, Online	eu-LISA annual conference 2020: Interoperability - Building Digital Resilience for the EU Justice and Home Affairs Community	Panels of representatives of EU Institutions and Agencies and practioners from Members States discussing the shift to interoperability.
03 November 2021, Online	eu-LISA industry roundtable 2020	Partners from industry such as WCC presented how their solutions might contribute to the developments of the infrastructure of the JHA domain.
27 October 2021, Online	eu-LISA annual conference 2021: Towards the Digital Schengen Area	The move in EU towards centralised and standardised information exchange in the domain of Justice and Home Affairs.

3.4.5 Other data

As previously mentioned, the sociotechnical evolution of technologies encompasses various pertinent social groups. To augment the analysis, supplementary data were obtained from other publications, such as news articles and press releases. These materials offered additional viewpoints, facilitating a comprehensive exploration of the historical progression and transformations concerning the WCC search and match software. The Nexis Uni[9] database and tool were used to search for articles that mentioned the company WCC and its software in the INDIGO system of the IND or the EU-VIS system. Nexis Uni allows users to search for articles in newspapers, online business publications, and other sources written in English and Dutch. For example, several articles published in Dutch IT business news websites provided information on IND's INDiGO system. Moreover, this database surfaced press releases pertaining to WCC's participation in the MITRE multicultural name matching challenge, featured in the analysis of Chapter 6. Beyond this, Nexis Uni enabled the discovery of newspaper articles, including a 2011 piece from "Het Financieele Dagblad," a prominent Dutch newspaper focused on economy and business. The article, titled "Utrechtse datatechnologie moet terroristen buiten de VS houden" (Data technology from Utrecht must keep terrorists out of the US), contributed to tracing the company's evolutionary trajectory within the realm of security.

Furthermore, additional publications were found via conventional search engines. For example, blog posts on WCC's and competitors' websites showed how experts in the field of name-matching devise and disseminate their knowledge. Consider, for instance, the blog post titled "Understanding Dari and Pashto names: a challenge to intelligence gathering in Afghanistan" (Basis Technology, 2012). This post originates from a WCC competitor engaged in developing a comparable data matching system. Within this blog post, a linguist elucidates the intricate technical hurdles associated with "Afghani names and how they challenge these software tools." Here, one can observe the practical application of linguistic expertise in security intelligence, exemplified by the assertion that "Afghani names pose a challenge to intelligence agencies." Another example is a WCC blog post titled "Biographic matching & UMF standards for EU interoperability" (Scheers, 2021). This specific post sheds light on a noteworthy dynamic, illustrating how WCC's software is being reinterpreted to align with a project's requirements set by the EU. The blog post delineates how the demands of the EU project are mapped onto the capabilities of WCC's solution, showcasing changes in design and interpretative flexibility.

[9] http://web.archive.org/web/20230606140543/https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page (Formerly LexisNexis Academic)

3.5 Techniques of data analysis

The data analysis in this dissertation employs various techniques that span qualitative and computational methods. In Chapter 4, a distinctive methodology is devised to compare data models across different authorities. This approach is utilized to analyze data models from information systems focused on population management. Moving to Chapter 5, a mixed deductive-inductive approach is adopted to analyze data matching practices derived from the interviews conducted at the IND. Lastly, Chapter 6 takes a slightly different approach by relying less on developing codes for theory development. Instead, the analysis involves writing integrative memos, which serve as a means to elaborate ideas and interconnect various pieces of data.

The data analysis for the first inversion strategy involved the development of a new method to compare data models collected in various formats and used by diverse systems, as no existing method fully met our requirements. Chapter 4 provides a comprehensive methodology and introduces a software tool called "The Ontology Explorer," designed to facilitate the comparison of data models collected in different formats and from various systems in two primary ways. Firstly, it supports analyses of information systems that define their data models, even if these systems are only occasionally comparable. Secondly, it systematically and quantitatively enables discursive analysis of "thin" data models by identifying differences and gaps between systems. The methodology involves extracting, analyzing, comparing, and visualizing heterogeneous data models. This structured approach was applied to the data models employed by diverse organizations, ultimately enabling the observation of differences and similarities within the data models of various data infrastructures. The data models used by the EU and Member States were analyzed using the Ontology Explorer methodology and tool. The results showed discrepancies and commonalities in the collected data, shedding light on the circulation of knowledge about individuals and the division of labor among different involved actors.(Pelizza and Van Rossem, 2023)

The data analysis for the second inversion strategy, focusing on data matching practices, involved utilizing the computer-assisted qualitative data analysis software ATLAS.ti for coding and analyzing the data. This data coding and analysis process was guided by the "Noticing-Collecting-Thinking" (NCT) method by Friese (2014), tailored to the ATLAS.ti software and comprises three interconnected steps. In the "Noticing" step, segments within the documents were labeled with codes. This phase encompassed both deductive codes stemming from research hypotheses regarding crucial aspects of applicant identification through the search and match tools and openness to inductive insights from the data. The deductive codes were primarily built around the three factors of search: query formulation, match computation, and manipulation of search results. Concurrently, new codes emerged based on the data's inductive findings, such as the processes of dealing with duplicate records. In the "Collecting" step, these codes were reviewed and grouped into similar categories. Subsequently, in the "Thinking"

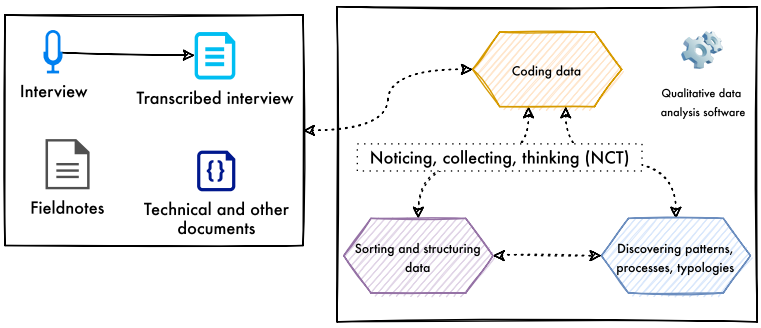


Figure 3.2: Schematic representation of the data analysis process for the second inversion strategy.

step, patterns, processes, and typologies were identified among the developed codes. Throughout the process, there was an iterative movement between the noticing, collecting, and thinking steps, enhancing the depth and richness of the analysis. Figure 3.2 visually represents this recursive data collection process and the application of the NCT steps. Chapter 5 offers practical examples of applying this methodology, including illustrative examples that led to the development of the interpretative framework for data matching practices.

The data analysis undertaken for the third inversion strategy, which focused on understanding sociotechnical change, involved writing integrative memos to elucidate concepts and connect fragments of data. Preliminary analyses and memos were already embedded within the interview transcripts and field notes. These fragments of analysis represent the researcher's initial development of ideas based on the data collected. Integrative memos amalgamate diverse data analysis components from interviews, field notes, or excerpts from external materials, such as news articles. The aim was to write these memos with the anticipation of future readers in mind, particularly those who may not be familiar with the research setting. As Emerson et al. (2011) highlights, "integrative memos provide a first occasion to begin to explicate contextual and background information that a reader who is unfamiliar with the setting would need to know in order to follow the key ideas and claims" (p. 193). For example, I mapped various actors and their involvements over different time points to track the sociotechnical evolution of the data matching software. Instead of relying solely on data coding, I found it more effective to construct narratives that inherently encompassed the required contextual details. These integrative memos played a pivotal role in advancing two key aspects. Firstly, they contributed to refining our understanding regarding the social groups involved and the interpretative flexibility of the software. Secondly, these memos aided in conceptualizing the software as a gateway technology that bridges connections among diverse actors and systems. This approach facilitated the organization of pertinent information into cohesive narratives, proving instrumental in meticulously tracing the complex sociotechnical

evolution of the data matching software.

Chapter 4

The Ontology Explorer: A method to make visible data infrastructures for population management

The chapter has been co-authored in collaboration with the Principal Investigator (PI) of Processing Citizenship and was originally published in 2022 as an article in Big Data and Society.[1] *The text keeps the original form as presented in the article. Because of its earlier publication, the original article has some missing links to the overall structure of the dissertation, when compared to the other work presented in this dissertation. To enhance the understanding of the chapter's relevance to the research goals, objectives, and methodological framework of the dissertation, additional information is included in this chapter, following the abstract.*

Abstract

This chapter introduces the methodology of the "Ontology Explorer" (OE), a semantic method and JavaScript-based open-source tool to analyze data models underpinning information systems. The OE has been devised and developed together with Annalisa Pelizza, who recognized a need to compare data models collected in different formats and used by diverse systems and diverse authorities at national and European level. The OE is distinctive firstly because it supports analyses of information systems that are not immediately comparable and, secondly, because it systematically and quantitatively supports discursive analysis of "thin" data models — also by detecting differences and absences through comparison. When applied to data models underpinning systems for population management, the OE enables the apprehension of how people are "inscribed" in information systems: which assumptions are made about them, and which possibilities are excluded by design. The OE thus constitutes a methodology to capture authorities'

[1] Van Rossem, Wouter, and Annalisa Pelizza. 2022. "The Ontology Explorer: A Method to Make Visible Data Infrastructures for Population Management". *Big Data & Society* 9 (1): 1–18. https://doi.org/10.1177/20539517221104087.

own imaginaries of populations and the "scripts" through which they enact actual people. Furthermore, the method allows the comparison of scripts from diverse authorities. This is exemplified by illustrating its functioning with information systems for population management deployed at the European border. The approach integrates a number of insights from early infrastructure studies and extends their methods and analytical depth to account for contemporary data infrastructures. By doing so, it triggers a systematic discussion on how to extend those early methodical innovations at the semantic level to contemporary developments in digital methods.

Contribution to research objectives *To introduce a new method and software tool to analyze the data schemas underpinning information systems in population management.*

This chapter contributes to the dissertation's research objective of analyzing data models underpinning information systems in population management. The chapter fulfills this objective by introducing the "Ontology Explorer" (OE), a methodology and JavaScript-based open-source tool for systematically examining and comparing non-homogeneous data models used in diverse information systems. Two key inspirations drove the development of the OE methodology. Firstly, it draws from schema matching techniques, which underscore the importance of finding correspondences between data models for data matching and integration across disparate data sources. Knowledge of the underlying data models is crucial, as, for example, records referring to the same person may be challenging to identify if represented differently across different datasets. Secondly, the OE is influenced by earlier STS research that aimed to uncover the less visible role of classifications and standards in shaping and segmenting the world. The OE method and tool thus contribute to the dissertation's research objective, enabling a deeper understanding of the assumptions inscribed in data models and their implications for population management and identification.

Infrastructural inversion strategy *First inversion strategy — Comparing data models*

This chapter employs the infrastructural inversion strategy of comparing data models to uncover the underlying design choices embedded within data infrastructures. Although integral to enacting populations, territories, and borders, these data models tend to become invisible over time. Unfortunately, data models offer limited information as their meaning often becomes invisible. The OE is a crucial tool to recover these meanings by relying on comparison for meaningful interpretation, thereby enabling systematic and quantitative comparisons of data models. By employing the OE, the less visible choices made within technical standards can be made visible, providing a deeper understanding of how populations are enacted across systems and infrastructures. The comparative approach

facilitated by the OE methodology helps invert imaginaries and expectations embedded within infrastructures' data models.

Contribution to research questions *RQ1: Which types of knowledge and assumptions about people-on-the-move are inscribed in data models of national and transnational security infrastructures? What implications does this have for how organizations can search and match identity data?*

This chapter addresses research question 1 by providing an illustrative use case for analyzing data models used in information systems aimed at identifying and registering migrants arriving at the European border. The OE allows for the systematic comparison and analysis of disparate data formats, making it possible to bring out implicit knowledge and assumptions hidden within these data models. As such, the analysis provides valuable insights into authorities' expectations and imaginaries about populations and allows for comparisons between different authorities. The findings reveal both constants and differences in authorities' assumptions about migrants, highlighting the conceptualization of identity data among authorities involved in population management and the exclusion of certain possibilities through design. Furthermore, by measuring the centrality of categories of data, the analysis highlights the importance of certain categories, such as "date of birth" and "nationality," in connecting identities registered in multiple systems. The chapter's illustrative analysis offers insights into how authorities conceptualize identity data in the context of population management and the possible implications for searching and matching such data.

Contribution to the main research question *How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?*

The chapter indirectly addresses the main research question, which asks how practices and technologies for matching identity data in migration management and border control shape and are shaped by transnational commercialized security infrastructures. The chapter unveils less visible assumptions and patterns embedded in national and European information systems by applying the OE to data models within the context of migration management and border control. It reveals how identity data properties influence the creation and circulation of data and the relationships between different authorities' data models. This understanding helps elucidate the reciprocal relationship between various authorities' data models and the links between identity data in data infrastructures for population management.

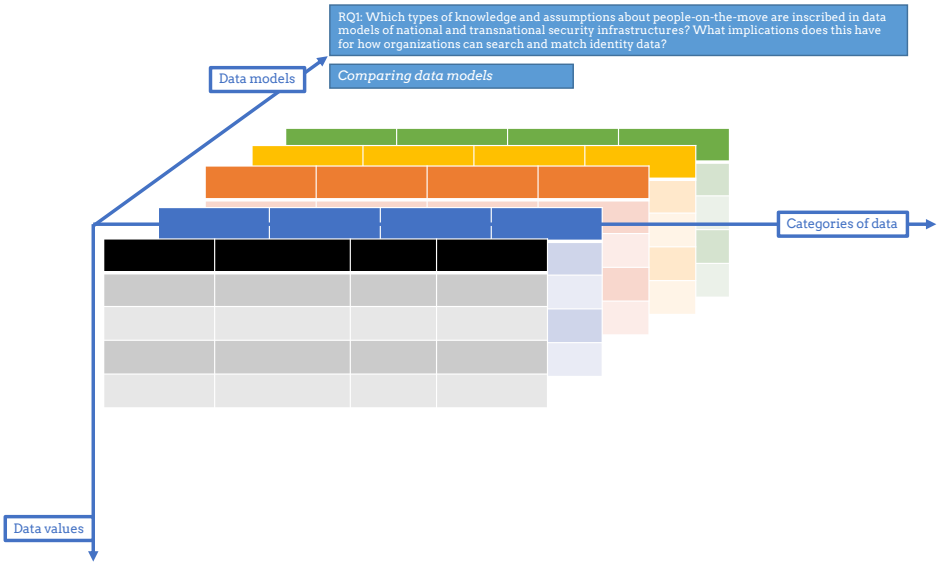


Figure 4.1: The axes pertaining to the methodological framework in relation to chapter 4.

4.1 Introduction

This chapter details the methodology of the "Ontology Explorer" (OE), a method and a tool devised and developed by the authors[2] to analyze data models which underpin information systems.[3] The OE is a semantic method and JavaScript-based open-source tool that compares data models collected in different formats and used by diverse systems. The OE is distinctive in two respects. Firstly, it supports analyses of information systems that define their own data models, even if these systems are not immediately comparable. Secondly, it systematically and quantitatively supports discursive analysis of "thin" data models, also by detecting differences and absences between systems. To achieve this, the method extracts, analyzes, compares and visualizes heterogeneous data models all at once. Applying this structured approach to the data models used by diverse organizations will make it possible to observe differences and similarities in the data models

[2] In collaboration with the Principal Investigator (PI) of Processing Citizenship.

[3] In this chapter the term "ontology" is used to refer to an abstract set of concepts, categories, and relations. Drawing on computer science's jargon, the term "data model" is used to refer to the material instances of those ontologies which are embedded in information systems. The distinction between the two terms is similar to the difference made by Ferdinand de Saussure between *langue* and *parole*, and his ideas on identifying the structures of language as an abstract system (langue) by studying the actual use of language (parole). In a similar vein, we can study ontologies of population management (langue) by looking at actual, materialized data models (parole).

of diverse data infrastructures.[4]

The OE integrates a number of insights from early infrastructure studies and extends their methods and analytical depth to account for contemporary data infrastructures. For many years, infrastructure studies have paid close analytical attention to classifications as an important force in shaping and ordering sociotechnical relations, as well as how categories may become contested or invisible. At the same time, these studies have noted how classifications are embedded into infrastructures, where they have a tendency to become taken for granted. More recently, data-oriented activists are well aware of the need to contest and intervene into the politics of data (Beraldo and Milan, 2019; Iliadis, 2018). Yet, to challenge prescriptive representations they first need to be analyzed using methods that can systematically do so.

Compared with existing digital methods, the OE features some unique characteristics that make such analyses possible. The method bears a resemblance to approaches that, for example, employ web scraping and correlation techniques to examine traces of data to find structures and patterns that may otherwise be hard to discern. However, the OE is distinctive in its attention to the technical details of systems (rather than the generated data) and in its ability to make non-homogeneous data formats comparable. In this manner, the OE supports discursive analysis, including analysis of differences and absences, and makes it possible to compare the expectations and imaginaries of diverse social actors, and forms of resistance to them.

Our methodological proposal aims to answer a broad set of research questions about how populations are enacted, and how authorities and countries differ in the way they enact them. In this way, the OE constitutes a methodology to capture authorities' own imaginaries of "intended" people, and the "scripts" through which authorities enact actual people as such (Pelizza and Van Rossem, 2023). It thus enables the apprehension of how people are "inscribed" in information systems: which assumptions are made about them, and which possibilities are excluded by design. For instance, analyzing population data collected by social security agencies can show in-built assumptions in the systems that automate decisions about receiving benefits (Eubanks, 2018). Another example is the COVID-19 pandemic and its attendant rapid changes in data practices, which have highlighted a need to analyze categories of data on population and health, and how presences and absences of data can cause uneven impacts of the virus on people (Leszczynski and Zook, 2020). In this chapter, an illustrative use of the method to analyze information systems deployed at the European borders shows that the method allows the comparison of various EU and Member State authorities' scripts about mobile populations.

The chapter is structured in the following order: in the following section we review

[4] To provide an example, such comparison can support the analysis of information systems run by diverse national and transnational institutions to manage mobile populations. This is indeed one of the goals of the "Processing Citizenship" research programme funded by the European Research Council (ERC), PI prof. A. Pelizza (2017-2023). This chapter is written as part of this programme and presents the methodological solutions devised to answer some of its research questions. While the proposed methodology is already utilized by researchers in the programme, it is also released in the public domain (Van Rossem, 2021).

the methodological approaches that have inspired our work; next, we discuss the design principles of the OE as a methodology to be, in the first instance, deployed in the field of population management. Then, after a brief discussion of the methods for data model collection, we describe the OE method and tool for data models analysis, followed by an illustration of how it functions in the real world by analyzing the data models underpinning selected information systems for border control management. Finally, these illustrative uses help us to reflect on the OE's contribution to a discussion on how to extend infrastructure studies' early methodical innovations at the semantic level to contemporary developments in digital methods for population management.

4.2 Review of methodological approaches

Critical studies on classifications have made it clear that information systems' categories shape and benefit forms of knowing, valuing and doing (Bowker and Star, 1999). This capability goes hand in hand with the ability of categories to make other forms undocumentable and, consequently, invisible. Data models embodied in digital information systems are, moreover, embedded in complex arrangements of – and relations between – actors. As a result, formalized representations of knowledge tend to become invisible even to actors who interact with the systems. Unsurprisingly, then, invisibility and transparency of data models and their information systems have inspired various strands of research that propose methods to make visible the politics of designing such models and systems, and their world-making effects when deployed (Kitchin and Dodge, 2011). It must be noted that, while this section reviews some of these related strands of research, it is not the intention to comprehensively review all such methodological approaches. Rather, to structure the discussion, we grouped similar strands together according to similarities in their approach and the characteristics that inspired our own method.

4.2.1 Classification and its consequences

Our method and tool draw on studies about classification that were popular in the late 1990s, epitomized by the well-known study by Bowker and Star (1999) about the processes of creating classification systems and their effects on social organization. The use and discussion of this early work is still wide-spread, which highlights its enduring relevance and topicality. Scholars in this field have shown how the politics of designing systems that standardize and classify parts of the world (people, objects, work, etc.) into discrete categories not only shape how information is represented but also affect what becomes both visible and, consequently, invisible (Bowker and Star, 1999; Hanseth and Monteiro, 1997; Suchman, 1993). Furthermore, most options that are objects of choices made during design processes easily become invisible, as only what is in the systems is usually (and at best) documented. It is thus not surprising that these studies have mainly used discursive methods, drawing on interviews, ethnographic observations, historical and

narrative accounts, textual analysis, with the added aim of retrieving any missing data (Clarke, 2005). In short, data absence has become as meaningful as data presence, and this is a crucial key insight that the OE has inherited from classification studies of infrastructures.

Another insight further developed by the OE concerns the design and standardization that are essential to information systems that link actors with their imprints (Hanseth and Monteiro, 1997). People, objects, relations, and actions must all fit into the categories and categorical values foreseen by information systems. In this way, categories used by states to classify their residents, for example, consequently shape relations between states and individuals, such as facilitating or restricting access to benefits (Ruppert, 2014). These influences are made durable and thus can have long-lasting effects. Specialists involved in the development of systems are, therefore, not only designing technical systems but making (im)possible certain actions, practices, and relations (Suchman, 1993).[5]

Overall, methods developed in this tradition have addressed the question of how to retrace the ethical and political work of otherwise unremarkable devices of representation. Researchers have used and combined various makeshift methods – from narrative interviews (Gazan, 2005) to discursive textual analysis (Caswell, 2012), and from participant observation (Meershoek et al., 2011) to archival and genealogical research (Gasson, 2006) – for the ethnographic and historical studies of information systems (Star and Ruhleder, 1996) and their classifications. It is necessary to take an interdisciplinary approach and pay close attention to the technical details of classification alongside the effects of classification (Kitchin and Dodge, 2011).

4.2.2 Representing and intervening

The rise of Big Data (and its effects on the volume, variety, and uses of data) has prompted research fields such as Critical Data Studies (CDS) to renew calls to make the ways in which data are generated, circulated and deployed more visible (Iliadis and Russo, 2016; Kitchin and Lauriault, 2018). While researchers investigating classifications and infrastructures mainly take a relational approach to study infrastructures and social order, CDS researchers prefer different approaches to examine power structures of data. They emphasize how data are never "raw" but always shaped by different choices and constraints (Gitelman, 2013). The two strands of research are thus in many ways complementary. Even so, new and important questions have been raised by CDS on how the links between different mechanisms and elements that constitute data shape power relations. For example, which aspects of everyday experience are datafied? Which ones are made invisible? And, who benefits?

[5] Since the Social Construction of Technology (SCOT) approach to more recent user studies (Hyysalo et al., 2016), we are well aware the interaction between designers and users plays a key role in the final design. This aspect is indeed addressed in another article under review. The current chapter aims to propose a method to analyze the final outcomes of such interaction.

Understanding the ways in which data are constituted, therefore, becomes especially important for critical approaches that aim to intervene on the "data politics" (Ruppert et al., 2017) of how subjectivities come into being. For example, analyses of information systems and their data models and standards in other fields have shown how such systems can enact young people as young offenders through various interconnected data strands (Ruppert, 2013), how different systems and standards enact assumptions about what constitutes a family in distinct ways (Cornford et al., 2013), or how data models translating a city into data can reshape the city (Lauriault, 2017). The goal for these researchers is, to a greater extent, normative to stamp out inequality and injustice in data systems: to examine unequal power relations and ask who benefits and, next, challenge these power relations using data (D'Ignazio and Klein, 2020). However, this field emerged "as a loose knit group of frameworks, proposals, questions, and manifestos" (Iliadis and Russo, 2016, p.3) and is still settling on its theories and methods. Systematizing methods to examine data models can therefore serve as an important step to enable interventions into data politics.

4.2.3 Distributed systems and infrastructures

Information systems are entwined in digital infrastructures and practices, instituting certain ways of knowing and working among distributed locales (see, for example, Bowker and Star, 1999). Data models designed to represent phenomena can thus be considered as a technology that enact particular kinds of knowledge, organizations and practices at different times and places (Bloomfield and Vurdubakis, 1997). In other words, data models underpinning information systems can be understood as part of infrastructures that connect and coordinate geographically distributed actors and practices.

Scholars studying infrastructures have provided different methodological strategies to make visible the distributed interconnections between technical minutiae and the politics of knowledge production (e.g. Bowker and Star, 1999; Edwards et al., 2009; Monteiro et al., 2013). Investigating distributed infrastructures has primarily been addressed by approaching infrastructures as a "relational concept" (Star and Ruhleder, 1996). Such relational approaches maintain that infrastructures are not objects of study in and of themselves, but show a preference for ethnographic methods to study how infrastructures emerge through interactive processes and practices (Karasti and Blomberg, 2018; Star, 1999). Methods include looking at moments of breakdown (Latour, 2005; Star, 1999), tensions in the emergence and growth of infrastructure (Hanseth et al., 2006), and material aspects (Ribes, 2019).

Debates still circulate on where to situate ethnographic and historical investigations of distributed infrastructures. Opinions differ on how to demarcate the site of where and how to investigate dispersed and distributed activities (Karasti and Blomberg, 2018): at a single site? Or, by comparing and connecting multiple sites? And on which (temporal) scale? An alternative approach to deal with these issues of scale is to take advantage of in-

formation that is produced by systems and members during their activities. Geiger and Ribes (2011), for instance, proposed to make use of various documentary traces, such as computer logging data that capture interactions between users and systems, to understand how distributed communities collaborate and coordinate. Their ethnographic approach turns these "thin documentary traces" into "thick descriptions" of communities and their activities. These approaches, however, do not specifically take into account the actual formats of data and the possibilities and constraints they entail: this is an issue that our method should instead consider.

Similarly to concerns raised by Geiger and Ribes (2011), data models can be considered "thin" traces: being rather standardized schemas made of categories and values, they are barely meaningful in themselves (Pelizza and Van Rossem, 2023). An analysis method should therefore consider how to turn these into "thick" data. On the other hand, given their implementation in diverse infrastructures, data models can be an excellent starting point to understand how geographically distributed sites are connected (Latour, 2005). Burns and Wark (2020) dubbed such an approach "database ethnography" and used traces left behind from a database as a site to analyze how social meanings of phenomena change over time – but, only using their own ad-hoc mix of methods.

So, while data models such as those implemented in databases are not always openly available for analysis, the use of documentary evidence can serve as a way to follow distributed phenomena. However, we should avoid an overly deterministic view on these abilities to enact particular kinds of knowledge (Hanseth et al., 2006). The strength to institute ways of knowing and acting depends also on data models' capacity to adapt to local circumstances and forms of resistance.

4.2.4 Script analyses

The prescriptive representations of data models can thus be understood as scripts that inscribe and enact certain expectations of their designers. The notion of a script was originally introduced in social studies of technology to refer to the implicit instructions to users or affordances that are embedded in artefacts (Akrich, 1992; Latour, 1992). A script requires users to adopt standard recommended behaviors, like grasping a coffee pot by the handle, and so assumes certain user skills, conditions and interests – in our example, the user does not wish to burn their hand when grasping the pot. To some extent, data models can be conceived as scripts, as they require system users to adopt well-defined behaviors to sort and select phenomena into standardized categorizations. As such, studying data models that inscribe and enact expectations in the materiality of information systems can provide insights about designers' expectations and their linked imaginaries (Pelizza and Van Rossem, 2021).

Furthermore, the conceptualization of a script introduces the possibility of accounting for resistance to technological artefacts in the form of a gap between expected uses, skills and abilities of intended users as they are assumed by artefact designers, and the actual uses and practices of empirical users. When intended and actual uses and skills

correspond, we have – to use script jargon – "subscriptions"; when they do not, the term is "dis-inscriptions" (Akrich and Latour, 1992). When there is a gap it is usually investigated by comparing artefact analyses with an ethnographic observation of actual use. Also, in the case of data models, the expectations and imaginaries that they entail may be resisted, and new ones may be proposed. Our methodology should thus systematically support an analysis of subscriptions, as well as dis-inscriptions.

The scripts may also show how the material properties of information (e.g., data records, data models, and databases) shape and structure how data are created and circulates, and shapes organizations and practices (Dourish, 2017). Paradigms and approaches for thinking about, modelling, and managing data all influence the creation of a data model (Thomer and Wickett, 2020). A data model will invariably vary depending on the kind of database that is employed, whether it be relational, object oriented, or graph based, and will need to subscribe to the expectations of the chosen database system. In the widely used relational model, for example, data are organized in tables with columns indicating data values (such as a surname) to be stored, and each row representing a relationship between these data values (for example, a link between a surname, date of birth, and nationality; or between a passport ID and a date of entry into a country). In this way the paradigms and technologies underlying data models influence what counts as data, how data can be used, and for whom.

Scripts and their material properties thus shape how informational representations inscribe and enact certain expectations for both categories and entities. The granularity of categories of data and their possible values, for example, will allow the capturing of only certain kinds of information about someone or something. To our knowledge, however, such a script approach has not yet been employed for digital methods that analyze technicalities of information infrastructures.

4.2.5 Digital methods

Drawing on a sociotechnical background similar to the one in which script analysis originally emerged, more recent developments have moved away from the technicalities of information infrastructures and prompted issue analysis as a field in need of innovative methodological solutions (Rogers et al., 2015). Issue analysis mainly relies on social media and web scraping techniques to analyze emerging topics, sentiments, and imaginaries on the web. However, data interventions also need to examine structures of data, such as the data models underlying the informational representations of various systems.

While it is not the aim of this chapter to foreground the epistemic and methodological challenges brought about by the digitization of social life (for a thorough analysis see Marres, 2017), it is noteworthy that the most common techniques adopted in the growing area of digital sociology combine scraping techniques of homogeneous web data, co-occurrence analysis of their lexicon, and some form of result comparison, usually through visualizations. Yet a number of research goals – like in the analyses of information systems for population management – require methods that go back to the tech-

nicalities of information infrastructures rather than focusing on (user-)generated contents. Details of diverse information systems' data models may need to be systematically compared by harmonizing non-homogeneous data formats to support discursive analysis. For instance, differences and absences in data models may suggest differences in the expectations and imaginaries of diverse authorities towards social actors, and forms of resistance to those expectations. This is what the OE is meant to achieve and makes it distinctive within the field of digital sociology and its current developments.

4.3 Design principles of the OE as a methodology: Script, comparison and resistance

Following the previous review of methodological approaches, in this section we describe three principles that have guided the development of the OE. First of all, however, we wish to make clear that we are not laying claim to the well-established achievements of others, as the OE is only novel in that it has integrated insights from existing classification and infrastructure studies, discourse analysis, and other strands described in the previous section. What makes the OE distinctive is the greater analytical depth into data infrastructures for population management, and the ad hoc combination of principles, requirements, and solutions.

The first principle followed in the development of the OE assumes that data models, as the actual encoding of an ontology in an information system, can reconstruct the discursive model of a domain that is operative in that system. As was discussed in Section 4.2 about data models as scripts, through "infrastructural inversions" (Bowker and Star, 1999) it is possible to retroactively reconstruct not only the discourse that informed the design of a system, but also expectations and imaginaries about the object of knowledge. In the case of data models underpinning information systems for population management, this means that the way populations are endogenously ordered and classified provides a "meta layer" of data for the observer. By looking at the assumptions that are made about people and examining the possibilities that are excluded by design, the analysis of data models can apprehend how people are inscribed in information systems.

In the case of information systems for population management run by authorities, an analysis of data models can capture authorities' own imaginaries about populations: their expectations, assumptions and patterns of exclusion. Such imaginaries can be assimilated into scripts, as they inscribe such expectations about people in the materiality of data models (Pelizza and Van Rossem, 2023). Our methodology should, therefore, systematically support a discursive analysis of data models to single out the scripts about individuals.

The second principle used in the development of the OE requires that such analytical support should proceed by comparison. Data models can be extremely "thin": by themselves, categories and values do not appear to reveal much evidence about the scripts they embed. Knowing the list of countries that work as values of the category "nation-

ality" might not lead to substantive outcomes. Yet, as semiotics has theorized, meaning emerges from comparison (see also Latour, 2005, and the methodological use that he makes of controversies). Meaning stems from the presence of (a set of) categories in some data models and absence in some others. Through the act of comparison we can see that the inclusion of a thin category like "profession" in one system but not another reveals authorities' differing imaginaries about people. For example, imagine two border control operations; one uses a migrant registration information system that includes only the category "language," while the other distinguishes between "native language" and "spoken languages." In the first case, people are enacted as monolingual, in the second case, as polyglots.

A method that wishes to capture the scripts embedded in data models should therefore proceed by comparing those data models in order to detect meaningful differences and absences. Furthermore, such principle is also relevant when it comes to analyzing "geographies of responsibility" (Akrich, 1992; Oudshoorn, 2012); that is, the division of labor between authorities involved in population management. As we have shown in our 2021 publication, collecting certain categories of data about immigrants at European borders can profile some Member States as asylum states. By design, not collecting this data can impede the possibility for EU agencies and other Member States to host and integrate newly arrived people.

The third principle in the development of the OE requires that the new methodology can account for resistance to the scripts identified. The original formulation of a script indeed foresees cases of "dis-inscription" in which users do not comply with their intended representation, adopt unforeseen uses of artefacts, and act in unexpected ways (Akrich and Latour, 1992). So, what are examples of unforeseen uses of data models? Again, examples come from the field of migration management. "Categorical stretching," for instance, has been defined by Pelizza (2019) as the behavior of border crossers who refuse to comply with Western categorical definitions of family, work, nationality, etc. Also, border officers might have levels of discretion (reliant on local knowledge) that resist certain standardizations, so influencing Schengen visa applications through street-level bureaucratic practices (Zampagni, 2016). In such cases, actual people resist the script embedded in data models and propose alternative ones. Similarly, our method should be able to conduct script analyses of intended people, while at the same time accounting for the possibility of practices of resistance exerted by actual people (to be further investigated through ethnographic research).

4.4 The OE as a method and a tool

As we've established, information systems define their data models in ways that are not immediately comparable. We therefore needed a method which would allow the extracting, analyzing, comparing, and visualizing of data models from heterogeneous sources. This section describes the different steps that we undertook to analyze data models sys-

tematically and quantitatively.

4.4.1 Data coding

As previously stated, documents reporting data models and their categories of data, i.e., the labels describing a state that can assume different values, can be very diverse: from regulations to design documents, from screenshots to interview transcripts. Such documents can be collected through desk research or fieldwork. However, to make such diverse descriptions of data models comparable we introduced a systematic approach to code, harmonize, and group all documents, categories, and values.

The data coding consisted of three phases: preliminary in vivo coding, code harmonization, and document grouping. All phases of data coding were conducted using computer-assisted qualitative data analysis software (CAQDAS). Preliminary in vivo coding entailed coding each category "in vivo," that is, using the category name as the code. Code harmonization was then necessary to enable the grouping of codes that refer to the same category of data but with minor variations among data models reported in different documents, or when categories were in different languages. Each of the code groups (e.g., "language," see Figure 4.2) gathers together similar categorical codes (e.g., "mother tongue," "native language," "communication language"). Values of categories were coded with the category name added as metadata in the code name and an additional meta-code. For example, for the category "nationality" we coded values that can be assigned, such as "Belgian" as "nationality: Belgian."

Document grouping was meant to identify data models and their source as independent variables. As one system's data model might be spread over several separately collected documents, it was necessary to group those documents together (Figure 4.2). This approach allows co-occurrence analysis to take place between data models (i.e., document groups) and codes or code groups, in order to obtain the frequency of a category in a data model.

Figure 4.2 outlines a use of the data coding technique for a hypothetical information system. Moving from top to bottom, a document group gathers three different types of documents that describe a data model; these documents show the occurrences of the categories and values coded in vivo. Then, values are distinguished through metadata in the code and an additional meta-code. Finally, these categories and values are grouped in code groups.

4.4.2 Data analysis

After coding documents, categories and values, we conducted analyses comparing diverse data models. The analysis was informed by discourse analysis criteria, like the requirement to focus on absences as well as what was visible and present, which led us to create three analytical categories: presence, absence, and frequency. By focusing on presence and absence, we aimed to address two questions:

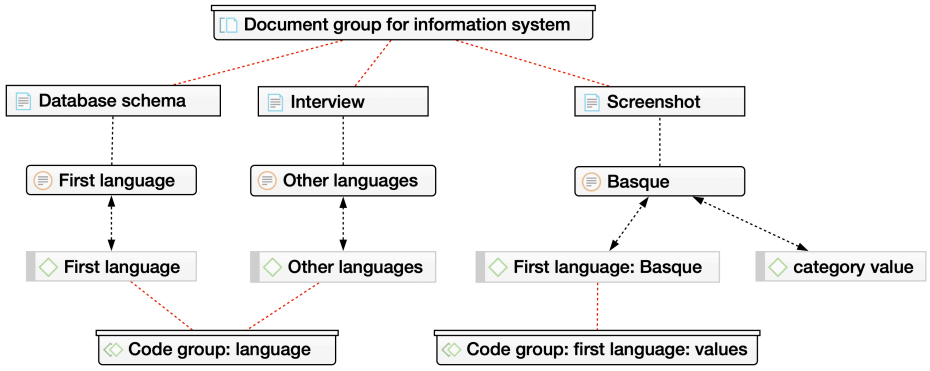


Figure 4.2: An example of the technique for coding data. From top to bottom there are four levels, one for each type of element: document group, document, quote, code, code group.

1. Which category or value used in one system is also present in another system?
2. Are some categories native or peculiar to a specific system?

By focusing on frequency, we aimed to address a third question:

3. In which systems do specific categories and values occur most frequently?

Our operationalization of the set of three questions was informed by graph theory, and used co-occurrence analysis and produced graph visualizations.[6] Such a combination of techniques has proven useful to visualize graph relationships and structures to help reason about complex networks (Hu and Nöllenburg, 2019). We decided to associate nodes with the data entities we coded as categories, documents, and groups; whereas links correspond to relationships of presence between nodes. The entire set of nodes and links results in a co-occurrence network. Figure 4.3 shows the whole process in pseudocode and using the notations introduced in the supplementary material. Our own analysis is primarily interested in the presence, absence, and frequency of code groups across document groups – corresponding to data models implemented by systems that may be spread over multiple documents.

4.4.3 Indications of presence, absence, and frequency

In order to support a discursive analysis approach – in particular, using comparison to trace invisibilities and absences, we analyzed subgraphs using set and graph theory

[6] Since the method described here is cross-disciplinary, we also take a cross-disciplinary approach to writing. The supplementary material is used to define concepts from the surrounding text in a formal way for readers that are familiar with notations and definitions of graph theory.

Input: The sets of Categories, Documents, CodeGroups, and DocGroups
$Categories = \{cat_1, cat_2, ..., cat_n\}$
$Documents = \{doc_1, doc_2, ..., doc_m\}$
$CodeGroups = \{codeGroup_1, codeGroup_2, ..., codeGroup_o\}$
$DocGroups = \{docGroup_1, docGroup_2, ..., docGroup_p\}$

Input: The mappings of categories to documents, of categories to code groups, and of documents to document groups
$Occurrence : (category \subseteq Categories) \mapsto (document \subseteq Document)$
$CodeGroup : (category \subseteq Categories) \mapsto (codeGroup \subseteq CodeGroups)$
$DocGroup : (doc \subseteq Documents) \mapsto (docGroup \subseteq DocGroups)$

Result: A directed graph: $G = (Nodes, Links)$ with elements as nodes, and direct and indirect links

$Nodes \longleftarrow Codes \cup Documents \cup CodeGroups \cup DocGroups$
$Links \longleftarrow \emptyset$

foreach $category \in Categories$ **do**
1 | **if** $(category \mapsto document) \in Occurrence$ **then**
| | $Links \leftarrow Links \cup (category, document)$
| **end**
2 | **if** $(category \mapsto codeGroup) \in CodeGroup$ **then**
| | $Links \leftarrow Links \cup (category, codeGroup)$
| **end**
end

foreach $document \in Documents$ **do**
3 | **if** $(document \mapsto docGroup) \in DocGroup$ **then**
| | $Links \leftarrow Links \cup (document, docGroup)$
| **end**
end

foreach $category \in Categories$ **do**
4 | **if** $(category \mapsto document) \in Occurrence$ **and** $(document \mapsto docGroup) \in DocGroup$ **then**
| | $Links \leftarrow Links \cup (category, docGroup)$
| **end**
5 | **if** $(category \mapsto document) \in Occurrence$ **and** $(document \mapsto docGroup) \in DocGroup$ **and**
| $(category \mapsto codeGroup) \in CodeGroup$ **then**
| | $Links \leftarrow Links \cup (codeGroup, docGroup)$
| **end**
end

return $G = (Nodes, Links)$

Figure 4.3: Pseudo code for creating the network and deriving direct and indirect occurrence relationships.

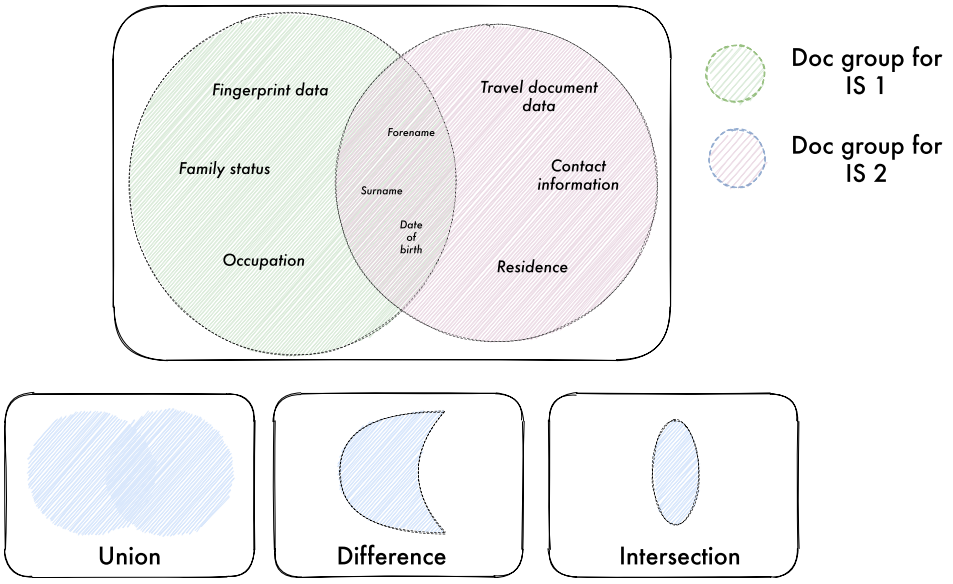


Figure 4.4: The different set operations.

to compute insights about presence/absence. In set and graph theory, *union* refers to the overall set that includes items from two separate sets. For example, the union of the separate sets {'first name', 'surname', 'age'} and {'surname', 'age', 'nationality'} is the set {'first name', 'surname', 'age', 'nationality'}. As well as union, other operators from set theory such as *intersection* and *difference* can also support a discourse analysis approach. The intersection of two sets is a set containing only elements that are part of both sets. So, the resulting intersection set from the previous example would be {'surname', 'age'}. A set difference shows the elements that are in one set but not the other. Using the operations from set theory on our graph model enabled us to analyze complex relationships of presence and absence of categories among different data models (Figure 4.4).[7]

Transforming data originally coded for qualitative analysis by means of CAQDAS to a co-occurrence network enabled us to further visualize relations of presence, absence, and frequency using special layout algorithms.[8] Techniques developed for graph visu-

[7] In the tool published at http://processingcitizenship.eu we provide a specific interface to conduct these kinds of set operations. After selecting the data models to include in the analysis and the appropriate set operations, we generate the results in the form of visualizations as well as in a table, which can then be exported to a file for further data analysis.

[8] We use Cytoscape.js, an open-source JavaScript graph theory library (Franz et al., 2016). As it is a JavaScript library, we could easily combine other web technologies, allowing our tool and results to be easily made shareable and publishable on the web. The generated network can also be exported for use in other tools such as

alization helped us to make qualitative observations about relationships and structures in the network. However, such observations became more difficult with larger networks so, for this reason, we needed additional measures to quantitatively express indications of significant nodes, groups of nodes, or network structures (Newman, 2018; Hu and Nöllenburg, 2019).

A class of visualization algorithms commonly used for co-occurrence networks are force-directed graph drawing algorithms. These kinds of algorithms use models from physics to metaphorically represent density of connection (Hu and Nöllenburg, 2019). The position of nodes in the drawing space is determined by forces of attraction and repulsion between these nodes. In this way, nodes occurring more often together are visualized as closer, while nodes occurring less frequently together are visualized as more distant. Such a spatialized visualization is known to help comprehend data (Kitchin and Dodge, 2011, p. 257-258). In our case, the generated spatial structure made it possible to visually analyze presence, absence, and frequency of categories in the co-occurrence network.

This metaphorical spatial representation is expressed numerically through the measures of *centrality* to identify the most important nodes according to a definition of importance (Newman, 2018). Two centrality measures that are useful to develop indications of categorical frequency are *degree centrality* and *betweenness centrality*. Degree centrality is a straightforward indicator that measures the importance of a node based on the number of links. A high degree centrality of nodes for code groups can indicate that these categories are used often by multiple systems; whereas a low degree centrality can indicate that a category is particular to a system. Betweenness centrality bases the importance of a node on the number of times it can be considered to form a bridge and potentially connect parts of the network. For example, in Figure 4.5 the node "date of birth" can serve as a bridge of the path from the node "surname" in one system to the node "nationality" in another system. In this way, these centrality measures allow us to find indications of important categories even if the network becomes too complex to analyze visually.

In the next section, we illustrate the actual functioning of the method by analyzing some data models of information systems for border management.

4.5 Illustrating potential uses: Information systems for population management at the European border

To illustrate and thus validate the use of the OE, this section provides a comparative analysis of data models utilized in national and international information systems for population management, which jointly work to support registration and identification practices at border zones in Europe.

Gephi.

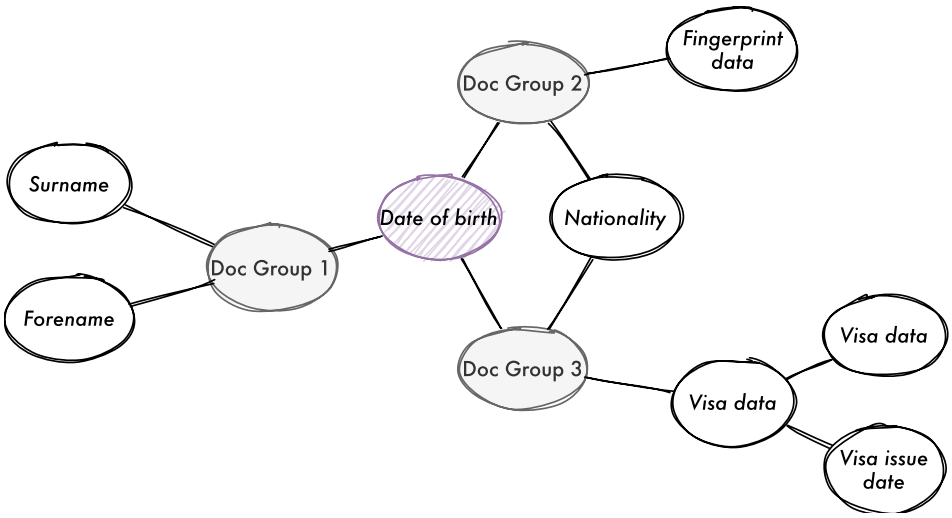


Figure 4.5: Network of categories. The node for the category "date of birth" will have a different centrality measures depending on if it is determined by degree or betweenness centrality.

4.5.1 Data collection

The data models used to design the method and tool proposed in this chapter were selected to answer the questions featured in Section 4.4.2, which in turn were informed by our three principles:

1. Is a category or value used in one system also present in another system? In the context of population management, this question intended to highlight differences and absences between systems run by different authorities, and therefore their diverse scripts, expectations, and imaginaries about populations.
2. Are some categories native or peculiar to a specific system? This question intended to identify categories belonging to a specific system in order to distinguish its mission.
3. In which systems do specific categories and values occur more frequently? With this question we aimed to identify which categories are used most in a range of different databases (e.g., national and international databases).

Data models were collected during fieldwork conducted as part of the Processing Citizenship project (2017-2023) at border zones in Europe.[9] Data collection included desk

[9] Data models were collected at different locations and in different European countries. Given linguistic constraints and the Project's task plan organized as a matrix, a few documents might have been collected by

research of European regulations, technical documents made available by European and national authorities, and systems screenshots collected at border zones in the Hellenic Republic from March to September 2018.

The comparison included three information systems developed by European Commission agencies: the European Asylum Dactyloscopy Database (Eurodac), the second generation Schengen Information System (SIS II), and the Visa Information System (VIS). All three systems have specific aims to support policing tasks related to undocumented migration, and cross-border crime and travel. Eurodac was established in 2003 to store fingerprints and other basic data of asylum seekers, and thus helps to identify asylum seekers this way, as well as by determining the Member State responsible for processing their asylum applications under the Dublin System.[10] SIS II aims to support external border control and law enforcement cooperation within the European Union. It stores alerts that contain information on people and objects, and information/instructions on what to do when they encounter these people or objects. This information can then be exchanged between law and border enforcement authorities of EU countries through the SIRENE cooperation network of national bureaus. Finally, VIS is a visa data exchange (including personal data and biometrics) that supports a common EU visa policy. VIS data checks are done via identification procedures at borders, and the data gathered in VIS can also be used by asylum authorities to determine which EU Member State is responsible for assessing the asylum application.

When looking at national systems, we compared the Hellenic Register of Foreigners (HRF) and the German Register of Foreigners (GRF). The HRF is the main system used at border zones in Greece to identify and register the arrival of foreign nationals. The system is used to support different tasks during the identification and asylum procedures: retrieving migrants' biographic and biometric data, conducting screening and asylum interviews, and assessing health conditions. Users of the systems are, therefore, wide ranging, including police, administrative personnel, and asylum officers.

The GRF contains a large amount of personal information on foreigners in Germany who have or have had a residence permit, as well as those who seek or have sought asylum, or are recognized asylum seekers (Bundesverwaltungsamt, 2021). This central register is accessed by various partner authorities and organizations in fields such as asylum, migration, and border control. The data sent to the GRF during the first registration is described in the XAusländer standard, a data exchange format that formalizes and enables data exchange between German immigration authorities (Bundesamt für Migration und Flüchtlinge, 2020). According to the description of the standard's motivation, it aims to facilitate exchange of such data between authorities in Germany in order to reduce data re-entry and to enable data reuse.

researchers employed on the project other than the authors. Yet the authors are the only responsible for the design and writing of this chapter.

[10] The Dublin System (Regulation No. 604/2013; also known as the Dublin III Regulation) establishes the criteria and mechanisms for determining which EU Member State is responsible for examining an asylum application.

4.5.2 Data coding

Categories and values were coded using in vivo coding; in the case of data models for migration and population management, categories included "country of origin," "profession," and "family name." We also coded category values, which are possible states associated with a category: "flight attendant," "teacher," "chef" are all possible values of the category "profession."

Code harmonization made it possible to group codes that refer to similar categories of data across different documents. Each such code group (e.g., "language") gathers together similar categorical codes (e.g., "mother tongue," "native language," "communication language"). This step allowed us to compare which types of categories of data are present and absent within each system.

Document grouping was also needed to run co-occurrence analyses for authorities in order to obtain the presence, absence, and frequency of categories. For example, the data models implemented by the HRF were collected via multiple screenshots and interviews. Being images, each screenshot constitutes a separate document. Therefore, grouping the individual screenshot documents and the interview transcriptions in one document group allowed us to view the whole data model as one unit of analysis.

4.5.3 Application

To test the OE method, we adopted two research questions that were crucial to the Processing Citizenship's project.[11] First, how does a diverse range of national and European authorities differ in their expectations and imaginaries about migrants? Second, what elements do such varying scripts share? These two questions addressed the need to account for different ways of enacting migrants' identities on different scales, despite the European Commission's effort to harmonize migration management. The hypothesis leading the first question assumes that obstacles in the standardization of migration management can depend on diverse expectations and imaginaries about border crossers and populations at large. The second question hypothesizes that, among categories that are common between authorities, there are some categories which may have the potential to connect identities across systems.

To answer these research questions, we operationalized the method in two ways. First, by using the analytical tools of presence and absence. Second, by using the analytical tools of frequency and centrality. In order to operationalize the first question, we identified authorities and their systems as the independent variable. Consequently, the original question was more operationally rephrased as: which groups of categories (code groups) are present in some systems and related data models, but not in others? As specified in Sections 4.2 and 4.3 above, this rephrasing assumes one of the basic method-

[11] The goal of this section is to test the validity of the method, so we do not report and analyze results of these questions in this chapter. However, this area of research is covered in Pelizza and Van Rossem (2023).

ological intuitions of discourse analyses, namely the need to pay attention to what is discarded.

Figure 4.6 and Table A.1 summarize the results of operationalizing the two questions.[12] The visualization of the network provides an overview of the characteristics of the different data models and their categories of data (see Figure 4.6). At the edges of the network are categories particular to each authority. Between the nodes representing the authorities are categories that are present in both data models. Finally, in the center of the visualization are categories shared by all three authorities.

In Table A.1, each row corresponds to a code group, while the columns for the three authorities indicate if the code group is present (blue) or absent (void) in the authorities' data models. The EU column is based on a combination of the Eurodac, SIS II, and VIS data models. The Greek column refers to the HRF and the German column to the GRF. Rows are categorized according to the code group's overall frequency in all selected systems. Categories of common data are therefore at the top of the table and include, among others, "nationality," "surname," "date of birth" and "sex." At the bottom of the table are more specific categories and code groups such as data related to integration or law enforcement, which only appear in a few systems. This table indeed allows us to see which categories are absent in some systems, like data related to language spoken or civil status.

The second research question can be operationalized through measures of centrality. The outcome of computing the betweenness centrality measure can be found in the last column of Table A.1, while degree centrality (frequency) is shown in the first column. The numbers computed through the betweenness measure are represented in this table as a line range to show their relative centrality. For example, categories common to all three authorities such as "nationality," "surname," "date of birth" and "sex" show differences in their betweenness centrality. This is explained by the fact that while some categories occur frequently, the nodes for these categories may be more or less important in connecting different parts of the network.

4.5.4 Discussion

One of the main goals of the OE was to allow the systematic comparison of diverse information systems, also by harmonizing non-homogeneous data formats. The coding scheme we developed successfully allowed us to compare diverse data model formats, even when developed by different organizations for different purposes, and in different formats and languages. In this way, the homogeneous comparison of data models (Table A.1 in Appendix) makes visible which categories are required by one authority and not by another. For example, we observed that only national systems collect data such as "ethnicity," "language," "temporary accommodation." Through the coding schemes it

[12] For the visualization, the network generated by the OE was exported to make use of the Gephi tool for adding the colors and other post-processing.

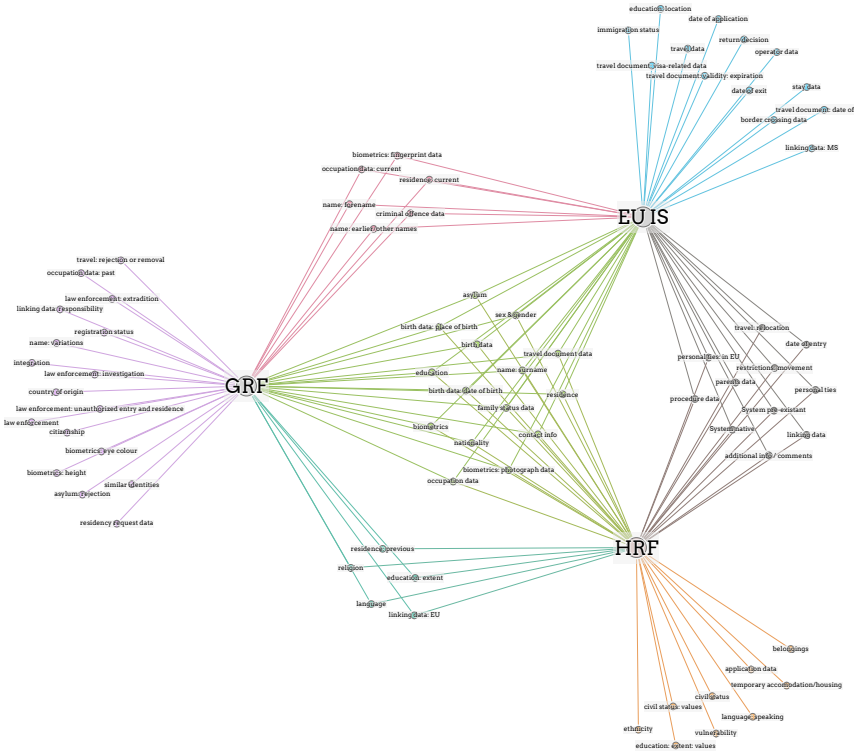


Figure 4.6: A visualization of the network of categories present for the three authorities. Color coding is used to highlight some of the characteristics of the network.

is also possible to see that not all authorities employ the same granularity. This is true, for example, of categories related to names such as "name: earlier/other names" which do not appear in the HRF. The method also allows values to be compared and thus enables differences in the granularity of values (e.g., category "education level") to be seen. We observed, for example, that the German data model is much more specific than the Hellenic model for categories such as "education level" or "marital status."

The analysis can, for this reason, uncover the scripts, possibilities and limitations, and thus the imaginaries of diverse authorities about people on the move. The example analyses show how national and European authorities differ in their scripts about migrants. Authorities can differ in the type and granularity of the data they collect, and in the amount of control the systems allow in adapting for local circumstances. These differences can help to further analyze how, for example, migrants are enacted as persons in need of integration (a category only included in the GRF, see also Pelizza and Van Rossem (2023)), or how their family relations are enacted – the HRF seems to focus on parents' data; the GRF family data can include children, parents, spouse, life partners. At the same time, there are also constants in how authorities enact migrants. Yet, despite these constants, measures of centrality reveal that some categories (such as "date of birth" and "nationality") could be more important than others to connect identities registered in multiple systems.

Furthermore, the analysis can aid understanding of how forms of resistance and adaptation can be accounted for. For example, the HRF reports categories like "mother tongue," "communication language" and "languages (other)." This last category suggests that the data model allows for some control for mismatching declarations, such as when the language of an applicant is not in the list of available languages. In the GRF we also observed the use of a particular data type for date of birth that may be incomplete; from our fieldwork experience, we know this may prevent frictional situations. When only a person's year of birth is known, officers might introduce inaccurate data by filling in a dummy day and month such as the first of January of that year. In general, the OE is tasked with providing a method and a tool to comparatively single out scripts about populations and their expected or intended identities. As Pelizza and Van Rossem (2023) have shown, people do not exert resistance in a vacuum but against a backdrop of these scripts. Script analysis can thus constitute the first analytical step to ethnographically uncover forms of resistance by diverse actors.

Finally, the method allows us to recognize some of the material properties of data models; for example, in the HRF data model it is possible to connect different individuals via the category "members of the case" and the use of internal identification numbers. We can also see how categories are arranged in relation to each other; for example, we can spot temporal alignments such as "date of application" and "date of exit," or alignments of the route a person has taken in categories such as "previous country of residence" and "date of entry in Greece." Such materialities enable and constrain how data can be created and used.

All in all, these illustrative analyses confirm that the method proposed complies to a satisfactory extent with the research goals and design principles. The analysis can give indications of where to focus our attention, but may need to be combined with further data to turn thin indications into thick descriptions; for example, by complementing the analysis of individual categories with fieldwork observations. There is one particular enlightening example: the table and visualization shows that HRF does not have actual fingerprint data as a category. Fieldwork observation has revealed that the system does not directly collect fingerprints, but it invokes a separate component to integrate with Eurodac (Pelizza and Van Rossem, 2021). We therefore decided to further group such categories of data which refer to other systems as "linking data" in order to understand how the EU systems link to Member States' systems, or vice versa. Based on this analysis we observed that only the national systems in our comparison have links to systems of international organizations.

4.6 Conclusions

In this chapter we outlined a novel methodology that integrates a number of insights from early classification, infrastructure and script studies and extends their methods and analytical depth to account for contemporary data infrastructures. By so doing, we hope to trigger a systematic discussion on how to extend those early theoretical and epistemological insights at the semantic level to contemporary developments in digital methods and for data infrastructures for population management. To achieve this, we validated the OE method and tool in light of the features of existing methods, as shown in Section 4.2. As we have outlined, the OE is inspired by earlier research which aimed to make visible the otherwise inconspicuous work that informational representations, classifications and standards do in shaping and segmenting the world. The OE builds on this tradition and supports analysis of differences and absences; it also provides quantitative comparative results that can inform further discourse analysis. Indeed, what makes our method distinct is that, by harmonizing non-homogeneous data formats, it systematically pursues comparisons of data models across diverse information systems. Such advanced comparison abilities makes it possible to detect meaningful differences and absences in the scripts embedded in data models.

We illustrated this method in action by comparatively analyzing data models used in information systems aimed at identifying and registering migrants arriving at the European border. This gives insights into, and allows comparison of, authorities' expectations and imaginaries about populations that are assimilated into scripts. Two research questions guided this analysis to compare the enactments of migrants by different European authorities. Our first findings reveal there are both constants and differences in authorities' assumptions about migrants, and show which possibilities are excluded by design. This provides a significant level of insight into the imaginaries and division of labor between authorities involved in population management. The findings thus showed

the usefulness of our method of combining some of the most robust features of both discourse and network analysis.

Furthermore, the OE method provides the possibility to account for resistance. The results obtained through the OE can indeed constitute the first element of critical analyses when put together with ethnographic observations of the actual use of information systems. The discursive analysis of "thin" data models can thus be used for detecting differences and absences between systems. The scripts found can be used at a later point to turn these indications into "thicker" descriptions.

The method can also contribute to current debates around the intervention into the politics of data and data-driven forms of governance. Although current developments in digital sociology mostly draw on user-generated content, interventions into data also need to examine extant structures, such as the data models underlying the informational representations of various systems. Presenting the data models in ways that make them easier to comprehend could support experimental forms of participation and speculation that would enable a rethink of collected categories of data with actors that are otherwise only data subjects (Grommé and Ruppert, 2021). This is because the OE constitutes a method that goes back to the technicalities of information infrastructures rather than focusing on high-level textual contents.

All in all, the method presented in this chapter shows that there is still plenty of potential for innovative methodological contributions towards the analysis of information systems' data models. However, it is important to stress that caution must be taken when interpreting the results of the analyses. Following Venturini et al. (2021), we reflected on how the network and visualizations we created should not just show connected complexity but help us to understand structures of networks. These authors rightly noted that ambiguities in network analysis and visualizations can actually be considered a strength, and we agree that this encourages further and deeper analysis into the ambiguity of empirical data. The method, therefore, can support a discursive analysis of information systems' data models, but it is not specifically designed to give definitive answers. In this way, we see our strategy of assembling different methodological approaches and traditions as, what Marres and Gerlitz (2016) calls, an "interface method." With this concept they urge paying close attention to the process of assembling different methods and how any tensions can, in fact, be productive. As with any other quali-quantitative analyses, care should be taken at each different step; especially the coding and code-grouping steps are crucial to obtain reliable results. The evidence obtained from the OE could be integrated with other observations, such as those from fieldwork.

Our future work on the OE will concentrate on applying the method to more in-depth and differentiated analyses of data models used by authorities responsible for border and migration control. More broadly, we also plan to further develop this method in other fields of application and expand the software tools so that they can serve as a base for future studies aiming to employ qualitative and quantitative methods for comparative analyses of data models.

Chapter 5

From registration to re-identification: Exploring the interplay of data matching software in routine identification practices

Abstract

In migration management and border control, identifying individuals across data infrastructures frequently demands intricate processes involving integrating and aligning diverse data across various organizations, temporal contexts, and geographical boundaries. While existing literature predominantly focuses on first registration and identification, this chapter takes a novel empirical route by investigating the "re-identification" of applicants in bureaucratic processes within the Netherlands' Immigration and Naturalization Service (IND). Re-identification is conceptualized as the continuous utilization and interconnection of data from various sources to ascertain whether multiple sets of identity data correspond to a singular real-world individual. Through the lens of re-identification, the research examines the iterative processes of identifying applicants across various stages of bureaucratic processes, drawing from fieldwork and interviews conducted at the IND and its data matching software provider. The study delves into the IND's designed infrastructure for applicant re-identification, particularly the tools for searching and matching identity data. By contrasting design and practical use, the research uncovers multiple forms of data friction that may hinder re-identification. Furthermore, exploring the costs stemming from failed re-identification, manifested through duplicate records and the labor-intensive deduplication process, highlights the evolving bureaucratic re-identification practices and their links with transnational security infrastructures. The findings contribute to debates about the materiality and performativity of identification in two ways. Firstly, they redirect attention from first registration and identification to encompass re-identification practices across data infrastructures. Through an interpretative framework developed in the analysis, re-identification is further demonstrated not as a

singular process but as a range of iterative practices. Secondly, the findings underscore that while integrating data matching tools for re-identification alleviates data friction, it inadvertently also comes with certain costs. This integration involves a redistribution of re-identification competencies and labor between the IND and the commercialized data matching engine and potentially shifting the burden of costs related to failed re-identification to different parts of the bureaucratic system.

Contribution to research objectives *To examine the relationship between identity data matching technologies and routine identification practices.*

This chapter provides an in-depth analysis of the interconnections between identity data matching technologies and routine identification practices, focusing on the operational intricacies of re-identification within migration management, such as in the context of residency or naturalization applications. Through empirical investigation, the chapter explores the utilization and interconnection of data from various sources to ascertain whether multiple database records correspond to individual applicants. By comparing the design of data matching tools in contrast to their real-world implementation, the comparison unveils multiple forms of data friction that can impede the IND's re-identification processes. One form of friction arises from variations in the precision and accuracy of identity data during its transformation across different mediums, which subsequently influences the formulation of search queries. Furthermore, data friction can arise from the opaque calculation of match results by the tools, making it challenging for IND staff to understand search results, leading to the need for refining search parameters and strategies. The investigation highlights the possible costs of these forms of friction by examining the consequences of unsuccessful re-identification, exemplified by duplicate records.

Infrastructural inversion strategy used *Second Inversion Strategy — Data Practices*

This chapter uses the second infrastructural inversion strategy to examine the practices related to identity matching and linking across data infrastructures. Using this strategy, the chapter's findings highlight a significant diversity in re-identification practices within the IND. This diversity emerges in two main aspects. On the one hand, re-identification practices are characterized by the information available to staff during the process. On the other hand, the precision criteria necessary for successful re-identification exhibit significant variation. Building on these observations, the chapter develops an interpretative framework that categorizes re-identification practices based on the demands of interpreting search inputs and results. The resulting matrix recognizes re-identification not as a solitary process but as a range of iterative practices. These practices encompass a variety of scenarios, including direct applicant interactions, staff managing phone conversations, handling application forms sent via postal services, and

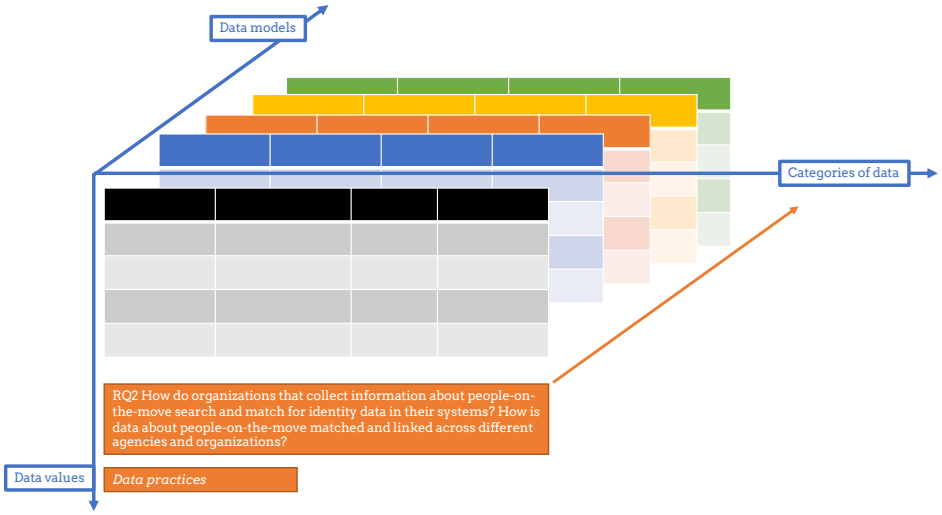


Figure 5.1: The axes pertaining to the methodological framework in relation to chapter 5.

automated re-identification processes.

Contribution to research questions *RQ2: How do organizations that collect information about people-on-the-move search and match for identity data in their systems? How is data about people-on-the-move matched and linked across different agencies and organizations?* This chapter addresses RQ2 by analyzing the various re-identification practices that encompass a range of scenarios reflecting diverse re-identification practices within the IND, the broader collaborative framework of the Netherlands' migration chain, and even extending to transnational data infrastructures. Internally within the IND, these scenarios encapsulate various contexts, from direct applicant interactions to staff managing telephone conversations and processing physical application forms. Across the broader migration chain, the utilization of the v-number identification numbers emerges as one element to match and link data across systems. However, the study also uncovers instances where re-identification efforts falter due to, for instance, inconsistencies in identification practices across the migration chain partners. Furthermore, the analysis of the deduplication process unveiled a connection to transnational systems, as the IND and migration chain partners harness data from prominent European Union information systems to forge connections among seemingly disparate records within their respective databases.

Contribution to the main research question *How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transna-*

tional commercialized security infrastructures?

This chapter addresses the main research question by investigating the interplay between the re-identification practices of the IND and a commercially developed data matching system. This integration of proprietary tools, designed by a private entity, for data matching not only signifies a transformation in the IND's re-identification approach but also underscores the redistribution of re-identification expertise and competencies as these are embedded within a proprietary data matching system. Furthermore, the chapter highlights a link between the deduplication process and transnational systems. The IND and migration chain partners effectively employ data from prominent European Union information systems, forging connections among apparently disparate records within their databases. Moreover, the findings demonstrated challenges faced by vendors in designing custom-made solutions versus standardized approaches for deduplication. The complexities of defining identity and duplicate records, inherently linked to the specific organizational context, became apparent during an upgrade of automated duplicate detection tools. These findings emphasize that re-identification processes and associated technologies are far from isolated; they are intricately entwined within broader commercialized security infrastructures.

5.1 Introduction

The stakes in identification can be high, and authorities' use of specialized technologies to search and match identity data can significantly mediate uncertain identification outcomes. An often invoked real-life example of the complexities of identification is when one of the "Boston bombers," Kyrgyz-American Тамерла́н Царна́ев, was *not* pulled aside for questioning when leaving from and returning to JFK Airport in New York for a trip to Dagestan in the Northern Caucasus in 2012 (an area considered as a high-risk travel destination by the US government). [1] In April 2013, he and his brother carried out a terrorist attack during the annual Boston Marathon. According to an investigative report for the United States House Committee on Homeland Security, which media outlets reviewed, he was mistakenly not identified as a person of interest nor questioned (Schmitt and Schmidt, 2013; Winter, 2014). In 2011, Russian authorities had already informed their American counterparts of his ties to terrorist organizations. Following this information exchange, the US government added him to various watch lists and databases, including the Terrorist Identities Datamart Environment, which contains information on over 1.5 million people who are either known or suspected to be international terrorists. Such watchlisting systems automatically compare data about individuals, alerting and instructing authorities on what to do when they encounter someone whose data matches a watchlist entry. Due to missing information regarding Mr Царна́ев's date of birth and variations in the transliteration of his name, "Tsarnaev" — "Tsarnayev," the system did not raise an alert in his case.

The case of the Boston bomber exemplifies at least three essential features of modern identification practices mediated by digital technologies: inherent difficulties, technological solutionism and flexibility in the application of regulation. First, at all levels of bureaucracy, from street-level bureaucrats to system-level bureaucracies, there are inherent difficulties in accurately identifying and confirming identities. Most organizations must cope with databases containing incomplete, not current, incorrect data — or even duplicate entries referring to the same real-world persons (van Keulen, 2012). Such difficulties undermine the possibility of building trust. For example, the European Union Agency for Fundamental Rights (FRA)'s investigation on the implications of data quality in EU information systems for migration and border control on fundamental rights (2018) has reported that "authorities often suspect identity fraud when cases of data quality are the real reason for concern" (p. 81). Hence, ensuring that data are correct, complete, and accurate and that they can be shared, used, and processed by different parties and information systems is deemed vital for the functioning of (bureaucratic) proce-

[1] Tech companies often use the Boston bomber's misspelt names as an example of why watch list screening systems need their data matching technologies (see also Basis Technology, 2021). Businesses can use this scenario in their sales pitch to show how their technology can handle the ambiguity in identifying and connecting individuals' identities. Despite the risk of perpetuating the securitization of identification, this case is instructive as a practical illustration of the interdependence of various government agencies, border guards, and watchlisting systems in identifying potentially risky travelers.

dures.[2]

Within the context of identifying individuals in migration management and border control scenarios, inherent difficulties can be explained by the need to interconnect and harmonize diverse datasets spread across different organizations, timeframes, and geographical boundaries. For alphanumeric personal data (such as surname, date of birth, and nationality), these data quality issues can have various, often unspectacular, reasons. The case of Тамерла́н Царна́ев touches on the fact that watchlist databases need Latin characters' names, yet, transliteration of a name can take many forms. Hence, working with different data sources usually brings challenges related to what I will term "re-identification." With this term, I intend to encompass a spectrum of iterative identification processes where data, whether sourced from within or across organizations and collected across diverse temporal and spatial contexts, are employed and interconnected to determine if multiple database records correspond to a single real-world individual.[3] Instances of such re-identification encompass diverse scenarios, ranging from cross-referencing an individual's passport details to access their visa records, correlating flight information to identify matches on watchlists, or linking migration and law enforcement databases to unveil potential suspect identities.

Second, new technologies keep being introduced in an attempt to solve "data friction" (Edwards, 2010), and re-identification should be perceived not only as being disrupted by data friction but also as a means through which technology can introduce friction. Research should consider how technologies for searching and matching identity data reconfigure re-identification practices. This point draws on materialist and performativity debate on identification (e.g., Fors-Owczynik and van der Ploeg, 2015; Leese, 2022; Pelizza, 2021; van der Ploeg, 1999; Pollozek and Passoth, 2019; Skinner, 2018). Following these debates, identification should not be understood as a problem of truthful representation between people and their identity data but of how data infrastructures for identity management and identification "enact" individuals as migrants, criminals, risky travelers. From this perspective, we can rethink the above quote from the Fundamental Rights Agency. Instead of asking if doubts about someone's identity arise from inaccurate data or mistrustful border control practices, we must also consider how re-identification en-

[2] The General Data Protection Regulation (GDPR), for instance, states the accuracy principle in Article 5(1)d. According to this principle, the personal data that organizations collect and use must be "accurate and, where necessary, kept up to date," and "every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay."

[3] In technical literature, the term re-identification is also used to describe processes of de-anonymizing data, i.e., revealing personal identities associated with anonymized data. Re-identification of previously anonymized individuals is, in fact, always a possible outcome of data matching processes. As Christen (2012) explains, re-identification is possible because "record pairs classified as matches in a data matching project can contain information that is not available in the individual source databases that were matched" (p. 189). By matching data from different sources, individuals in those databases may be (un)intentionally identified and disclosed even with incomplete identifying information. Consequently, there is an undeniable connection between re-identification as de-anonymization and the practices and technologies described in this chapter.

acts subjects as potential identity frauds. For example, a case of potential identity fraud could be discovered by automatically matching similar biographical data. Such a materialist and performative approach can replace discussions of identity as faithful representation with accounts about how re-identification introduces novel forms of suspicion. So far, however, literature has focused on the materiality and performativity of first (and often biometric) identification and registration (e.g., biometric refugee registration), and there needs to be more discussion about how people are re-identified and enacted throughout bureaucratic practices and data infrastructures.

Third, the literature on street-level bureaucracy emphasizes that public employees put government policies into practice through their regular interactions with citizens to deal with complex situations that do not always fit neatly into the rules and regulations made by legislators (e.g., Lipsky, 2010). Re-identification can be ambiguous, and there is often a lack of clear guidelines; as a result, there is considerable room for discretion. However, the problem of re-identifying applicants during bureaucratic procedures has received little attention in the literature to date. A helpful example of an identification encounter where the tension between systems, policies, and local circumstances is apparent is provided by Pelizza (2021). She describes the back-and-forth between an applicant, a police officer, and a translator to convert the applicant's name from Arabic to Latin characters during first registration at a Greek border. The name that emerges from this identification encounter is, in Pelizza's words, the result of a "chain of translations" of the migrant's name from oral to written to finally end up in the information system to serve as the official version to be used to re-identify this person in future administrative procedures. The process Pelizza described is very different from an example I encountered in The Netherlands. In this case, when there are doubts or refusal to give a name, a person will be assigned a label that serves as a name and includes details about the applicant's sex as well as the time and place of registration (e.g., "NN regioncode sex yymmdd hhmm"). In both examples, public servants re-identify people by tailoring their actions to the individual involved, all within the constraints and affordances of a given sociotechnical setting.

However, while the street-level bureaucracy literature has long debated the constraining or enabling effects of new technologies, such as those related to automated decision-making (Bovens and Zouridis, 2002; Buffat, 2015), it has been less specific about the entangled technologies. Nonetheless, it is clear that the expectations and materialities of data shape identification encounters. The designs and data models of technical solutions, such as those used to search for a person's record or to determine whether two identity data records refer to the same person, embed many assumptions about those data (see also Pelizza and Van Rossem, 2023; Van Rossem and Pelizza, 2022), which shape bureaucratic re-identification. In the case of identity data, such tools assemble knowledge and enact equivalences between otherwise disparate naming practices. For example, the male and female family name forms might each have a slightly different final syllable, but they could still be considered equivalent. By examining how

applicants of bureaucratic procedures are re-identified, this chapter intends to answer RQ2:

> How do organizations that collect information about people-on-the-move search and match for identity data in their systems? How is data about people-on-the-move matched and linked across different agencies and organizations?

The research seeks to answer these questions by empirically studying re-identification at the migration and naturalization service in The Netherlands (IND). The analysis draws on data gathered through fieldwork — interviews, documents, field notes — at the data matching software supplier and the IND agency itself. The research hypothesized that the design of search and matching tools incorporates assumptions about databases and their data records, which influence and are influenced by bureaucratic re-identification practices. Assumptions like these include the possibility of incompatible naming practices and conventions, meaning databases could never be entirely accurate. Therefore, I formulated the hypothesis that utilizing data matching software could result in a redistribution of responsibilities and capabilities driven by the inherent affordances and limitations of the software itself. Specifically, I anticipated that government agents would rely less on their identification expertise and instead rely more on automated matching algorithms to retrieve identity information.

The chapter aims to contribute to the literature on the materiality and performativity of identification, particularly within the intersection of science and technology studies (STS) and critical migration, security, and border studies. The findings of this chapter can contribute to these scholarships in two ways. Firstly, while prior investigations have predominantly concentrated on initial registration, often involving biometric data, this chapter's findings illuminate the often-overlooked processes of re-identification that transpire throughout bureaucratic procedures. The chapter sheds light on lesser-known practices of dealing with uncertain alphanumeric biographic data in migration management. Secondly, by examining routine re-identification interactions embedded within specific sociotechnical contexts, the findings demonstrate how incorporating data matching tools, intending to curb data friction, sometimes shifts the costs associated with managing ambiguous data to other actors or entities within bureaucratic systems.

The next section of this chapter will commence with a review of the background and related work that has been instrumental in conceptualizing re-identification as a bureaucratic practice, emphasizing its intersections with the materiality and performativity of identification processes. Following this, an overview of the case and methodology adopted for examining matching systems and applicant re-identification at the Netherlands' Immigration and Naturalization Service (IND) will be presented. The chapter's empirical case and findings will be presented across three sections. Two empirical sections will juxtapose the designed applicant identification infrastructure and its practical implementation at the IND, recognizing three forms of data friction that can im-

pede the re-identification process. The third empirical section will delve into the costs of unsuccessful re-identification, focusing on the problems related to duplicate records and the labor-intensive deduplication process. Finally, the chapter will answer the chapter's research question by synthesizing the diverse re-identification practices encountered throughout the empirical sections into an interpretative framework, highlighting a range of re-identification scenarios.

5.2 Conceptualizing re-identification: Bureaucratic contexts and the dynamics of identity data

Many interactions between migrants and public authorities involve forms of identification to establish or verify applicants' identity in different steps of bureaucratic processes relating to granting asylum, issuing residency permits, naturalization, and so forth. As the literature on street-level bureaucracy has shown, public-service workers in charge do not just carry out relevant policies; they are also actively involved in interpretative work through the discretion workers use (e.g., Lipsky, 2010; Collins, 2016). What does it mean to regard re-identification practices as part of routine bureaucratic procedures?

5.2.1 Re-identification as a bureaucratic practice

Michael Lipsky's book "Street-level bureaucracy: dilemmas of the individual in public services" (published in 1980) is widely credited with popularizing the concepts of street-level bureaucracy and discretion. According to this widely held view, diverse frontline public service workers influence public policy through regular interactions with the general public. As Lipsky (2010) states: "street-level bureaucrats have considerable discretion in determining the nature, amount, and quality of benefits and sanctions provided by their agencies" (13). For instance, a border patrol agent may have discretionary authority to grant entry to a traveler based on the results of an identification encounter with the traveler at passport control. Lipsky's original argument, however, required updating in light of the increased use of digital technologies in government and the rise of e-government. Essentially, the digitalization of government agencies has compelled scholars to reconsider the role of street-level bureaucrats and their daily interactions (Bovens and Zouridis, 2002; Buffat, 2015; Busch and Henriksen, 2018; Snellen, 2002). With digitization, discretionary power is sensibly reduced as decisions are delegated to automated systems. This argument has implications for identification, which can occur not only in direct interactions between applicants and bureaucratic officers but also increasingly in automated processes mediated by digital tools.

The transformations brought about by information and communication technologies were conceptualized by Bovens and Zouridis (2002) in an influential article as occurring first at the "screen-level" and then at the "system-level" of bureaucracy. "Screen-level bureaucracies" refers to how interactions between officials and citizens have become in-

creasingly mediated through computer screens. For instance, the personal data of a residency permit applicant is filled out using electronic template forms in a case management system. Alternatively, increasingly applicants themselves are also provided access to government information systems (Landsbergen, 2004). Meanwhile, decision trees, business rules, and algorithms that model the policies and regulations will guide the decision to grant the permit in this example. "System-level bureaucracies" refers to an even higher level of automation and digitization when collecting data and carrying out routine tasks. The following is the author's idealized description of the practitioners' new roles in such an organization:

> The members of the organization are no longer involved in handling individual cases, but direct their focus toward system development and maintenance, toward optimizing information processes, and toward creating links between systems in various organizations. Contacts with customers are important, but these almost all concern assistance and information provided by help desk staff. After all, the transactions have all been fully automated. (Bovens and Zouridis, 2002, p. 178-179)

Within the bureaucratic processes, individuals applying for asylum may often find themselves subject to iterative identification procedures that traverse the realms of street-level, screen-level, and system-level bureaucracies. It typically commences at the street and screen levels, where front-line bureaucrats serve as the initial point of contact. These bureaucrats collect and input applicant information, funneling it into the complex interfaces of the bureaucratic systems. Subsequently, further decisions regarding data management, including updates, linkages, and corrections to applicant information, may be made at the system-level. This iterative process through bureaucratic layers underscores the importance of consistent and precise identification practices across all levels.

The concept of "re-identification" highlights the entanglement of street-level procedures and the crucial role in automated systems within system-level bureaucracies responsible for processing applications from individuals seeking services or assistance. Automated processes must correctly re-identify the distinct applicants of bureaucratic processes to make the right decisions. System development will thus be required to automatically re-identify individual cases and to ensure that data are accurate and up to date, that no duplicate entries exist, and so on. Moreover, as applicants themselves are provided access to government information systems, such as through the filling out of digital application forms, they can be assumed to become more involved in the re-identification process. This change brings an additional layer of complexity, as the accuracy and consistency of the data provided by the applicants also contribute to the success of re-identification processes. In a system-level bureaucracy, re-identification will be linked to verifying and connecting individual records across systems and organizations. [maybe this is the place to defi ne re-identification]

In the literature about the entanglement of street-level, screen-level, and system-level bureaucracies, the desirability of automation for fairness and efficiency is usually weighed against its potential negative impact on human judgement and autonomy. Buffat (2015) categorizes these debates into the "curtailment thesis" and "enablement thesis." The former argues that information and communication technology (ICT) limits frontline officers' discretion, transferring it to other actors. The "enablement thesis," on the other hand, suggests that technologies play a more nuanced role by shaping interactions between technologies, workers, and citizens. A more recent perspective, the "digital discretion" literature, proposes the use of "computerized routines and analyses to influence or replace human judgement" (Busch and Henriksen, 2018, p. 4) to adhere to policies and ensure fair and consistent outcomes. This chapter takes a further distinct STS-influenced approach, emphasizing how re-identification is intertwined with technology's affordances and constraints that shape bureaucratic realities. Such an STS lens prompts us to be specific about the use of technologies, such as how the design of data matching systems, their embedded algorithms, and their interfaces affect the daily routines of those involved in re-identification processes and ultimately shape the re-identification outcomes.

The literature suggests that when analyzing the interplay between re-identification, discretion, and varying levels of bureaucracy, two key elements should be taken into account. Firstly, the literature suggests that identification policies are executed by public workers in their daily routines, often influenced by their discretionary powers. Secondly, it emphasizes the need to view routine identification practices within bureaucratic frameworks in the context of broader changes in their sociotechnical systems. The concept of re-identification can help make sense of the interactions between bureaucratic organizations and applicants when these interactions combine street-level interactions, screen-level processes, and system-level bureaucracies. When there are uncertainties regarding precise identification, the discretionary components of procedures can become more important. In such scenarios, the interplay between human judgement and automated mechanisms could enhance or impede re-identification. It also raises questions about potential challenges associated with unsuccessful re-identification attempts, including subsequent consequences and necessary corrective measures.

Re-identification, as introduced here, is a concept that can offer insight into the entanglement between street-level, screen-level, and system-level bureaucracies. In the realm of bureaucratic processes, data and information frequently traverse these distinct levels, presenting both challenges and opportunities for the handling of applicant data. This concept of re-identification aims to untangle the complexities that arise when individuals engage with government systems and personnel, necessitating multiple rounds of identification and verification across bureaucratic contexts. By emphasizing the iterative character of identification, re-identification highlights the recurrent need for verifying individuals' identities. Furthermore, it underscores the pivotal role of technology, interfaces, and organizational structures in shaping identification processes within bu-

reaucratic systems.

5.2.2 Materiality and performativity of re-identification

A different body of literature further recognizes the importance of identification as intermingled with the government's obligations and rights (e.g., citizenship, residency), as well as coercive measures (About et al., 2013a; Caplan and Torpey, 2001). As recalled in Chapter 2, scholars have conventionally placed a significant emphasis on the interconnection between the formation of modern nation-states and the development of registration and identification systems, such as the creation of civil registers or passport documents (Breckenridge and Szreter, 2012; Caplan and Torpey, 2001; Torpey, 2018). An often preferred term for the state's capacity to identify its citizens is the notion of *legibility* of Scott (1998). Scott noted how the increased interaction of states and their population (e.g., for purposes of taxation) went hand in hand with projects of standardization and legibility as attempts to identify its people unambiguously. So, in the example by Scott, while cultural naming practices are very diverse and can serve local purposes, the standardization of surnames "was a first and crucial step toward making individual citizens officially legible" (p. 71). In these practices, the identity of the person is not a problem of representation between a person and information captured about them but one of reducing multiplicity while mutually enacting subjects, states, and institutions (Lyon, 2009; Pelizza, 2021). What needs to be clarified is how such concept of legibility and reducing multiplicity also intersects with the notion of re-identification, as the state's ongoing endeavor to ensure legibility involves not only initial identification but also successive processes of verifying and connecting data over time and across various contexts.

The growing body of literature at the intersection between STS and Critical Security Studies has added an important dimension to the discussion on identification by accounting for the materiality and performativity of devices and practices (Cole, 2001; Gargiulo, 2017; Skinner, 2018; Pelizza, 2021; Suchman et al., 2017). Bellanova and Glouftsios (2022), for instance, have studied the actors and practices involved in maintaining the EU Schengen Information System (SIS). The SIS system allows authorities to create and consult alerts on, among others, missing persons and on persons related to criminal offences. By looking at how these alerts "acquire the status of allegedly credible and accurate information that becomes available to end-users through the SIS II" (p. 2) they make evident its role in conditioning international mobility. Fors-Owczynik and van der Ploeg (2015) have shown how three systems in the Netherlands translate and frame risk categories to identify potentially risky migrants and travelers. Building on this literature, re-identification can be understood as intricately connected with the materiality and performativity of devices, shaping the evaluation of data, individuals, and organizations as accurate and trustworthy. Drawing on findings from the politics of mobility literature (Cresswell, 2010; Pallitro and Heyman, 2008; Salter, 2013), such as observations regarding the expedited processing of certain passenger classes through trusted traveler programs at airports, these disparities can lead to divergent outcomes. In cases where

discrepancies arise, individuals may be subject to heightened scrutiny and additional security measures. Conversely, consistent information across systems has the potential to expedite their passage through border controls.

Surprisingly, despite the significance of re-identification in contemporary bureaucratic practices, there remains a noticeable gap in our understanding of how practitioners navigate the complexities arising from ambiguities in personal identity data during re-identification. A case in point is highlighted in a report by the European Court of Auditors, which outlines that "when border guards check a name in SIS II [the Schengen Information System], they may receive hundreds of results (mostly false positives), which they are legally required to check manually" (ECA, 2020, p. 31). This operational challenge, rooted in the technology's approach to computing and presenting matching data, exemplifies how the concept of re-identification intersects with the practical realities of border control. The abundance of false positives generated by the system raises questions about how re-identification encounters are negotiated when dealing with such ambiguities and how technologies might influence these interactions.

5.2.3 Conceptualizing data friction in re-identification

Critical data studies have made it clear that data are never "raw" (Gitelman, 2013) and "contain traces of their own local production" (Loukissas, 2019, p. 67), and that work is therefore needed to put data to use. For example, a European Court of Auditors report mentions that a prominent EU information system supporting border control contains millions of potential data quality issues, such as first names recorded as surnames or missing dates of birth (ECA, 2020). Many such discrepancies are likely related to work practices and issues of fitting local circumstances to global standards (Bowker and Star, 1999). As Loukissas (2019) remarks, databases might contain various errors and "local knowledge [is needed] to see that such errors are not random" (p. 67). In this sense, data serve as evidence of the local conditions of their production, which, for future re-identification processes, must be linked across space and time. If data quality problems and uncertainty are facts of life (van Keulen, 2012), then bureaucratic organizations must cope with this uncertainty in re-identification practices.

As highlighted in Chapter 1, multiple technical mechanisms exist for dealing with such uncertainties, such as determining whether two or more data records pertain to the same real-world individual (Batini and Scannapieco, 2016). Data matching techniques will compare attributes of data records and use classification methods to determine matches (Christen, 2012). There are numerous classification techniques to determine matches: some are based on adhering to specific rules, while others take a more probabilistic approach. Metrics can, for example, calculate the similarity of two sequences of characters based on the number of operations required to transform one into the other. In this way, the names "Sam" and "Pam" may be considered closely related (for instance, was it a typo?). Other approaches may even calculate such similarities by comparing how names are pronounced (in English). When matching personal data,

rules-based matching may include ignoring honorifics and titles (e.g., Mr., Ms., Dr.). Although these technical mechanisms for data matching are widely recognized, their practical implications in the processes of re-identification remain less evident.

The insights gleaned from Critical Data Studies indicate that investigating technical data matching mechanisms not only reveals local conditions of data productions and operational dilemmas but also offers valuable insights into re-identification. This is exemplified by the specific case of data matching aimed at discovering and resolving duplicate data records. For instance, a migrant might inadvertently be registered multiple times in a database due to technical glitches. Typically, a deduplication process (for example, as detailed in Batini and Scannapieco, 2016) periodically compares each record with all others in the database to identify records pertaining to the same individual. A domain expert usually intervenes to make decisions regarding whether these matches do indeed pertain to the same individual to consolidate the multiple data into a single one.

Following Loukissas (2019), the process of "normalizing" duplicates can be "a key to learning about the heterogeneity of data infrastructures" (p. 60). Loukissas gives the example of software that identifies digital copies of books, newspapers, and objects in a digital library collection. He challenges the software's intention to eliminate these copies, suggesting that delving into the duplicates' origins could be more instructive. This discussion on duplicates holds relevance for re-identification in two ways. Firstly, deduplication can offer similar insights into multiple re-identification practices. Secondly, the presence of duplicates prompts another question: what are the implications for applicants and organizations of data matching failures and unsuccessful re-identification?

The complexities arising from impediments in the seamless flow of identity data may indeed be at the heart of unsuccessful data matching and re-identification processes, which can be aptly conceptualized as manifestations of "data friction" (Edwards, 2010). Data friction, according to Edwards (p. 84), "refers to the costs in time, energy, and attention required simply to collect, check, store, move, receive, and access data." Data friction signifies the barriers that disrupt the smooth flow of data across different actors, organizations, and material forms. As noted by Bates (2017), data friction is "influenced by a variety of infrastructural, sociocultural and regulatory factors interrelated with the broader political economic context," all of which influence the movement or hindrance of data. Pelizza (2016b) explains the process of addressing data friction as a dynamic interplay between aligning and replacing infrastructural elements that facilitate data movement, where changes in one aspect impact the other. In her study of the Dutch land registry, Pelizza (2016b) portrays data friction as conflicts revolving around finding the best configurations of actors, institutions, and resources to ensure dependable data. As she emphasizes, even in complex systems designed to mitigate friction, complete removal is often unattainable; instead, the associated costs tend to shift to alternative actors, organizations, or material forms. Consequently, we may hypothesize that data frictions concerning re-identification present associated costs, such as organizational labor, interpretive activities, task complexification.

Identity data takes on diverse forms as it navigates through different actors and organizations, and these transformations may entail associated costs when utilized for re-identification purposes. I propose that the concept of data friction can be extended to the movement of identity data across organizations and different material forms. For instance, as identity data transitions from a physical passport to a digital database record or moves between the systems of different organizations, barriers may emerge, leading to friction in the smooth movement of identity data and, consequently, re-identification. Regulatory constraints undoubtedly influence the movement of identity data between organizations. However, discrepancies may also arise due to variations in naming conventions, differences in date of birth formats, or inconsistencies in the use of characters like hyphens or spaces in surnames across organizations. To illustrate, let us consider my experience while applying for a Russian visa in 2018: my passport information was copied into various systems, leading to an error where my second name was mistaken for a patronymic in the application. Moreover, my first name was inadvertently listed with the letter "v" instead of "w" in the machine-readable zone of the visa due to the absence of the letter "w" in Cyrillic. These confusions stemmed from differences in naming conventions and ambiguities in the transliteration process. Even seemingly minor discrepancies like this can create complexities in the re-identification process.

Mechanisms aimed at mitigating data friction in the context of re-identification are likely to bring about shifts in associated costs. For instance, one hypothesis could suggest that the circulation of identity data and the presence of data friction is closely linked to the proliferation of duplicate records. Consider the scenario where various organizations share a common database: disparities in identifying individuals, like variations in naming conventions or data formats, could lead to registering multiple entries for the same individual. In this and similar cases, streamlining data frictions by integrating data matching tools into bureaucratic systems simultaneously redistributes costs. Integrating such tools, for example, might necessitate organizational adjustments, as staff may need to allocate additional resources to manage other aspects of re-identification, such as the labor-intensive task of detecting and resolving identity discrepancies.

This section has reviewed literature that has been instrumental in conceptualizing re-identification as an iterative bureaucratic practice, emphasizing its materiality and performativity dimensions. The conceptualization of re-identification as a bureaucratic practice underscores its significance within the interactions of bureaucratic organizations with applicants, particularly as these interactions become increasingly digitized and automated. As a bureaucratic practice, it underscores the potential links of re-identification with the exercise of discretion by bureaucrats in routine re-identification practices at the street-level and screen-level. Re-identification can be further contextualized at the system-level within the realm of the materiality and performativity of devices, ultimately influencing the evaluation of data, individuals, and organizations, shaping their credibility and reliability.

Additionally, delving into the technical mechanisms of data matching not only uncov-

ers operational intricacies but can also serve as a means to gain insights into the diversity of re-identification processes. Integrating data matching tools within bureaucratic systems with the aim of reducing data friction can inadvertently shift the associated costs. As such, the literature discussed here supports the idea that the re-identification of individuals throughout bureaucratic processes and data infrastructures is a crucial but understudied area of research. An iterative approach to identification underscores that re-identification is not a one-off event but an ongoing, multifaceted process spanning diverse bureaucratic tiers. It encompasses street-level interactions, screen-level engagements, and system-level operations, unfolding across both spatial and temporal dimensions. This approach enables us to hypothesize that data friction is an inherent element within these iterative identification processes, potentially leading to less-than-optimal re-identification outcomes that entail organizational costs. The next section will describe the empirical case and the methods used to investigate re-identification.

5.3 Case and method: Empirical analysis of the interplay between data matching systems and applicant re-identification

The investigation into re-identification within migration management draws upon data collected through fieldwork conducted in person and remotely between July 2020 and July 2021. While a comprehensive methodological framework is outlined in Chapter 3, this section offers additional specific details tailored to the context of this chapter. Throughout this fieldwork, I established a collaborative partnership with the Dutch company WCC Group, specializing in developing data matching and deduplication software. In the context of this chapter, the focus will be on the software's use by the Netherlands' Immigration and Naturalization Service (IND). The IND, entrusted with responsibilities such as processing residency and nationality applications, utilizes the ELISE software for searching and matching applicants' identity data within the back-office system and also assists in managing data anomalies, such as duplicate records.

The fieldwork delved into applicant re-identification procedures at the IND, specifically examining their interplay with WCC's "ELISE ID platform." By design, the matching system aims to circumvent errors in both the database and the search criteria. For instance, it can automatically accommodate instances where the date of birth is incomplete or date and month values have been inadvertently interchanged in the search query or database records. Using the ELISE system thus facilitates re-identification when discrepancies may arise from difficulties in matching personal data from different locales, scripts, and cultural contexts. The re-identification of an applicant through a search, based on factors like their name, nationality, and date of birth, results not in a simple roster of exact matches but rather a compilation of applicants with an associated value signifying the likelihood of a match between identity data records. In short, the data matching engine integrates diverse algorithms into a cohesive system that aims to address uncer-

tainties in identity matching, thus supporting the IND's operational processes.

Methodologically, the research considered discrepancies between tool design and practical usage. This was accomplished by comparing how different organizational actors within the IND employ data matching capabilities in their daily re-identification tasks. The study hypothesized that for re-identification to be effective, for IND staff to use the search and interpret the results effectively, there must be some alignment between the system and its users. The research placed a specific emphasis on understanding the challenges faced by IND staff while employing search and match functions in IND's systems, revealing underlying assumptions and expectations of the data matching system. Indeed, discrepancies between intended use and actual application might underlie challenges in re-identification.

5.3.1 Data collection

The data collection process was facilitated by my involvement as a temporary member of WCC's ID team. This collaboration enabled me to access essential technical documentation and engage in informative meetings, including some conducted on-site at the company's headquarters in Utrecht, The Netherlands. The collected technical documents regarding integrating the ELISE data matching system into the IND systems can be classified into three categories. Firstly, some documents cover the ELISE system's overarching technical specifications. Independent of any specific organizational implementation, these documents provided insights into the data matching software's overall design and intended applications. Secondly, a trove of technical documents, meeting minutes, and presentations delved into the precise implementation of the ELISE system within the context of the IND. These resources helped analyze the search and match software integration into the IND's systems. Thirdly, the collected documents encompassed public communications, such as online news aimed at ICT professionals and official reports like government audit findings. These sources contributed an additional layer of context regarding the evolution and establishment of the IND's information system. The initial reading of these documents, accompanied by annotation and note-taking, served as a jumping-off point for structuring the questions to be asked during the interviews with IND staff.

Following the document analysis, I conducted semi-structured interviews to gain insight into the development and use of the search and match tools. As detailed in Chapter 3, interviewees can be divided into two main groups based on their themes and used different protocols for each group. The first group was centered on IND staff — the users, whose duties include looking up and matching identities in their databases, which necessitates using the ELISE ID platform. The analysis in this chapter mainly draws on data from interviews with IND staff and notes from the briefing meetings with WCC ID team members. The second group was WCC staff involved in the software's development, deployment, and maintenance as designers. Although these interviews with WCC staff provided additional context about ELISE, they are not directly featured here and will be

more important for Chapter 6. With the aid of the ID Team, initial contact with the IND was established, yet pandemic constraints necessitated interviews via online meetings or phone calls. In five interviews, each spanning roughly an hour, participants shared their experiences with the search and match tools at the IND.

The interview protocol for IND staff was designed to explore various aspects of re-identification within the IND organization. Interviews began with general inquiries about the interviewees' roles within the IND, providing context for understanding their experiences and tailoring subsequent questions. The initial questions centered around three key factors influencing the searching and matching of applicant data. These factors are how search queries are formulated, the computation of matches, and the handling of search results. Regarding the first factor, questions included IND personnel's approach to formulating search queries, including the data categories they input, their knowledge of data elements that yield better search results, and their utilization of match features such as wild cards. For the second factor, questions were tailored to uncover their expectations regarding the match results and their understanding of the match engine's functionality. For the third factor, the questions probed on their processing of search results, addressing their perception of result quality, the ranking of matches, and the interpretation of match scores. Next to those questions, the protocol inquired about duplicates and the deduplication process, inquiring about the criteria used to identify duplicates and the organization's approach to resolving them. Lastly, the protocol investigated the participants' use of additional systems and data to support the re-identification process. Overall, the interview protocol aimed to provide comprehensive insight into the search and match procedures and the challenges IND personnel face in re-identifying applicants.

5.3.2 Data coding and analysis

For analyzing the fieldwork data, I followed standard methods for coding and analyzing qualitative data. After collecting and preparing data from documents and interviews (including transcription), I coded and analyzed the data using the computer-assisted qualitative data analysis software ATLAS.ti. The data coding and analysis drew inspiration from the three interconnected steps of the "Noticing-Collecting-Thinking" (NCT) method by Friese (2014), which follows a standard qualitative data analysis but is tailored for the ATLAS.ti software. The Noticing phase involved both deductive codes and openness to inductive insights from the data. These codes were reviewed and organized into similar categories in the Collecting step. The third step, Thinking, led to identifying patterns, processes, and typologies among the developed codes. Figures 5.2 and 5.3 provide a simplified overview of this process, illustrating broad deductive themes on the right, more inductive findings in the middle, and representative interview excerpts on the left.

In more detail, the first step of the data coding process began with deductive coding, aligning with the key factors influencing the search and matching of applicant data, as outlined in the interview protocol. These factors encompassed the formulation of search

queries, match computation, and handling of search results. The data coding utilized several of these predefined codes, falling under broad categories like "search query," "search engine," and "search results." This coding method began by applying these predetermined codes to relevant excerpts. However, these codes were subsequently refined through an inductive approach, recognizing patterns within the interview excerpts.

The second step of refining and collecting codes proceeds through adding a colon ":" to the code and names to introduce inductive sub-codes. For instance, "search query: use of data: amount of data available" corresponds to a deductive category concerning the types of data employed in crafting search queries ("search query: use of data"). The inductive aspect ("available data") emerged from the interviews and was consistently used for quotes referencing the quantity of data available concerning formulating search queries for re-identifying applicants.

The third step of the process involved further discerning patterns, processes, and typologies within the developed codes. Two illustrative examples encompassed an in-depth analysis of friction associated with search query input and output, visualized in Figures 5.2 and 5.3. In the diagram illustrating search input, three main challenges were identified, each linked to underlying codes associated with challenges in constructing search inputs due to typographical errors in strings, numbers, and dates; complications in transcribing or interpreting data; and uncertainties about how data should be input (e.g., determining the extent of data to input and which combinations to use). Conversely, the figure illustrating search output captured different broad types of challenges, again connected to more specific underlying codes tied to processing results, including instances where the output yielded an excessive or insufficient number of results, as well as instances where the results were unexpected.

5.4 Exploring the designed infrastructure for identifying applicants

This section starts the empirical analysis by situating the re-identification of applicants[4] at the Immigration and Naturalization Service (IND) within its software architecture and inter-organizational frameworks of the Netherlands' migration policy. Understanding the organizational and software architecture helps situate the intricacies and challenges of the IND's re-identification processes. The IND's operations rely on its information systems called INDiGO, designed to manage the identification and registration of applicants applying for various purposes, including residency or naturalization. Additionally, INDiGO interfaces with various partners and stakeholders within the migration chain

[4] In this chapter, the term "applicant" is employed to refer to the individuals that submit formal requests or applications to the government agency. However, the prevalent term used by interview participants from the agency in Dutch is "klant" or "cliënt." This term would translate to "client," a person receiving the benefits or services of a government agency. Interestingly, the first term could also be translated to "customer." I have chosen to utilize "applicants" to refer to these individuals as "applicant" emphasizes the act of making a request or application, whereas "client" emphasizes the recipient of a service or assistance.

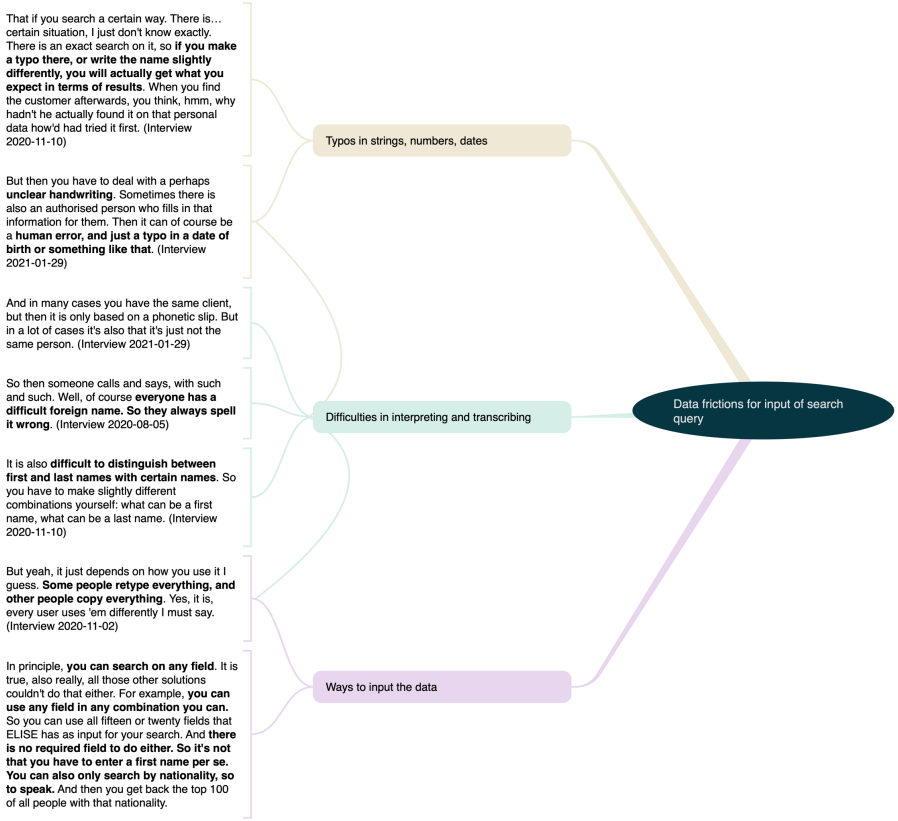


Figure 5.2: This diagram shows how friction with search query input were found by analyzing interview data.

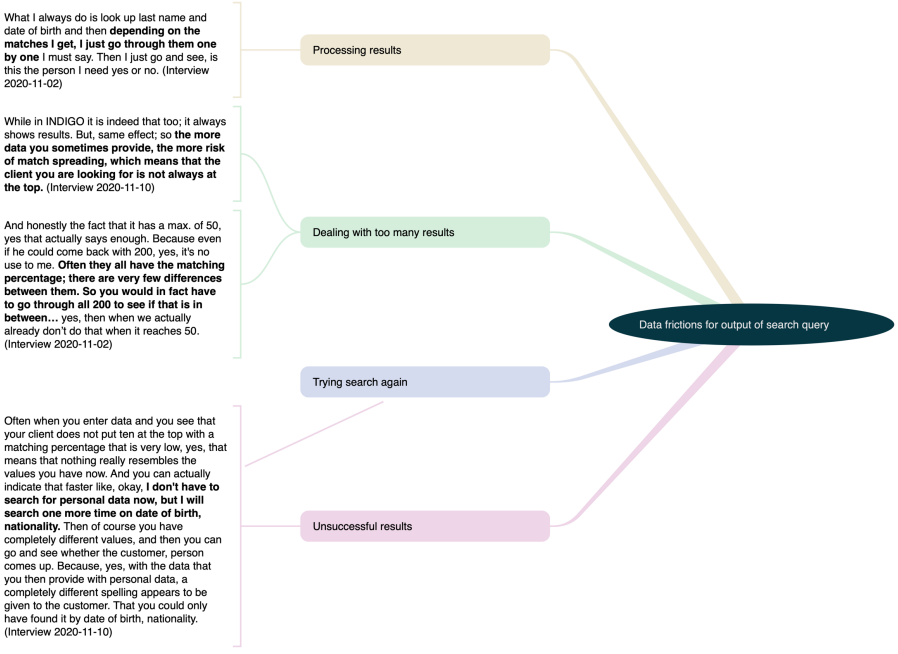


Figure 5.3: This diagram shows how friction with interpreting search results were found by analyzing interview data.

whose systems and databases also play a role in the IND's processes. This initial examination of the organizational and software architecture will provide the foundation for the following sections, in which this architecture and software designs will be compared to their practical implementation.

5.4.1 Migration chain and identifying chain partners

The processes and practices involved in re-identifying applicants for the IND need to be understood within the larger context of the information infrastructure for handling foreigners in The Netherlands. The IND is just one link in the "migration chain" (*migratieketen*), a collaboration between various governmental and non-governmental organizations in The Netherlands. Each link in this chain, known as a "chain partner" (*ketenpartner*), is responsible for different processes foreign nationals in The Netherlands go through, including entering the country, obtaining a residence permit, naturalization, and departure or expulsion. These partners exhibit interdependence since their decisions often necessitate information from others, facilitated by an interconnected information infrastructure.

The information infrastructure of the migration chain can be traced back to a subfield of information science called "chain computerization" or *keteninformatisering* in Dutch, which has been influential in Dutch academia and government digitalization in The Netherlands (e.g., Grijpink, 1997). Chain computerization pertains to the information infrastructure of the networked chain of interdependent organizations without a formal hierarchy (Oosterbaan, 2012). These entities collaborate and exchange information to execute a shared process, exemplified in this context by the handling of foreign individuals within the Netherlands. The migration chain, as outlined in Zijderveld et al. (2013) that describes its architectural framework, defines principles and objectives aimed at enhancing information exchange. It addresses identification challenges, particularly those contributing to duplicate registrations among chain partners, and constitutes an important focal part within the architecture.[5]

Foreigners within the migration chain are assigned a unique identifier known as the "v-number," which is used for identifying individuals throughout the chain. This unique identifier is issued through the *Basisvoorziening Vreemdelingen* (BVV) system, functioning as a centralized repository for sharing and consulting information about foreign nationals among the various chain partners. Upon the first contact with a foreign national, the relevant organization must ensure that the individual has not been previously registered and, consequently, already possesses a v-number. The BVV database can be updated and enriched with identity data, travel information, identity documents, biometric characteristics, and status data, such as asylum application outcomes, originating from the

[5] A new architecture called MIRA is being developed for the migration chain. Unlike the current approach of automating data streams between chain partners, this new architecture suggests creating an information platform that offers data services for chain partners to use in achieving their processes, services, and systems (Ministerie van Justitie en Veiligheid, 2023).

chain partner's systems. Even though every partner in the chain has their database and information regarding migrants, the use of BVV and the v-number can enable the linkage of information by utilizing the v-number as a shared and unique identifier (ICTU, 2015).

The processes of first registration and identification of foreign nationals are further directed by the "Protocol identification and labeling" (PIL, "Protocol identificatie en labeling" in Dutch) The various chain partners use this protocol; it standardizes the process of identifying and registering foreign nationals as a way to ensure that "unique, unambiguous personal data of optimal quality are available in the migration chain" (Ministerie van Justitie en Veiligheid, 2022, p. 9). As implied by its name, the protocol also includes a labeling provision. If someone is hesitant or unwilling to reveal their name, they will be assigned a label. This label will serve as their name and contain details about their gender and the date, time, and place of registration. For example, the label could be *NN regioncode sex yymmdd hhmm*. Therefore, the protocol can be interpreted as being designed to minimize identity multiplicity and streamline subsequent re-identification by providing clear guidelines for recording individuals' data. Nonetheless, as elucidated later in this chapter, the presence of individuals possessing multiple v-numbers indicates that this ideal scenario may not always hold in practice.

The key takeaway from the architecture of the migration chain is that the unique identification of individuals is deemed crucial for re-identifying foreigners in The Netherlands among all the chain partners. Specifically, the IND, as one of the chain partners, relies on the BVV systems and v-number for effective identification and re-identification of applicants. Through this re-identification process, the IND is tasked with confirming whether an applicant has not already been initially registered by other chain partners such as the "Vreemdelingenpolitie" (national police) or the "Koninklijke Marechaussee" (national gendarmerie). Next, we will delve deeper into the systems employed by the IND and explore further their interactions with the BVV and other chain partners.

5.4.2 Unpacking the IND and INDiGO infrastructure

Shifting the focus from the broader discussion of the identification of foreign nationals in the migration chain, this section delves into a more specific examination of how applicants are identified in the information systems employed by the IND. The central pillar of this information infrastructure for application and identity management is the INDiGO system.[6] The implementation of the INDiGO was part of a more extensive digitization project and data transfer from a previous system called INDIS. The distinct manner in which the upgraded system technically compartmentalizes various facets of organizational operations is of particular significance. This division is primarily manifested in the separation of policy implementation, which involves the application of business rules

[6] The system is based on Oracle's Siebel system, a multinational computer technology company. While it is interesting to consider how this generic Siebel case management system has influenced the IND's operations, it is out of scope for this discussion.

aligned with the Dutch Aliens Act, from information management tasks, including data storage, searching, and matching (KPMG IT Advisory, 2011).[7]

The information infrastructure of the IND, as outlined in Figure 5.4, can be characterized as a form of system-level bureaucracy (Bovens and Zouridis, 2002), given that INDIGO places significant emphasis on information management and the automation of decision-making in the processing of digital dossiers.[8] At the system level, the identification of applicants unfolds through automated data exchanges, where applicants are re-identified in processes to update their applications and dossiers. Re-identification also extends across multiple bureaucratic tiers, spanning the street- and screen-levels. At these levels, IND staff, both at the front and back offices, interact with the graphical interfaces of the information systems to confirm and verify applicant identities while processing their applications. The following section will examine how INDiGO utilizes the ELISE software for searching and matching applicant data across all these bureaucratic levels.

5.4.3 Applicant re-identification and matching with ELISE software in the INDiGO system

Throughout the evolution of INDiGO, the ELISE software for searching and matching applicants has been applied in various scenarios, which can be categorized into the following three cases. First, the ELISE software was initially used during the transition from the old IND information system (INDIS) to the new INDiGO. During a transition period when INDIS and INDiGO ran in parallel, the IND used the ELISE software to migrate legacy data by re-identifying matching applicant identities between the two systems.[9] The ELISE software's second and most prominent use is to facilitate applicant data searches. While the underlying case management system provides a "traditional" search, this was deemed insufficient because it would fail to return results when search criteria are too strict or contain errors. For this reason, the ELISE system was added to use the software's fuzzy search algorithms to provide more advanced and reliable searching capabilities (Interview 2020-08-05). Third, the ELISE software searches the database for possible duplicate applicant data. The software attempts to match all recently created applicants to all other applicants in the database. Potential duplicate matches that meet specific criteria will then be flagged and investigated further. In all three uses, the software calculates match scores calculated by the software based on the likelihood that an

[7] INDiGO documentation refers to this as the "flow" (information management) from the "know" (policy implementation). Technically, this separation is accomplished by adopting a Service-Oriented Architecture (SOA) design, a software architecture strategy that aims to modularize system functions into relatively independent services.

[8] The architecture of the INDiGO system received praise during its initial development in 2009 (Toet, 2009), but the rollout of the new system in subsequent years did face several delays and problems (Bergsma, 2013).

[9] Technically, this was made possible by running the software's algorithms on data that has been *replicated* from both INDIS and INDiGO (i.e., regularly copying data from both sources into the ELISE in-memory database). The initial rollout of the INDiGO system started in 2009 and was completed in 2013.

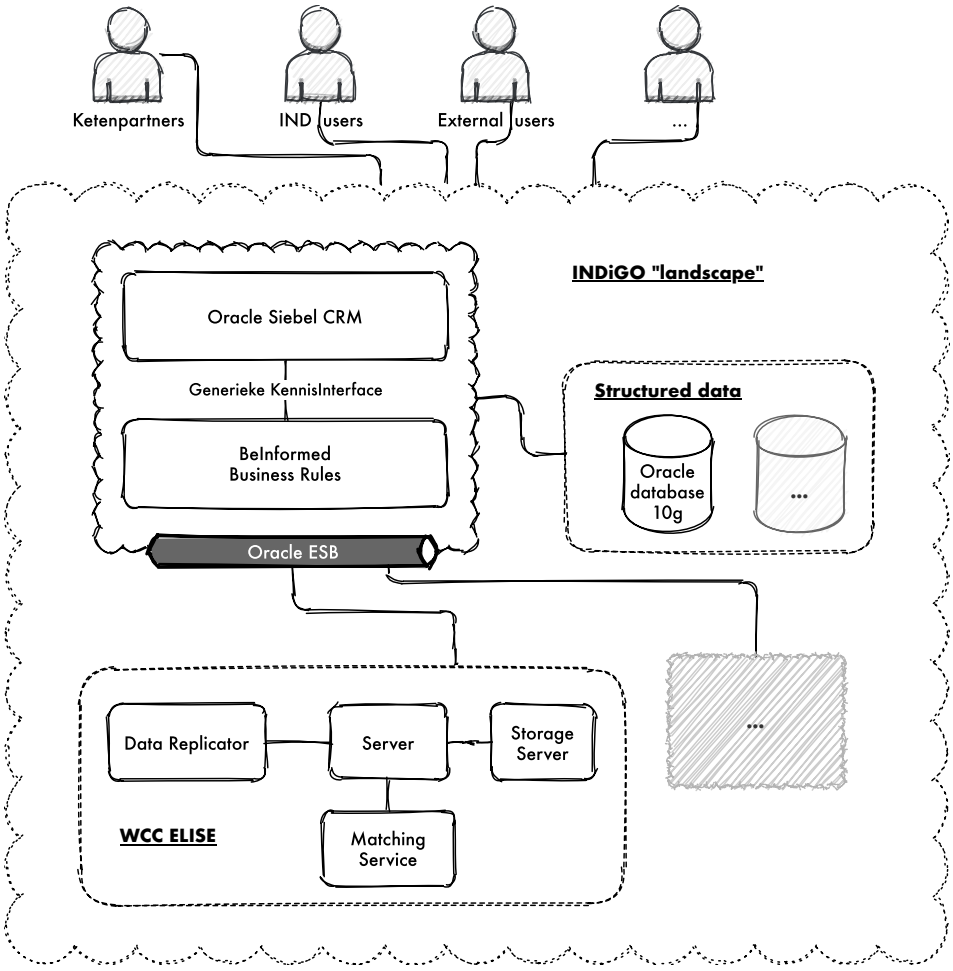


Figure 5.4: A schematic representation of the IND's information infrastructure, illustrating its role in facilitating tasks related to application interactions, and inter-organizational collaboration with MC partners.

applicant in the database meets the given search criteria.[10]

Following my analysis of the technical architecture and utilization of ELISE in INDiGO, as detailed in design and technical documents, I suggest categorizing the searching and matching into three essential components: query input, processing by the matching engine, and results output. Designed as a generic and decontextualized component within the INDiGO system architecture, ELISE is intended to function independently, receiving input from various sources. This input might originate from the INDiGO graphical user interface or other automated processes. According to the documentation, the system is designed with the recognition that both search queries and database records can contain errors. This design thus accommodates scenarios such as typos, the inadvertent swapping of first and last name fields, and similar errors, whether they are already present in a record within the database or introduced during the formulation of a search query. Within INDiGO, the search and match functionality operates without specific user distinctions, instead relying on the ELISE system as a data matching service configured universally and integrated into various components of the INDiGO system.[11] This prompts the question: what implications do this absence of user-specificity have for the re-identification practices of the IND?

Per the system's technical documentation, queries originating from IND end-user applications are channeled to the ELISE service, which employs diverse algorithms to compute matches. In practice, the data matching engine assesses the similarity between the input query and all database records, generating a corresponding similarity score. The computation of this match score can be adjusted through system configurations, allowing certain factors to be weighted more or less significantly. The matching process encompasses deterministic data matching algorithms that calculate similarity by considering variations in name spellings, utilizing methods such as name initials and even intelligently accounting for transposed numbers, evident in dates of birth or identification numbers. Additionally, the engine leverages name data databases to facilitate advanced matching techniques based on rule-based or domain knowledge, such as accommodating name transliterations and recognizing variations like "Aleksandra" and its diminutive form "Ola." Furthermore, the system incorporates probabilistic matching mecha-

[10] It is noteworthy that initially, INDiGO utilized ELISE's fingerprint-matching capabilities to verify whether an applicant was already known, achieved by comparing their fingerprints during their first registration with the entire database. However, this "one-to-many" matching practice is no longer in use. Fingerprint data now exclusively serves to confirm an individual's identity, particularly during face-to-face interactions with applicants at service counters. A simplified process involving distinct software is employed for this purpose, enabling "one-to-one" fingerprint matching. Although biometrics play a role, this chapter's primary focus revolves around backstage re-identification and the utilization of non-biometric data. This emphasis corresponds with the identified gap in the existing literature.

[11] This also influenced by the way INDiGO utilizes the ELISE system. ELISE has the capability to dynamically adjust the behavior and importance of search criteria for specific services, which isn't the case in the INDiGO implementation. The GUI lacks the feature to configure this for queries, and it seems that the criteria used in data matches are not displayed either. This is likely a result of balancing user-friendliness, configurability, and complexity of implementation.

nisms, including a feature termed affinity matrices, which involve attributes like the "soft matching" of birth years within a reasonable range. For example, if a search specifies a birth year as 1990, the system can be configured to consider birth years slightly earlier or later, covering a span like 1988 to 1992. By employing these diverse matching features, the system assigns a match score that gauges the likelihood of a match between the search query and the corresponding database entry. These system functionalities raise questions about the influence of ELISE on the re-identification expertise of IND personnel, potentially shifting the locus of expertise from street-level bureaucrats to the system.

The system returns a set of records ranked based on their closeness of match to the query, as opposed to offering a single match in response to a database lookup. By design, the data matching process always yields results, even if no exact match is found. The number of matches returned is also adjustable within the system. It is important to note that, due to its modular structure, the results are sent back to the point of origin of the search, such as being displayed through the INDiGO graphical user interfaces. Consequently, the searching and matching service has limited insight into where within the INDiGO system and process the call is initiated or who is making the query. Subsequent sections will explore how this user-agnostic approach aligns with the actual usage patterns for the IND's re-identification practices.

5.4.4 Architectural and system design influences on applicant re-identification at the IND

This section has described architectural and system design elements of the IND composite information environment, and highlighted how they can influence the re-identification of applicants within the IND. The discussion has delineated two main dimensions that shape the IND's data infrastructure and subsequently impact the re-identification process.

Firstly, operating as a node within the migration chain, the re-identification practices of the IND are not isolated but rather interconnected with the diverse partners comprising the migration chain. During the initial interaction with an applicant, the agency is tasked with validating whether the individual is already within the records of the migration chain. The v-number, functioning as a unique identifier for foreign nationals in the Netherlands, facilitates the process of re-identification and linkage of applicant data across the chain partners. However, re-identifying applicants will become considerably more intricate when this identifier is unavailable. The introduction of mechanisms like the PIL strives to establish standardized initial registrations to facilitate smoother re-identification in subsequent stages. In the following section, we will examine the practical aspects of these interactions and delve into the complexities of re-identification, including the interaction between different systems and the challenges faced by IND personnel.

Secondly, the IND's strategies for re-identification are intertwined with the ELISE data matching system. This system acts as an intermediary in the re-identification of

applicants, addressing various uncertainties surrounding identity data. Positioned as a loosely coupled module within the broader INDiGO system and its accompanying databases, ELISE aims to facilitate the process of applicant re-identification. The system's design acknowledges the inherent uncertainties in search queries and database accuracy. In the upcoming section, we will delve into the empirical data gathered from fieldwork, elucidating three distinct types of data friction that can hinder re-identification, stemming from disparities between the intended designs of systems and their actual practical utilization.

5.5 Putting the design into practice: Investigating the practical application and challenges in the IND's identification processes

5.5.1 Friction 1: Navigating diverse re-identification approaches

This section explores scenarios that emerged during the interviews, which are representative of challenges of re-identification within the IND and the migration chain. Firstly, we look at how verifying an individual's existing record within the migration chain works in practice by focusing on how IND personnel consult and link data from the Basisvoorziening Vreemdelingen (BVV) system. Next, a real-world example will be presented, highlighting a particular challenge in re-identification regarding the automated information processes for residency updates between municipalities in the Netherlands and the IND. Together, these scenarios illuminate a first form of data friction that can emerge between standardized identification practices and the idiosyncrasies of institutional procedures and data matching technologies. This finding underscores the complexities involved in re-identifying applicants across different administrative entities.

The following illustrative interview quote from an IND staff member offers an insightful glimpse into the broader process, providing a representative portrayal of the general approach to re-identification through the BVV system. As mentioned previously, chain partners such as the IND must ensure that the individual has not been previously registered and already possesses a v-number. INDiGO thus allows personnel to search for matching personal data on the BVV system. The interview quote's context pertains to the IND's procedures upon receiving a new application, such as one for a residency permit. In this scenario, the initial phase entails confirming whether the applicant's presence is already recorded within the migration chain, necessitating linking their data to the BVV system. This verification process also entails checking the applicant's status in the IND system and determining whether existing applicant files need to be updated or new ones created. As the interviewee remarks, this process involves scrutinizing the applicant's personal data against the BVV records. When disparities arise between the information stored in the IND's database and the BVV, these discrepancies are duly noted for future investigation and resolution. Here is how the interviewee describes the process:

> We actually search first on the system called BVV [...] We click on a button, and then a search is made for the personal details that then appear. If we have a hit, it means, for example, that either the Royal Netherlands Marechaussee, Foreign Affairs, or the police have ever registered the applicant. Well, then, the data only occurs on the system called BVV. And if so, well, we'll make a link. Then we click on a button, and then there is a connection between the data from the BVV with the data we have received from the municipality. And if that is not the case, for example, you can find the applicant in the BVV and our IND system. That's when you press another [search] button. And when it turns out that the applicant appears in the BVV and the INDiGO system. Well, then we check in the INDiGO system whether the names match completely, for example. In case of small changes in the name data, we also look further into the file. And if we do come to the conclusion that "this is the same person," then we also make the connection, so we register the applicant. We link the data together. Well, then you only have one applicant file, and then nothing is wrong. However, when there is a difference in personal details, for example, we have to report this in the system. (Interview with IND staff member, January 29, 2021)

The interview quote highlights two insights regarding the connections with the BVV system for applicant re-identification. At the outset, although my initial inquiry aimed to solicit an account of the applicant data retrieval process at the IND, the interviewee's response accentuates a primary phase wherein cross-system validation assumes precedence ("we *actually* search first on the system called BVV," emphasis added). Thus, the interview quote effectively emphasizes an ordering in which different data sources assume priority in the re-identification of applicants. The hierarchical sequence underscores the necessity of referring to the most authoritative source early in the re-identification process to streamline re-identification and minimize errors. However, this ordering raises questions about how conflicting data from different sources is handled and which source takes priority in resolving such conflicts.

A second observation from the interview excerpt concerns the associations with the BVV system in the context of IND applicant re-identification. It revolves around establishing and maintaining links among applicant data within these distinct systems. As elucidated by the interviewee, once corresponding identity data within the BVV and IND systems are identified, IND personnel can utilize designated buttons within their user interface to establish links between applicant data across these systems. However, this process is not entirely streamlined or automated. As connections materialize between the systems, the interviewee underscores the manual effort required to ensure the alignment of applicant data. They point out instances where disparities in personal data between BVV and IND records emerge and necessitate reporting. Despite standards like the Protocol Identificatie en Labeling (PIL), which aims to ensure consistent registration of personal data, challenges may still arise due to these variations. Let us examine a real-

life scenario to understand how these discrepancies can manifest.

The following re-identification scenario relates to a situation wherein a migrant initiated a residency application with a municipality. Subsequently, the municipality employs an automated data exchange mechanism to notify the IND to update the applicant's particulars within the IND's systems. However, as the interview quoted below demonstrates, a challenge arose, as evidenced by the failure of the automated message exchange and unsuccessful re-identification. The problem described by the interviewee is characterized by a complicated interplay of naming conventions, identification practices, and policies for registering individuals. In this example, the problem is a divergence between the IND and the municipality regarding how an individual's registration is established - whether based on their birth certificate or passport data. Here is how the interviewee describes the problem:

> In principle, the municipality only registers applicants who submitted such an application [for a residence permit] to the IND. A condition for registering with the municipality is that applicants must identify who they are. So that can be done, for example, with a birth certificate, a copy of a passport, or an identity document, or other documents, so to speak. The municipality does have a different kind of policy on identification than, for example, the IND. They have a different ranking of pieces that they, well, consider important to have.
>
> For example: we — the IND — see a copy of a passport sufficient, or an ID card, or even a laiser-passer. The last one is a kind of document issued by the embassy if the applicant does not have a passport or ID card. But the municipality, [...] the most important document to register someone is actually a birth certificate. And then you sometimes have differences because, for example, applicants from, well, for example, from Ukraine. They have, say, a name and then a patronymic. That [patronym] actually refers to the name of their father. And then the family name. And, well, that patronym is often included in the registration by the municipality.
>
> But the IND, on the other hand, does not necessarily register based on the birth certificate data; because those data were once given at birth, but of course, they may have changed after many years because it is possible, by the way, that you take your marriage name, for example. So, if the applicant submits a passport with the marriage name, the IND will register the applicant based on the passport data. While the municipality uses the birth certificate data. So you already have a difference. And we may then receive an automatic message [from the municipality], which the system cannot automatically link to an applicant. (Interview with IND staff member, January 29, 2021)

This example is a poignant illustration of the challenges confronted by the IND in re-identification processes. These challenges become particularly evident when the agency attempts to ascertain whether an applicant is already known to the agency or in the migration chain. For instance, naming conventions like variations in spelling, name order, or the inclusion of middle names or initials can make re-identification more complex. Additionally, there may be differences in institutional practices, where certain documents or data categories are given more significance, which can further complicate re-identification.

Conceptualizing these challenges as a first form of data friction introduces an additional theoretical layer to our understanding of the re-identification concept. In this context, re-identification goes beyond mere technical exercise of matching and linking data; it manifests a negotiation between diverse identification practices. Re-identification assumes an additional theoretical dimension, reflecting practices through which entities like the IND act as mediators within the web of bureaucratic operations. The illustrative example highlighted this dimension by demonstrating how IND staff engage in re-identification tasks. They need to identify the source of discrepancies in the identity data, confirm whether these data pertain to the same individual, and update their residency status. This process ensures that the information about the individual remains accurate and up-to-date within the context of the IND and the broader migration chain. This mediating role requires aligning multifaceted and sometimes disparate identification practices. The upcoming sections will investigate further re-identification at the IND through its use of tools to search and match applicants.

5.5.2 Friction 2: Balancing precision and accuracy in the movement of identity data

This section, along with the next one, will center on three primary aspects of the search and matching process hypothesized to play a role in the successful re-identification of applicants: the formulation of search queries, the calculation of matches, and the handling of search results. This division aligns with the interview protocol's structure, as elaborated in the section on data collection. The analysis will commence by exploring the challenges associated with formulating search queries. As emphasized in the design discussion, the absence of user-specificity in the search and match functionalities design means that these functions are intended to operate uniformly and generically for all users within the IND, without distinct configurations or adaptations tailored to specific user roles or preferences. This design choice might have implications for re-identifying applicants at the IND, as it raises questions about how effectively the system can cater to the diverse needs and practices of different users and departments involved in the re-identification process. The interview excerpts and scenarios presented in this and the subsequent sections are selected as they are deemed the most illustrative of potential data friction that could impede smooth re-identification when using the search and matching tools, as revealed through the analysis.

To start, it is essential to acknowledge that for respondents the lack of user-specificity

is not a prominent concern, primarily due to the simplicity of formulating search queries. Participants underscored that the most fundamental and frequently utilized method for searching and re-identifying applicant data within the IND adopts distinct identifiers like the v-number. For instance, applicant re-identification might occur while processing new information related to an ongoing application in back-office settings or direct interactions with applicants at front-office counters. In these cases, the procedure is ongoing at the IND, and staff can thus execute such searches on the system relatively seamlessly, as the applicant is already known to the IND and the v-number can be employed.

However, as highlighted in the subsequent interview excerpts, IND staff members frequently encounter situations where formulating seemingly straightforward search queries becomes complex due to various factors. These scenarios often involve dealing with intricacies stemming from data inaccuracies, misunderstandings, and human errors, especially when data is received or input through diverse means like handwritten documents or phone conversations. In specific departments, such as those handling handwritten documents sent via postal services, employees often grapple with accurately deciphering and transcribing handwritten data, consequently introducing an element of uncertainty into the re-identification process. To illustrate, the following example from an interviewee sheds light on the challenges arising from processing handwritten application forms and managing errors or ambiguities in the provided date of birth:

> But then you have to deal with perhaps unclear handwriting. Sometimes there is also an authorized person who fills in that information for them. Then it can, of course, be a human error and just a typo in a date of birth or something like that. (Interview with IND staff member, January 29, 2021)

Another instance of this issue involves the potential confusion between first and last names. IND personnel may encounter situations where the data provided by applicants does not clearly distinguish between these two components. This ambiguity can create uncertainty for the IND staff when using the search fields in INDiGO, as highlighted by the following interview excerpt:

> It is also difficult to distinguish between first and last names with certain names. So you have to make slightly different combinations yourself: what can be a first name? What can be a last name? (Interview with IND staff member, November 10, 2020)

There are instances where the perceived match between applicants might originate from phonetic errors or other subtle variations. For instance, envision a situation where a name like "Rousseau" is misheard or misspelt as "Russo," resulting in what the interviewee below refers to as a "phonetic slip":

> In many cases, you have the same applicant, but it is only based on a phonetic slip. But in a lot of cases, it's also that it's just not the same person. (Interview with IND staff member, January 29, 2021)

As previously outlined, the ELISE data matching system is intentionally designed to address such potential errors in search query inputs, aiming to account for possible typos, mixed names, and similar variations during matching. Nonetheless, when I inquired about this aspect during the interviews, there appeared to be a general lack of specific knowledge about how this mechanism functions for search query inputs. However, as noted in the following excerpt, the mechanism that accounts for search input could occasionally lead to confusion in the search results – a subject we will revisit in our subsequent discussion regarding the computation of matches and the handling of search results.

> [I]f you search a certain way. [...] I just don't know exactly. There is an exact search on it, so if you make a typo there or write the name slightly differently, you will actually get what you expect in terms of results. [But] When you find the applicant afterwards, you think, hmm, why hadn't it actually found it on that personal data I had tried first? (Interview with IND staff member, November 10, 2020)

The formulation of search queries can become more ambiguous and error-prone, as evidenced by the excerpts from IND staff members. These issues are particularly notable when dealing with phone calls or handwritten documents, where elements like v-numbers, dates of birth, or applicants' names might contain typos or be challenging to read. In these contexts, data friction arises as identity data transitions between different material forms, arising from difficulties in comprehending or accurately transcribing information. Consequently, this can hinder the successful re-identification of an applicant. Interestingly, this challenge aligns with the design of the data matching system, which anticipates potential errors in search queries and compensates for such uncertainties. Here we can see how the data matching system thus functions as a mechanism to alleviate data friction as identity data shifts across various media. However, there is a potential disconnection between users' expectations of search input accuracy and the system's ability to accommodate errors and uncertainties in the query.

The findings on the formulation of search queries contribute to our understanding of re-identification in two ways. Firstly, they underscore the distinction between different forms of data employed in the re-identification process. While certain data, such as identification numbers like v-numbers, demonstrate a higher capacity to traverse various material forms and actors, personal data like names and birthdates are more susceptible to discrepancies, errors, and interpretive challenges. Nevertheless, both data types are vulnerable to errors, whether through typos, ambiguities, or variations, potentially complicating the re-identification process. Secondly, the findings indicate a disparity between the system's intended flexibility, designed to accommodate errors and

uncertainties in search queries, and users' assumptions regarding the accuracy of their inputs. Users may expect accurate input, while the system automatically rectifies any errors they make in their queries. This disparity can create friction in re-identification, and introduce frustration and inefficiency into the re-identification process when the system includes unanticipated results. This finding underscores that while the mechanism designed to mitigate data uncertainty in formulating search query inputs is effective, it can inadvertently lead to ambiguities, particularly when users are unaware of its functions.

Another aspect of search formulation that warrants investigation is the amount and combinations of input data required for successful re-identification. While the ELISE data matching system is designed to operate optimally when the search query includes as much information as possible, the interviews also probed into the specific combinations and amounts of data actually utilized in search queries. However, this aspect is intricately interlinked with the expectations surrounding search computation and the subsequent results, which will be discussed in the subsequent section.

5.5.3 Friction 3: Deciphering opaque match results for successful re-identification

Continuing the investigation of the three aspects thought integral to the re-identification of applicants through the search and matching process, this section now delves into the practical complexities associated with match calculation and the subsequent handling of search results. By delving into these aspects, we can address the questions and hypotheses that emerged during the discussion of the system's design. By looking at these two aspects, we can examine the implications of ELISE's functioning on applicant re-identification, considering its potential to complement or replace the re-identification expertise of IND personnel. Additionally, we can analyze how the presentation and handling of search results can either facilitate or hinder the re-identification of applicants.

In general, interviewees showed an awareness of these data matching features during their searches. However, there was also a prevailing sentiment that additional features might operate beyond their explicit knowledge. When probed about more advanced search functionalities, one interviewee indicated that while users may not fully comprehend the intricacies of the search process, they can typically accomplish applicant re-identification. The following quote captures a common sentiment among interviewees — a sense of uncertainty regarding the exact calculation of the match score:

> [N]ow and then it is very hazy how [the search] exactly works. For example, sometimes a letter seems to be more important than at other times, depending on where it is located. But you usually see if you misspell such a letter that you then do not get a hundred per cent hit. But then you still get sixty per cent or so. In some cases, it is also higher than that percentage. But you usually find [the applicant]. (Interview with IND staff member, August 5, 2020)

As mentioned previously, the ELISE data matching system is designed to work best when the search query includes as much information as possible. By doing so, the data matching algorithms can utilize all this input data to calculate match scores. However, as the comments below demonstrate, there was a mixed sentiment among interviewees regarding the usefulness of providing more input data. While some perceived that additional data did enhance results, for others, it did not necessarily guarantee improved outcomes:

> If you have more data, you look at what more you can put in it. So you're actually trying to make it as broad as possible. If you have a date of birth, you have a street name, or you have something else, to increase the matching percentage. And then you actually also look — if there are multiple search results — then you actually look first at the highest matching percentage.

> My experience with searching for personal data is that the more data you enter, the more difficult the result will be. And the worse the result actually gets. So I often build it up. I [input] less data, and if necessary, I add some data if there are too many results. (Interview with IND staff member, November 10, 2020)

One interviewee aptly described this situation, stating that INDiGO staff sometimes feel compelled to "play" around with the search tools until they successfully identify the desired applicant.

> [...] what you often see in how they work is that hey, they use it first with one type data. And if they still get too many results, or they don't see it, they try with an extra piece of data. Or they try it with another kind of data. So you see, to find a person, they sometimes do five searches in a row. Also, a little, OK they could enter everything at once, but you can see they play with that a little bit. (Interview with IND staff member, August 5, 2020)

The interviewee's statement hints at the possibility that IND personnel engage in various permutations of data categories when conducting searches. On the one hand, this highlights a tension between the expected use of ELISE, which encourages providing as much search input as possible, and the practical experience that entering excessive query information might occasionally worsen results and introduce ambiguity in re-identifying the correct applicant. However, another interviewee sheds light on an alternative approach to mitigating uncertainty in re-identification. This method involves experimenting with different combinations of data or deliberately omitting certain information. Remarkably, by not inputting certain details, these omitted elements can later be utilized to cross-check results, potentially reducing uncertainty in re-identification. The interviewee outlined their process as follows:

> And there is also a kind of self-check in [the search process]. So I often search by first name, last name; to start with. [...] But I often try not to do too much and see if that result is there. And on that basis, OK, the date of birth also matches the date of birth that I have. So I don't always deliver what I have available as information. But I also partly use it as a checkpoint for the search result that I then get to the top. It also works a bit more efficient for me. Because it makes no sense to enter much more data. Because you can find the applicant anyway, also sufficient on the basis of first and last name. And you get insight so that you immediately know that you have the right one. (Interview with IND staff member, November 10, 2020)

Upon sharing this scenario with a senior developer from WCC, their surprise at the idea of deliberately withholding input data to reduce uncertainty underscored the friction that can emerge between the intended design of a system and its actual real-world application. For IND personnel, the data matching process can sometimes feel like a black box, amplifying the uncertainty around determining the optimal input of information into the data matching system. Choosing the appropriate amount of information to provide, what to include or exclude, and the potential consequences of different combinations requires additional efforts in the re-identification process, pushing staff to allocate more time and resources to refine their searches.

The complexity surrounding the functioning of the matching process came to the forefront when matches were generated utilizing information stored in "historical fields." These specialized fields exist within the IND's data model to accommodate the storage of multiple values for the same data category, thereby allowing for the retention of various historical information about an individual. This aspect was highlighted during interviews with multiple participants who grappled with comprehending the logic behind including certain applicants in the match results. The confusion arises from the fact that only the most recent values from these fields are considered in the search results. To illustrate, imagine a scenario where a search is conducted using an individual's pre-marital name, yet the results display an applicant with an entirely different post-marital name due to a subsequent name change. An interviewee aptly captured the complexity of deciphering matches based on historical data in the following manner:

> And what is actually very interesting in [the case of the IND] is that someone does not just have one address but can have several addresses, for example, or even several names. And he may have changed his name, for example. So then the old name is also saved. You actually have a primary field, for example, for name or address. And you have historical fields. And they are all searched with ELISE. [...] So we actually have the history of every field. That can contain one value, but it can also contain ten values. And if you match that. I think that there is also an interesting point with user expectations.

> […] I think they're not always aware of that. That if they find someone, it can also be based on an old date of birth, which has been entered incorrectly, or based on an old name. (Interview with IND staff member, August 5, 2020)

When specific categories within the IND's data model contain multiple values, it can both facilitate and hinder re-identification. On the one hand, this approach makes it easier for the agency to re-identify individuals by considering both their current and past data. However, this flexibility also poses a challenge. Scrutinizing matches using historical data becomes more complicated, adding to the workload of personnel. During an interview, one interviewee shared that they acknowledged that such matches are likely to happen and emphasized the importance of carefully examining the results to understand why the match was included:

> […] experience has taught me that often the name has been changed. So there is also a history; so if someone has a different name, if a name is changed — and that is sometimes changed considerably — then you will indeed get it as matching. And when clicking through on the history of the name, you see that the logic comes from there, that it knows from history that it was called differently. And that is why it is shown. That is my experience. (Interview with IND staff member, November 10, 2020)

Likewise, a comparable challenge arises when matching relies on interconnections between applicants or other entities. The INDiGO system also records data about various entities, such as lawyers representing applicants or affiliated organizations. Hence, as elucidated by the subsequent interviewee, an effective strategy for re-identifying an applicant entails exploring linkages within the system. Their strategy also involves searching for applicants registered under a specific affiliated lawyer if a direct search for the applicant proved unsuccessful:

> Then you try to see if you cannot find another way to find the applicant. It may sound a bit strange, but sometimes you can see […] to which lawyer it was submitted, via which lawyer it was ever submitted. And then you can look through the lawyer and which applicants he has under him. That way, you can also indirectly find out which applicant it is. But that is the difficulty when it comes to finding applicants. (Interview with IND staff member, November 10, 2020)

By design, the top results with the highest match percentage should ideally present the most relevant matches, making the matching applicant readily identifiable among these top results. However, in practice, this ideal scenario is not always realized. Responses from the interviewees reveal that users have developed various strategies to decipher results, particularly in distinctive cases. As one interviewee succinctly put it, IND staff are, at times, "actually trying to fine-tune [the search] so [they] can get the right

person up" (Interview 2020-11-02). This phrase highlights challenges in re-identifying applicants and ensuring their appearance among the foremost search outcomes. In this context, it is reasonable to consider that re-identification issues emerge when a disparity arises between the user's intended goal (successfully re-identifying an applicant) and the actual outcome presented by the system (a collection of search results). These difficulties introduce a third form of data friction in the re-identification process, stemming from the interplay between the system's ranking of search results and the efforts of IND staff to re-identify and retrieve applicants' data. This friction is effectively demonstrated through the struggle to retrieve an applicant's data and ensure its prominent positioning within the search results.

These findings on the practical challenges of calculating matches and the subsequent handling of search results further add to our understanding of re-identification. The somewhat enigmatic nature of result calculation highlights a nuanced interplay between human interactions and the capabilities of data matching tools that both facilitate and hinder re-identification. While the interviewees exhibit a certain degree of understanding regarding aspects of the data matching system, such as basic fuzzy search techniques, it becomes apparent that the system employs additional features and autonomous functionalities that often operate beyond their explicit awareness. On the one hand, these features can streamline re-identification efforts, even if the users are not fully aware of the underlying mechanisms. Conversely, disparities in understanding can introduce an additional layer of intricacy, prompting staff to invest substantial time and resources in refining, comprehending, and critically assessing search results to ensure accuracy and dependability. As we will see next, the repercussions of failed re-identification extend beyond these challenges, often triggering issues such as creating duplicate records, which will be explored in the next section.

5.6 The costs of failed re-identification: Duplicates and the labor of deduplication

As established in the preceding sections, the interplay between the design of standardized identification practices and the utilization of data matching technologies can introduce forms of data friction in the re-identification process. This section will explore their broader consequences, focusing on a twofold outcome of unsuccessful re-identification: the presence of duplicate records for applicants and the ensuing labor-intensive deduplication process.

5.6.1 The deduplication memo and resolving identity duplicates

As highlighted previously, the issue of duplicates arises when an organization has multiple disjointed database entries for the same individual, a challenge that many organizations commonly face (Christen, 2012; van Keulen, 2012). As we have explored forms

of data friction influencing re-identification in the preceding sections, it is evident that unsuccessful re-identification can be connected to the proliferation of duplicates. When an applicant's data already exists in the database but is not correctly re-identified, for example, due to one of the forms of friction, the possibility of generating a new, duplicate record is real. The excerpts presented in this section will explore how duplicate records can arise within the IND's operations and give insights into the subsequent work needed to rectify them.

The following interview excerpt underscores how certain departments within the IND are particularly susceptible to generating duplicate identity records, often attributed to factors such as the high volume of applications they handle and disparities in knowledge and experience with the employed search tools:

> Yes, specific departments within the IND have that [problem of creating duplicates], which can indeed create a duplicate applicant more often. For example, counter staff can do that. Due to lack of, well, yes, just having less experience with the system. At least with search [functionalities]. Of course, they have to work quickly because they have the applicant in front of them, so to speak. So maybe there is a bit more time pressure. And besides, my department, which is more trained in searching… we may generally be searching a little better in the system. […] We also have the postal department, which is called DRV, digital registration and preparation. And, of course, they have many more applicants [to process] daily, so there is a good chance that they will create a duplicate applicant. But also my department, even though we are very trained in this. We are also still creating duplicate applicants. More than we'd like. (Interview with IND staff member, January 29, 2021)

The interviewee highlights that even in their department, where they possess a solid grasp of the tools, the creation of duplicate records persists. This observation highlights the multitude of factors that may contribute to the generation of duplicates, even among knowledgeable staff. Discussing this matter, another interviewee added that instances of duplicates could also arise when a migration chain partner has previously established a record for an applicant. However, this record does not get properly re-identified:

> We [the IND] also have contact with migration chain partners such as the Vreemdeling Politie [national police], the Koninklijke Marechaussee [national gendarmerie] and the like. They can also create applicants themselves. Buitenlandse Zaken [foreign affairs] can also create applicants. And it often happens that sometimes an applicant has already been created by the Vreemdelingen Politie, and then [the applicant] comes to our counter, that [the data] is created again and that it has been created again in a different way. (Interview with IND staff member, November 2, 2020)

In this scenario, an applicant might not have been successfully re-identified from the record of a migration chain partner due to minor disparities in the data. This, in turn, can result in the creation of a separate and distinct record. In that same interview, the interviewee also cited an issue arising when a considerable amount of time has lapsed between applications from the same individuals. They presented a scenario where IND staff processing the application fail to locate the applicant's old record, subsequently generating a new one. The interviewee gave two examples of how such duplicate records could be identified. One way is during the final stage of the applicant evaluation process, where a higher-up in the organization conducts extra checks before making a final decision (such as granting a residency permit). Another way is if a new document related to the application arrives, which then requires re-identification. In either case, two records are identified as possibly referring to the same individual and will need to be resolved (the "deduplication process"). Here is how they described the process:

> So there is a moment someone creates an applicant. And then a new document arrives. And the person after that could be a year later, five years later, he's looking and actually finding nothing. But the applicant still comes first. And if he then creates it again, he often does not immediately realize it himself, but I get that back or from the decision process from the applicant that occurs twice. Or someone else also receives a document that then comes in and starts looking for the applicant, only to find out that it occurs twice. Then we have a deduplication process that then turns the two applicants into one applicant again. (Interview with IND staff member, November 10, 2020)

To put it concisely, when identifying duplicate records, as indicated in the interview quote, they are flagged and processed through a "deduplication request." The "Titles and Identity" (T&I) department processes these requests, taking the necessary steps to take the records in question. Deduplication requests can come from various sources, including IND departments and migration chain partners, with requestors completing a designated form for submission to the T&I department. The form outlines the specific data elements that prompted the requestor to identify duplicates and explain the reasons behind the deduplication request. One interviewee from the T&I department shed light on the evolution of this "deduplication request memo" (*ontdubbelverzoek memo* in Dutch):

> Internally, within the IND, such a deduplication memo must be sent. And for those requests, we have a kind of standard analysis. So basically an analysis based on system facts. For example, what comes in it is: we want to know if there is an identity difference between the applicant. Because we've actually been working with that since — let's see — October 2019, when we also focused on identity differences. We call this "identiteitsvraagstukken" [identity question/problem] internally. That needs to be taken care of, and that was never done before. Until we got a case with a difference that we

> were aware of, hey, if we actually deduplicate files, we choose to make specific identity leading while this is not correct. Because my department just doesn't have that authority. And that authority is laid down in policy documents, among other things. (Interview with IND staff member, January 29, 2021)

The interviewee explained that the protocols for deduplicating records lacked a standardized approach. However, as the complexities of the deduplication process became more apparent, efforts were made to streamline and standardize the process. The interviewee emphasized the need for such standardization, especially after instances of incorrect decisions. In situations involving two duplicate entities, a decision is required to determine which of them should be designated as the "leading" identity data based on evidence from the systems ("system facts," further elaborated below).

The following excerpt from another interview underscores the significant weight attached to the decision-making process of deduplicating data records. Within the organization, this process is sometimes metaphorically likened to designating a "survivor" and a "loser" record. Essentially, the "loser" data record is rendered inactive, a potentially impactful action that could lead to real-world complications for the affected individual if carried out inaccurately, given the interconnectedness of documents associated with these records. The interviewee highlights the challenges of losing document and case data traces during deduplication, highlighting the necessity for traceability in the decision-making process within the system for future procedures:

> [...] you choose who... we call it "the survivor." You choose who will be "the survivor" and "the loser." It sounds very harsh, but that's how we choose who will be the survivor. And all data from the survivor remains leading. And you can possibly still find somewhere — if you search very well — some personal details, data of when he was born and the like. But certain data and file documents. You really can't do anymore; unless it is really stated in the file document under which applicant it was. But other than that, you really can't figure it out anymore. Cases only if file documents are linked to them and that file document contains a case number. But otherwise you really wouldn't know anymore. There is nowhere that a copy is stored of this was the situation. None of that is there. That is also the reason why people fill out that memo. So that you at least — should it go wrong — that you notice a little, that you can figure out a little. (Interview with IND staff member, November 2, 2020)

To ensure traceability and consistency in the decision-making process within the deduplication, as highlighted by the interviewee, the "Titles and Identity" (T&I) department utilizes a template to gather information from the IND and other relevant systems upon which the decision will be based. This template outlines the various types of

"evidence" that IND personnel can gather to make a decision. These accumulated pieces of evidence are meticulously documented within the same memo. Upon analyzing the template and the deduplication process, I found a classification of evidence into two distinct categories, termed "weak evidence" and "strong evidence," as denoted in the template itself. Weak evidence encompasses elements such as matching addresses, which can be considered a weak indicator due to potential outdatedness or shared addresses among certain groups like asylum seekers residing in the same housing center. Conversely, strong evidence comprises instances where two records possess identical identification documents, a robust indicator owing to the uniqueness of such documents and their clear association with an individual. While I cannot disclose all the specific types of evidence, this categorization highlights a specific ordering of data types, where certain types hold more significance in determining whether the records pertain to the same individual.

It is important to note that some strong evidence can also originate from sources outside the IND, including migration chain partners and other international information systems. An illustrative case is the justification for deduplication when the National Police finds duplicate records by establishing a connection between data from Eurodac or EU-VIS. In such cases, they may leverage data from prominent European Union information systems to establish a link between seemingly separate records within the IND's database. In this case, even international systems can take priority in the re-identification process. These scenarios of duplicate resolution demonstrate that even on a local level, resolving questions about personal identity data can involve the engagement of international systems.

Based on my analysis, the IND's deduplication process can be conceptualized as having four stages. In the first stage (1), multiple records potentially referring to the same individual are identified. These records can be discovered through various means, encompassing internal IND and external re-identification procedures or even through an automated process, which will be elaborated on shortly. Subsequently, in the second stage (2), relevant parties are promptly alerted to the existence of these duplicate records through the deduplication memo. In the third stage (3), the T&I department's staff gathers evidence by investigating the records in question. They use a systematic approach to determine whether the identified duplicates belong to the same individual. To assist in this process, they use a structured form that outlines different categories of evidence that can support their decision-making. Upon the collection and evaluation of evidence, the fourth and final stage (4) involves the actual execution of the deduplication process within the system. In this step, the records will be processed, documents or data will be updated, and one of the identified duplicate records will be marked inactive. Additionally, these steps are logged in the memo, documenting the entire deduplication procedure and the decision-making process, offering transparency and traceability to subsequent procedures. Figure 5.5 visually presents these different stages in the deduplication process.

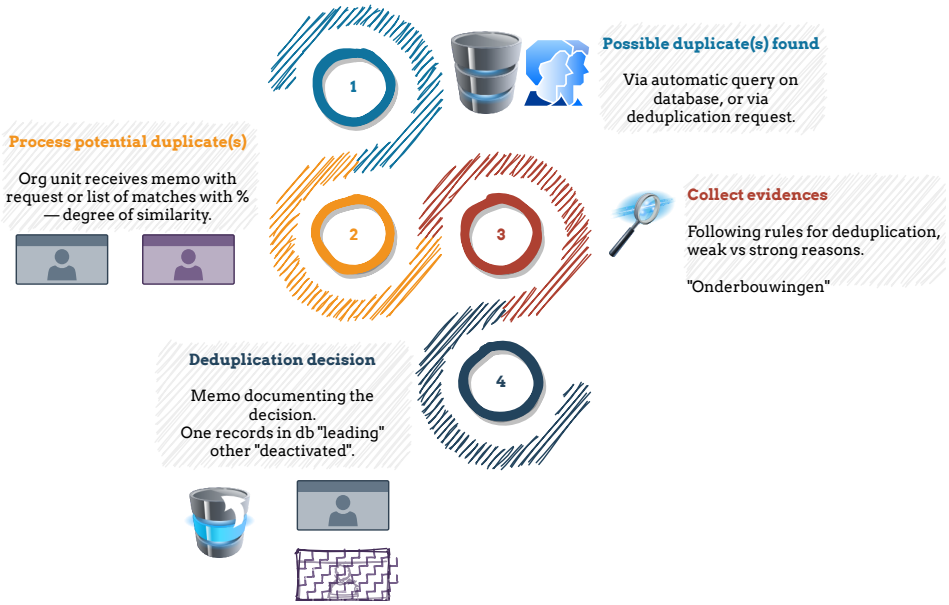


Figure 5.5: The author's conceptualization of the IND's deduplication process.

The process of deduplication resonates with the insights articulated by Loukissas in his exploration of the idea that "all data are local," suggesting that the act of "normalizing" duplicates offers a lens through which to examine the diversity inherent in data infrastructures (Loukissas, 2019). Echoing Loukissas' insights, we can consider IND's deduplication process as a lens demonstrating the heterogeneity of re-identification practices. The emergence of duplicate records within the IND's database is a result of various factors, ranging from the pressures of workload in specific departments to the complexities of handling diverse forms of information, such as handwritten documents or historical data. Moreover, discrepancies in the practices of migration chain partners can lead to the creation of duplicates when their registrations diverge and subsequently fail to re-identify the same individuals. Examining the process of resolving duplicates not only elucidates the operational intricacies within the deduplication process but also offers insights into the challenges posed by heterogeneous re-identification practices.

Additionally, the deduplication process exposes the changing bureaucratic dynamics associated with re-identification practices. Insights gleaned from the interviews suggest that there may have been a time when IND personnel wielded greater bureaucratic discretion in resolving duplicates, including deciding which record to render inactive. However, this discretionary approach may have resulted in inaccurate decisions. Consequently, the introduction of the deduplication request and memo suggests the shift

towards a more standardized approach, complete with clearly defined weak and strong evidence criteria. This shift signifies the organization's endeavor to streamline the deduplication process, minimize the reliance on discretion, enhance transparency and traceability across the entire process, and ensure more consistent and equitable outcomes. Furthermore, this shift suggests a potential redistribution of roles within the organization, particularly regarding the responsibility for thorough applicant re-identification. In some instances, such as due to challenges like time constraints, staff may not have the opportunity to conduct thorough re-identification of applicants, which can result in the creation of duplicate records. Consequently, the department responsible for deduplication has expanded tasks of correcting wrongly re-identified applicants.

While we have delved into the manual identification of duplicates, an unexplored aspect within the deduplication process is an extension of the initial stage: the automated identification of potential duplicate records within the IND's database. This automated process entails leveraging the capabilities of the ELISE data matching engine to identify duplicate entries based on specified criteria proactively before they escalate into more complex issues. In the following section, we will delve into the mechanics of automated deduplication, providing a lens into the broader complexities inherent in defining identity for re-identification.

5.6.2 Automated duplicate detection and the challenges of defining duplicates in re-identification

The prior interview excerpts have revealed recurring causes for duplicate records being introduced into the INDiGO system. Factors such as time constraints, limited familiarity with the search tools, inadequate training, knowledge gaps, and system integration with external organizations contribute to the creation of duplicates. Multiple records for the same individual pose a risk for the IND to make erroneous decisions, and detecting these duplicates early is therefore considered crucial for the IND (Fieldnotes 2020-07-06). Consequently, the data matching system is employed not only for applicant searches but also to identify potential duplicates using distinct criteria tailored for this purpose.

Upon reviewing available documentation, I uncovered two distinct approaches that could have been adopted for automated duplicate detection within the IND's re-identification process.[12] The first approach involves implementing automated checks during the initial registration of applicants. The second entails periodic automated scans of the entire database to identify potential duplicates. Notably, the IND chose to forego the first option, which would involve automatic checks while creating applicant records. Instead, the organization has opted for the second approach, relying on automated database queries performed at intervals to detect potential duplicates. In practice, this service systematically compares recently created applicant records and all

[12] INDiGO's technical design documents refer to these two approaches to finding possible duplicates as online and offline deduplication.

pre-existing database entries. This process aims to identify potential duplicates between a newly generated record and an older one. The IND has distinct criteria for calculating match scores within the ELISE data matching system for this automated process, which differ from the standard criteria used for regular applicant searches. If the computed match score between two records surpasses a predetermined threshold, both records are flagged as potential duplicates, warranting further investigation.

During fieldwork, I had the opportunity to gain further insights into the configuration of the data matching system designed for identifying potential duplicates. These insights were acquired through discussions with the staff of the WCC regarding the development of a new version of the duplicate resolution system (Fieldnotes 2020-07-06). At that time, the previous version encountered performance issues, prompting the development of an updated version to be rolled out. In meetings with WCC staff, I was privy to the considerations and deliberations surrounding configuring data-matching rules for duplicate detection. Although I cannot disclose all the rules, I can share some general principles. Notably, matching criteria included factors such as shared surnames, nationalities, and birth years (with less emphasis on matching months and days of birth). For instance, I was given an example of twins with different names but the same initials and birthdates, which could trigger a false positive in the duplicate detection mechanism if the rules were not carefully set up. This illustrates the complexity of developing matching criteria that are both sensitive enough to identify potential duplicates but also specific enough to avoid incorrect identifications.

What stood out in these discussions was the dual consideration of custom-made solutions versus standardized software. Specifically, the company engaged in developing the deduplication tool was weighing the merits of creating a solution tailored precisely to the IND's needs versus a more standardized solution that could potentially be deployed for other customers in the future. A member of the project team outlined the contrasting features of these alternatives. On the one hand, a standardized solution would offer a consistent approach to detecting duplicates, making it easier to implement across various customers of the company. However, it might be less adaptable to the specific organizational contexts in which it is deployed. On the other hand, a highly customizable solution could be crafted, albeit requiring significant tailoring to adapt to individual customers' unique contexts and deduplication requirements. Ultimately, the custom-made solution was deemed the most fitting choice. A concern that was repeatedly raised was the difficulty in creating clear and comprehensive definitions for identity and duplicate records, consistently portrayed as deeply entwined with the specific organizational context.

Delving into these automated mechanisms not only uncovers the complexities of identifying and managing duplicates but also underscores their entanglement with street-level, screen-level, and system-level bureaucracies. Duplicates often emerge due to failed re-identification, which can occur at any of these three levels. Moreover, the resolution of these duplicates can occur at any of these three levels. Offline

deduplication, characterized by periodic reviews of potential duplicate records, aligns more with system-level operations. In contrast, online deduplication operates in real-time, affecting street and screen-level bureaucrats immediately during their daily workflow. Furthermore, the choice between a standardized and customized deduplication approach underscores how identity and duplicate record definitions relate to organizational contexts. Chapter 6 will delve deeper into the intricacies of constructing data matching systems to work across organizations. However, before we proceed, the upcoming section will introduce an analytical tool designed to interpret the diversity of re-identification practices discussed thus far.

5.7 An interpretative framework for re-identification scenarios based on search input and results

This section sets out to synthesize the chapter's findings to address the research question that has guided this chapter's investigation: "How do organizations that collect information about people-on-the-move search and match for identity data in their systems? How is data about people-on-the-move matched and linked across different agencies and organizations?" As we navigated the multiple facets of re-identification at the IND and the migration chain, these questions can now be answered by synthesizing the findings into a typology of re-identification scenarios. As such, this analysis serves a dual purpose: it unveils the operational intricacies of data matching in migration management's bureaucratic procedures while delving into the challenges, friction, and complexities that weave through these procedures.

In light of the chapter's findings, it becomes evident that re-identification within IND can take different forms. On the one hand, re-identification is characterized by the diverse information available to staff during the process: more or less skills can be required to interpret information to be inputted in the system. On the other, the precision requirements for successful re-identification differ significantly: more or less skills can be required to interpret search results. Building on these observations, I propose an analytical framework for interpreting re-identification that categorizes practices based on the demands of interpreting search inputs and results. This classification yields four distinct combinations, visually depicted as a matrix in Figure 5.6. This interpretative framework serves as a tool for unraveling the intricacies of re-identification within street-level, screen-level, and system-level bureaucracies, enabling a clearer understanding of the diverse impacts of data frictions and the ensuing costs that arise in the event of re-identification failures.

In the upper left quadrant of the matrix, re-identification scenarios are depicted where the need for scrutinizing both input and results is relatively low. This often pertains to routine tasks involving unambiguous data like identification numbers, frequently conducted at service counters where direct interactions with applicants occur. An illustrative instance is when staff need only the v-number to access existing

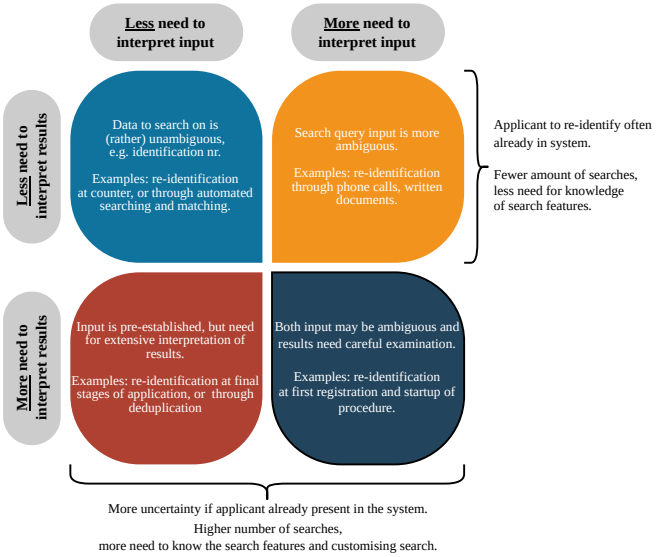


Figure 5.6: This matrix shows the four combinations arising from differences in the need for interpreting the input with the need for interpreting the results in re-identification.

applicant data, leading to straightforward search results to retrieve the applicant's data. This quadrant also includes instances of automated search and matching processes, such as the automated exchange of residency data. In these cases, the search query is predefined, meaning it's set in advance without the need for interpretation by IND staff, as the data originates from another source. Moreover, the system doesn't interpret the results like a human would; instead, it simply selects the outcome with the highest match score from the available matches.

The upper right quadrant pertains to re-identification scenarios wherein the input may have a slightly higher level of ambiguity, yet the necessity for reviewing output remains relatively modest. This quadrant encompasses scenarios involving tasks like handling phone calls or processing written documents; it includes scenarios where the applicant being searched for is typically already present within the system, leading to fewer searches and a decreased demand for comprehensive familiarity with search functionalities. In this context, the primary challenge for re-identification originates from the necessity to interpret the input. For instance, this might involve deciphering an applicant's v-number provided over the phone or transcribing data from a written form into the search query. This process can introduce errors that affect the search results, for instance, due to audible misunderstandings or typographical mistakes.

The lower-left quadrant focuses on re-identification scenarios characterized by a diminished requirement for reviewing input, as it has typically pre-established, but a heightened necessity for scrutinizing results due to the careful interpretation they demand. This quadrant encompasses instances like the processing of automated deduplication matches and the tasks by what is termed as "decision makers" within the organization. In automated deduplication, the input is also predefined, but considerable effort is necessary to verify potential duplicates meticulously. The notion of decision makers pertains to individuals within the organization, often superiors, who conduct final assessments before arriving at decisions concerning applicants. In both these scenarios, a rigorous validation process is essential to ascertain whether an individual is already present in the system, leading to a higher volume of searches. Consequently, this demands a thorough familiarity with search functionalities and customized search methodologies.

Finally, the lower-right quadrant encompasses re-identification scenarios characterized by ambiguous input and a need to examine search results thoroughly. Scenarios falling within this quadrant are exemplified by the initial registration and commencement of a procedure for an applicant, particularly when dealing with processing written application forms. Here, staff may encounter the challenge of deciphering potentially unclear and error-ridden forms. Simultaneously, thorough inquiry is required to ascertain whether an individual is not already enlisted within the IND or migration chain. This is necessary to ensure that an applicant has, for instance, not been previously registered under a different name before marriage. Consequently, as revealed by the findings, such cases demand a familiarity with the search and match tools and often prompt staff to

devise their own search strategies.

The matrix consisting of four distinct combinations functions as an interpretative framework, enabling us to comprehend the nuances of various re-identification scenarios within the entanglement of street-, screen-, and system-level bureaucracies and how different scenarios interact with and respond to different forms of data friction. The framework categorizes these differences as contingent upon specific levels of input and output review requisites. In the upper two dimensions of the matrix, there is a higher certainty that the applicant is already within the system. This leads to scenarios characterized by fewer searches, less reliance on intricate search features, and potential friction primarily related to processing search results. In contrast, the lower two dimensions signify greater uncertainty about the applicant's presence in the system. More searches are needed in these situations, a comprehensive understanding of search features is crucial, and customizing search parameters becomes essential. The potential forms of friction in re-identification are likely more influenced by the complexities at the input stage and by differences in re-identification practices.

Re-identification scenarios in bureaucratic settings exhibit a complex interplay that extends across street-level, screen-level, and system-level operations. Street- and screen-level bureaucrats often initiate the data entry process, gathering and inputting information into the system. This data then undergoes subsequent processing where applicants are re-identified for verifying, updating, and correcting applicant details. Throughout this process, these bureaucrats grapple with data friction stemming from disparities between the information provided by applicants and the data present in official records. Their ability to achieve accurate re-identification relies on interacting with system interfaces that shape the process at various stages – from entering search queries to interpreting the ranked results. Meanwhile, system-level decisions revolving around matching algorithms, criteria, and technical configurations wield influence over the outcomes of re-identification efforts.

5.8 Conclusion

This chapter commenced by highlighting the significance of re-identification processes, even in high-stake cases like that of the Boston bomber, where specialized technologies are employed by authorities to navigate ambiguous re-identification outcomes, as evidenced by the ambiguity surrounding the perpetrator's transliterated name. Notably, participants in the study also highlighted security considerations, such as the role of duplicate record detection in identifying individuals with malicious intent. However, a more comprehensive exploration of the connection between data matching and security will be addressed in the subsequent chapter. Shifting away from such high-stakes contexts, the investigation delved into re-identification processes as routine bureaucratic practices supported by technological tools aimed at minimizing uncertainties. Through a comprehensive exploration of the Netherlands' Immigration and Naturalization Ser-

vice (IND) and its interconnections with the Dutch migration chain, the research set out to examine the multifaceted practices of re-identifying applicants.

The inquiry began by examining the IND's designed infrastructure for applicant re-identification, including the tools for searching and matching applicants' identity data. Next, this designed infrastructure was compared through its practical implementation and use, enabling us to unravel the challenges inherent in the IND's identification processes. The analysis identified three prominent forms of data friction that may hinder applicant re-identification: friction between standardized identification and the differences in institutional practices, friction from variations in the precision and accuracy of identity data during its transformation across different mediums and use in formulating search queries, and friction arising from the opaque calculation of match results and the need for thorough interpretation and fine-tuning of search results. These forms of friction, in turn, prompted a closer examination of the costs arising from failed re-identification, as exemplified by the existence of duplicate records and the labor-intensive process of deduplication.

The findings enrich our theoretical understanding of re-identification processes in three dimensions. Firstly, re-identification is more than the technical mechanisms of matching data; it encompasses complex negotiations between diverse identification practices. It involves the IND acting as a mediator in the complex web of bureaucratic operations, aligning multifaceted identification practices for successful re-identification. Secondly, the data matching system's standardized design interacts dynamically with individual user needs during practical application. This interaction underscores the complexity of translating standardized designs into effective practice, particularly within the intricate context of the IND's operations. These findings emphasize that successful re-identification requires a confluence of designed features, user adaptations, and real-world intricacies. Lastly, the expertise needed for successful re-identification is through a synergy of human and technological data matching expertise. The interviewees' comprehension of the data matching system was partial, as computer systems employ autonomous functionalities beyond their explicit awareness. Moreover, the specific match criteria often remain opaque. This distribution of knowledge signifies a division of labor, where human operators rely on computer systems to enhance their re-identification efforts. However, successful re-identification also rests on the tricks of the trade of human operators in refining, comprehending, and critically evaluating search results.

Examining the costs of failed re-identification through the existence of duplicate records and the labor-intensive process of deduplication contributed two insights into re-identification. Firstly, delving into the process of resolving duplicates not only unveiled the operational intricacies of deduplication but also offered a lens to understand the evolving dynamics of bureaucratic re-identification practices. Interviews indicated that there may have been a period when IND personnel exercised greater bureaucratic discretion in resolving duplicates, potentially leading to inaccurate decisions. Conse-

quently, the shift towards a standardized approach, marked by pre-defined evidence criteria, reflects the organization's pursuit of streamlining deduplication, reducing discretionary elements, enhancing transparency and traceability, and ensuring more consistent outcomes. Secondly, the automated duplicate detection process, driven by the data matching engine, presents a proactive solution to enduring re-identification challenges. However, delving into these automated mechanisms not only reveals the complexities of duplicate identification but also underscores the broader intricacies inherent in defining identity. The struggle to establish a universal deduplication method underscores that such definitions are inherently tied to an organization's context, defying easy application to other organizations.

These findings were synthesized to create an interpretative framework that conceptualizes re-identification practices according to the demands of interpreting search inputs and results. This resultant matrix of diverse re-identification scenarios effectively serves as a bridge to address the gaps identified in the literature review. The literature review unveiled two overarching themes regarding the conceptualization of re-identification as a bureaucratic practice. We recognized re-identification as a substantial yet relatively unexplored aspect of bureaucratic interactions with applicants, particularly as these interactions shift from traditional in-person settings toward more digitally-mediated and automated processes, characterized in the literature as a transition from street-level to screen-level and system-level bureaucracies (Bovens and Zouridis, 2002). The matrix of re-identification scenarios illustrates the multiplicity of practices, showcasing the diverse ways data matching tools shape the processes of re-identifying individuals within bureaucratic contexts. These scenarios encompass various situations, ranging from direct applicant interactions to staff managing phone calls and handling application forms sent via post.

Furthermore, in the literature discussing materialist and performative viewpoints of identification and the challenges arising from data uncertainties and data friction, a noticeable knowledge gap exists concerning how practitioners effectively manage the intricacies of uncertain personal identity data during the identification processes. This gap prompted exploring the effects of technologies designed to address data uncertainties on bureaucratic re-identification practices, potentially enhancing and limiting their efficacy. The matrix visually represents the varying input and output reviews required across distinct re-identification scenarios. Through its typology, the matrix categorizes uncertainties and ambiguities present not only in search inputs but also in search results. Consequently, the findings emphasize that the approach practitioners adopt in navigating uncertainties within personal identity data during identification encounters is context-dependent and shaped by the tools and mechanisms of data matching.

This chapter's findings also address the dissertation's central research question: "How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?" Through examining applicant re-identification practices within the

Netherlands' Immigration and Naturalization Service (IND) and the larger migration chain context, this chapter demonstrated the interconnection between a government migration agency's re-identification practices and commercial data matching systems, exemplified by the ELISE software. Furthermore, the examination of the deduplication process reveals a crucial link with transnational systems, as the IND and migration chain partners leverage data from prominent European Union information systems to establish connections between seemingly disparate records within their databases. The findings underscore that re-identification processes and technologies are not isolated; they are entwined with wider commercialized security infrastructures.

Several questions remain unresolved regarding the interplay of matching identity data within migration management and border control and their interactions with transnational commercialized security infrastructures. The integration of commercial and proprietary tools, crafted by a private entity, for data matching, suggests an influential shift in the IND's re-identification knowledge. This integration intertwines the core of re-identification knowledge with a specialized system, resulting in a scenario where significant expertise resides within this proprietary system. However, the recent upgrade of the IND's deduplication tools illustrates the tension between crafting generic versus context-specific data matching systems. Such tensions raise questions about the potential role of the generic design of ELISE and software vendors' strategies for "generification" (Pollock and Williams, 2009). Regrettably, the impact of software vendors and their broader market strategies remains largely unexplored in current research. Thus, the subsequent chapter will illuminate the software's evolution, shedding light on how knowledge and technology related to matching identity data traverse diverse organizational boundaries.

Chapter 6

Uncovering the long-term development of identification infrastructures: A multi-temporal perspective

Abstract

Systems and infrastructures for identifying and registering mobile populations have many facets and long development histories, and researchers' partial perspectives shape their understanding of the technologies and practices involved. To overcome methodological partiality, researchers frequently study infrastructures at multiple sites or including human and non-human actors that shape identification encounters. As a further option, this chapter suggests that researchers can use multi-temporal sampling methods to understand the long-term development of identification systems and infrastructures. The chapter proposes two heuristics for selecting contingent moments in the lifecycle of identification technologies. The first heuristic employs the Social Construction of Technology's concept of "interpretative flexibility" to pick out moments when social groups challenge, change, or close down the meanings of identification practices and technologies. The second heuristic employs infrastructure studies' concept of "gateway moments" to pick out moments when heterogeneous identification software systems and infrastructures intersect. These two heuristics were tested through the analysis of data gathered at an IT vendor of software for matching people's identity data. This chapter makes two contributions to the research agenda of long-term perspectives on identification software development. The first contribution demonstrates how the contingent interpretation of data matching system corresponds to diverse problematizations of identification and its securitization. The second contribution demonstrates how "gateway moments" make it possible to see the compromises necessary when building identification infrastructures and adapting globally honed technologies to new settings. Together, these findings shed light on the activities of under-the-radar actors, such as software vendors, whose distribution and reuse of data matching systems have long-term implications on identification practices and infrastructures, not only in the security field.

Contribution to research objectives *To investigate the long-term development of identification systems and building of transnational data infrastructures by identifying crucial moments in their lifecycle to explore how data matching expertise travels and circulates.*
This chapter contributes to the research objective by employing two guiding heuristics: the interpretative flexibility of data matching systems and gateway moments associated with their integration into broader infrastructures, to select and analyze contingent moments. The first heuristic focuses on the system's changing interpretive flexibility, which allows us to see actors' varying problematizations of data matching and identification. The second heuristic uses "gateway moments" when systems and infrastructures intersect, which makes it possible to see the compromises necessary when integrating identification systems into broader infrastructures. Consequently, the chapter highlights the myriad factors influencing such systems' development, adaptation, and dissemination. Of particular significance is the exploration of how data matching expertise traverses and circulates, exemplified by the development of name matching expertise that finds application elsewhere and instances where this circulation faces hindrances due to infrastructural challenges, like backward compatibility issues. The chapter's analysis offers a nuanced perspective on the intricate and contingent interplay between technological advancements in data matching and their enduring impacts on transnational data infrastructures.

Infrastructural inversion strategy *Third inversion strategy — sociotechnical change*
The strategy is operationalized using a "multi-temporal sampling" approach, offering an alternative to conventional longitudinal studies for exploring the complex processes shaped by and shaping identification systems and infrastructures. In this context, the analysis of the data matching system assumes a dual role, serving as both the subject of investigation and a resource. This dual role enables a comprehensive exploration of the intricate and contingent processes underlying the development of WCC's ELISE data matching system within broader transnational data infrastructures. The ELISE system was repeatedly adapted to the evolving demands and challenges within the field of data matching, exhibiting phases of openness and closure in its design. As a resource, the method is used to address the research question of the chapter, which revolves around the investigation of the circulation of data matching knowledge and technology across various organizations, thereby providing insights into the dynamics of how data matching knowledge travels and circulates within the realm of identification systems and infrastructures.

Contribution to research questions *RQ3: How do knowledge and technology for matching identity data circulate and travel across organizations?*
By examining the interpretive flexibility of data matching software, the chapter sheds light on the roles of often-overlooked actors in this knowledge and technol-

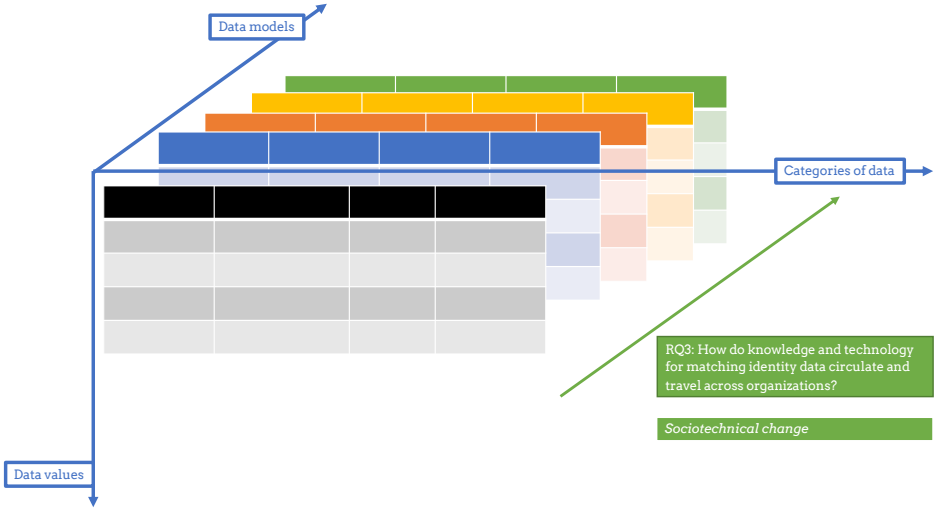


Figure 6.1: The axes pertaining to the methodological framework in relation to chapter 6.

ogy circulation. The analysis highlights instances where international professional networks and government agencies have influenced the trajectory of data matching technology, emphasizing how such factors drove adaptations and innovations within organizations like WCC. Furthermore, the chapter explores gateway moments, illustrating that the circulation of data matching knowledge and technology is not deterministic but contingent on specific contexts. The integration of ELISE into both the EU-VIS and IND systems serves as a compelling example, showcasing how different implementation approaches influence the circulation of data matching technology and expertise across organizations.

Contribution to main research question *How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?*

Through an analysis of interpretative flexibility moments, this chapter traces the evolution of identity data matching technology, conceived initially for versatile applications but later refocused on the identity and security sector due to geopolitical shifts and heightened security concerns. This shift led to expanding data matching capabilities, including biometric data matching and specific name matching for security agendas. Additionally, the chapter emphasizes the commercial nature of these technologies, detailing their extensive network of WCC's partners and collaborators. Furthermore, examining gateway moments illustrates how transnational infrastructures influence data matching practices. The integration of ELISE

into EU-VIS showcases member states' control over matching criteria and the challenges of such integration. At the same time, incorporating data matching into legacy systems highlights opportunities for standardization. Overall, the chapter provides a comprehensive understanding of how identity data matching practices and technologies are shaping and being shaped by transnational commercialized security infrastructures.

6.1 Introduction

What we know about technologies for identifying people is inextricably linked to what we know about the devices and practices involved (Garfinkel, 1964; Latour, 1986; Marres, 2017; Pollock and Williams, 2009). Every year, authorities identify millions of people crossing the EU's external border, including migrants, tourists and asylum seekers. In order to identify people, authorities frequently cross-reference and link personal information from different databases. Examples include checking visa applicants' identities, preventing identity theft by linking identities across worldwide law enforcement databases, and screening passengers for potential security risks by comparing their personal information to government watch lists. By facilitating or hindering particular forms of mobility and excluding others, these processes contribute to the fragmentation of mobilities (Olivieri, 2023; Sparke, 2006). However, what do we know about the histories of identification systems, and how are the technologies known to us? Most importantly, through which methods?

One significant challenge in researching technologies, often stemming from their development over extended periods and involving multiple components, lies in the inherent limitations of researchers' partial perspectives. For example, identification systems can be extensive, involving numerous interconnected components, databases, and networks, making it difficult for any researcher to grasp every aspect fully. Furthermore, the evolution of technological systems over time adds another layer of complexity. Many technologies have a long developmental history, undergoing continuous changes, updates, and adaptations. Investigations might focus on a specific snapshot in time, but this limited perspective might not capture the system's history or its future trajectories. To address this challenge, multi-site sampling methods are often employed to study technologies in diverse contexts. While these methods offer valuable insights into the varied practices and actors shaping identification technologies across sites, they have limitations in tracking how these practices and technologies evolve over time. By recognizing the significance of understanding temporal changes, incorporating a multi-temporal perspective becomes crucial. Tracking changes over time allows unraveling the dynamic nature of identification practices and technologies, unveiling the influences, adaptations, and challenges that occur throughout their lifecycle. By adopting a multi-temporal approach, researchers could uncover the intricate interplay between social, technological, and contextual factors that shape identification systems and infrastructures.

As employed in this chapter, the term "multi-temporal" does not imply the coexistence of multiple linear temporal dimensions running in parallel. Rather, it pertains to the practice of sampling distinct moments in time. Adopting such an approach can unveil a diverse range of moments within the lifecycle of identification technologies, each offering a unique perspective on these technologies and their associated practices. For instance, tracing the emergence of identification systems often involves following key

actors and scrutinizing the records they generate. This can encompass examining processes such as the design, construction, and maintenance of systems and policies, as seen in initiatives like the EU's "smart borders" project (Bigo et al., 2012; Jeandesboz, 2016) or the EU's Visa Information System (Glouftsios, 2019). Of course, chronicling a system's entire lifecycle may be impractical and not always necessary. Researchers should instead make well-informed choices based on their familiarity with the subject matter and research objectives (Pollock and Williams, 2010). These choices might involve focusing on one or more specific moments in the lifecycle of a sociotechnical identification system (e.g., planning, analysis, design, procurement, implementation, operation, and maintenance), recognizing that omitting certain stages will inevitably result in partial descriptions (Pollock and Williams, 2010). Sampling multiple moments allows researchers to explore various stages of the development and operation of sociotechnical identification systems, depending on their research goals and objectives.

Within STS, there is an ongoing recognition of the contingency involved in the development and practical use of technological artefacts (e.g., Akrich, 1992; Bijker, 1993; Suchman, 2007). This position questions deterministic perspectives on technological development and recognizes the role played by multiple factors, actors, and structural constraints in shaping uncertain sociotechnical outcomes. In a similar vein, in the overlap of STS and security studies, it is acknowledged that practices and security encounters are contingent (Amoore, 2013; Kloppenburg and van der Ploeg, 2020; Pelizza and Aradau, 2024). This approach emphasizes the divergence of security apparatuses from a perceived coherent structure, as various social, political, and technological forces introduce contingency. By emphasizing contingency, I specifically aim to spotlight moments where outcomes in a data matching software development are not predetermined but are instead influenced prominently by the circumstances, factors, and decisions made by individuals and entities involved.

By concentrating on how sociotechnologies of identification move and circulate across organizations, influencing identification problems and solutions, this chapter addresses the dissertation's goal to investigate the long-term development of identification systems and building of transnational data infrastructures by identifying contingent moments in their lifecycle to explore how data matching expertise travels and circulates. It does so by proposing two heuristics to identify contingent moments to study the evolution of identification-related sociotechnologies by drawing on concepts and theories from long-term and genealogical analyses of data infrastructures and social constructivist accounts of technology. Overall, the chapter aims to offer a methodological solution using sampling moments to address research question 3:

> How does knowledge and technology for matching identity data circulate and travel across organizations?

To address this research question, the chapter first considers methods suitable for tracking the spread of identification practices and technologies, and then provides

heuristics for identifying moments where these circulations of knowledge and technology are prominent. The next section begins by classifying the various research strands that have investigated sociotechnologies of identification according to the types of sampling methods used. First, much ethnographic research has been influenced by debates about expanding the number of research sites and connecting observations to gain insight into more encompassing phenomena, such as the construction of a border identification regime that criminalizes migration. Second, scholars have begun considering the complex web of human and non-human actors that make up border identification encounters. However, neither of these approaches can show how practices and technologies have changed over time. Thus, there is a need to provide detailed accounts that include the multiplicity of sites and actors and the multiplicity of moments in time (Hyysalo et al., 2019). Most often, longitudinal research is used to answer this need. This chapter introduces an alternative approach to longitudinal research by identifying analytically valuable moments in the evolution of sociotechnologies of identification.

This chapter suggests utilizing two heuristics drawn from the literature to investigate the evolution of sociotechnologies related to identification, focusing on moments of interpretative flexibility and gateway moments. Long-term and genealogical studies of information systems and infrastructures have convincingly demonstrated that temporal approaches make it possible to avoid a teleological view of the design of technologies (Edwards et al., 2009; Karasti et al., 2010; Ribes and Finholt, 2009; Williams and Pollock, 2012). Instead, temporal approaches allow the inclusion of often overlooked actors and moments in the development of identification sociotechnologies, such as the numerous interactions between government actors and technology consultants as they work together to develop the problems and solutions for identification. However, as an alternative to longitudinal studies, this chapter will propose "multi-temporal sampling" to provide another approach for examining of the evolution of identification in technologies in border and migration control. Before going over the two heuristics for multi-temporal sampling, we need to comprehend how researchers typically study identification, which occurs across multiple sites and involves both human and non-human actors.

6.2 Sampling methods for dealing with the scale of sociotechnologies of identification

Ongoing scholarly debates in Science and Technology Studies (STS) about the intertwining of methods and outcomes of investigations may provide inspiration to the research of identification technologies. In general, the discussion has drawn attention to an incompatibility between research methods that adopt a limited perspective and the understanding that technologies are shaped over time, in various contexts, and by various actors (Hyysalo et al., 2019; Silvast and Virtanen, 2023). This incompatibility has led to investigations that fail to account for the multiple and contingent lives of technologies,

due to research designs that limit the scope of technology analysis. Such assessments are significant, especially when combined with the realization that methods do not merely describe the world, but also have the power to shape specific realities (Garfinkel, 1964; Latour, 1986; Law and Urry, 2004). Therefore, it is necessary to acknowledge the limitations of methods used to study identification technologies, as they may inadvertently shape the specific realities they seek to describe.

Methods, in this view, can be thought of as devices that bring bits and pieces of the world together to enact certain realities (Latour, 1986; Law and Urry, 2004). As a result, researchers must answer the question of what kind of "ontological politics" (sic.) they participate in and what kinds of realities they contribute to. There are compelling reasons, for example, to use migrants' practices and experiences as a starting point for understanding the forms of discrimination and unpredictability embedded in border technologies. However, this focus can cause one to miss other phenomena, such as the rise to power of a small oligopoly of technology companies and IT consultancy firms in the development of technologies for identifying people (Lemberg-Pedersen et al., 2020; Jones et al., 2022).

What we know about sociotechnologies of identification is primarily based on empirical studies that investigate encounters between people and technologies or based on desk research and document analysis. This section suggests how methodological decisions in studying identification sociotechnologies can shape the research outcomes. Particular attention will be paid to the growing body of literature exploring the information systems that regulate the entry of people into EU territory. There are two main reasons why these systems were selected. First, there is much research on these EU systems, which are among the largest identification systems in the world. Second, the software I looked into, and will expand on later, is directly connected to one of the systems (the Visa Information System). In reviewing the literature on sociotechnologies of identification, it is possible to highlight thematic and methodological similarities and differences that affect the reviewed research findings.

This section first examines how research has dealt with the large-scale nature of identification systems. In other words, how research has dealt with the many people, places, and things involved in making, deploying and using these systems. As we will see, it is possible to discern three different sampling methods to address these scale issues. First, researchers can multiply the number of research sites to account for the dispersed nature of the studied phenomenon. Second, researchers can increase the number of actors in their analysis. Third, and most importantly for this chapter, different moments in a technology's lifecycle can be compared and analyzed. Much of the current literature on identification sociotechnologies focuses on the first and second approaches for dealing with scale issues. Section 3 of this chapter will cover the theoretical foundation and empirical value of using a multi-temporal sampling approach to broaden the analytical reach of IT for border and migration control.

6.2.1 Transverse sampling, or situating and tracing connections across sites

It is common conceptualizing information systems and database technologies that store data about mobile populations as part of larger structures, regimes, systems, infrastructures, and assemblages that bring together border and migration-related practices, rules, and meanings. Researchers are then confronted with the methodological conundrum of localizing and investigating these more comprehensive phenomena. One approach is to multiply the research sites to provide multiple accounts of the connections between sites and unravel distributed phenomena. Such approaches are inspired by multi-sited ethnography (Marcus, 1995). This theoretical framework improved upon the limitations of earlier ethnographic methods, which relied on researchers physically "being there," or observing and interacting with a specific group of people or community in a bounded field site for an extended period. While these accounts may be rich in empirical data, focusing on particular fieldwork locations was deemed insufficient for comprehending globally overlapping phenomena. When ethnographers trace links between sites, they do more than just put one site in a broader global context. In contrast, an ethnographer can decide to focus, for example, on consumption processes within a capitalist political economy by following connections between sites.

For instance, the "Mig@Net-Transnational Digital Networks, Migration and Gender" project (Tsianos and Karakayali, 2010) used a multi-sited ethnographic approach to investigate European border policies and practices by bringing together data from sites and contexts across Europe. Triangulating observations from various actors (migrants, policy experts) and locations (Greece, Germany, and Italy) allowed the team of researchers to show how, for instance, classifications of asylum seekers do not only follow a coherent system of governance. In principle, the Eurodac system categorizes asylum seekers into one of three categories based on whether they (1) regularly apply for international protection, (2) are discovered illegally crossing the border, or (3) are discovered illegally present within a Member State. However, based on interviews with officials conducted in 2011, the researchers discovered that there are differences in how these categories are applied and understood at a national and institutional level. A German contact for the Mig@Net project, for example, asked the researcher why Greece primarily employed the second rather than the third category. The German practitioner's apparent ignorance runs counter to the Eurodac system's purported role in the development of a uniform EU asylum policy. Additionally, by emphasizing the inconsistency and unpredictability of border and migration control, this case highlights how approaches that sample and aggregate observations from multiple sites can provide counter-evidence to the view of coherent migration and border control practices and policies.

Multi-sited research can cut across preconceived groupings such as local sites and larger phenomena (Marcus, 1995). However, caution should be taken not to postulate the existence of actors and phenomena beforehand. The problem may arise when researchers begin with a theoretical construct — e.g., the existence of an EU border regime that criminalizes migration — and investigate it by triangulating results from different sites and

established actors. We differently wonder whether a sense of global phenomena can instead emerge from tracing paths between heterogeneous actors. Actor-network theory, in particular has cast doubt on a priori methods' capacity to demonstrate how scale is accomplished in practice, and argued, instead, that "scale is the actor's own achievement" (Latour, 2005, p.185) and that, hence, theoretical divisions between micro and macro should be dropped (see also, Latour, 1999; Law, 2006). In this way, one valuable methodological insight is to "localiz[e] the global" (Latour, 2005, p.173) by following the "connections leading from one local interaction to the other places, times, and agencies through which a local site is made to do something" (p. 173). In other words, researchers can allow any notion of providing actors and locations structure to come from following connections rather than assuming specific orderings.[1]

Following the multiple circulating standards and categories is one effective way of tracking down these links (Latour, 2005). For example, Pelizza (2021) has documented how regulations for adopting FBI biometric identification standards in equipment used at the Greek border "create trans-national associations with the EU Commission, corporate contractors and the US security regime" (p. 16). Another example is provided by Donko et al. (2022), who have demonstrated how European migration management technology for identification stretches beyond the external EU borders. The authors describe how border checkpoints between Burkina Faso and Niger are linked to EU agencies such as the European Border and Coast Guard Agency (Frontex) via IOM border management information systems that record people's biographic and biometric data. Furthermore, the EU-funded West Africa Police Information System connects these border checkpoints to all INTERPOL member nations via the global police communication system. In order to see the formation of such novel and distinct "flat 'networky' topographies" (Latour, 2005, p. 242) of interactions between actors, it is crucial to avoid a priori postulating them in advance.

So, if we imagine this sampling strategy as a line that crosses and extends across research sites, we can refer to it as *transverse sampling* for studying large-scale data infrastructures. Far-reaching relationships might appear by exploring the lines between local encounters and other places. Tracing these connections highlights the different interconnected actors involved in large-scale data infrastructures.

6.2.2 Perpendicular sampling, or incorporating ecologies of interlinked actors

The need for solutions to scale up the number of research locations to examine large-scale infrastructures runs parallel to the challenge of who should be included in research designs. Scholars have debated the field's proclivity to investigate socially, politically, and geographically marginalized individuals. Contrary to the abundance of research on marginalized individuals, less research has been conducted to "study up" on the wealthy

[1] Ontological stances similar to these have their roots in earlier forms of social theory, such as Gabriel Tarde's "monadology" (see, for example, Latour, 2002).

and powerful (Nader, 1972; Gusterson, 1997). Similarly, many published studies on sociotechnologies of identification focus on the experiences of marginalized persons at the border and the related power imbalances. For instance, a growing body of literature (e.g., Kloppenburg and van der Ploeg, 2020; Kuster and Tsianos, 2016; Olwig et al., 2019) has investigated the contentious processes of collecting migrants' biometric data in border zones. Following the studying up/down analogy, researchers interested in ethnography have tended to concentrate on the perspectives of people who are identified and controlled down at the border and less on those responsible for creating and maintaining these large-scale identification infrastructures.

Alternatively, scholars can arrive at more balanced portrayals of the numerous human and non-human actors involved by concentrating on moments and places where diverse actors interlink and impact one another. Examples in the literature show that researchers can thus include a more diverse set of actors: from interviews with local administrators and officers of international organizations of migration centers (Pollozek and Passoth, 2019) to interviews with officials and experts from European, national, and international institutions (Glouftsios, 2018; Pelizza and Loschi, 2023; Trauttmansdorff and Felt, 2023), and professionals in the security domain at industrial fairs (Baird, 2017). Studies influenced by STS and ANT highlight the significance of non-human actors alongside these human ones (e.g., Pelizza, 2021; Pollozek and Passoth, 2019). Suppose we visualize these approaches as highlighting the multitude of actors whose paths cross at sites. In that case, we can refer to this strategy of increasing the number of actors in studies as *perpendicular sampling*. [2]

The decision of whom to include in the study reflects the questions and politics of the researcher. The "Autonomy of Migration" (AoM) approach, for example, begins with migrants' practices of subverting and appropriating mobility regimes and contrasts them with the Fortress Europe discourse (De Genova, 2017; Scheel, 2019). It has been argued that depicting migration and border controls as fortifications fosters a narrative that ignores the diversity of experiences on borders and migration (Mezzadra and Neilson, 2013), thus instilling a paternalistic view of migrants as helpless victims in need of protection. Authors in this body of work seek to destabilize such tropes by centering on the experience of migrants and illustrating how migrants can circumvent and subvert restrictive migration and border control mechanisms. One may wonder, however, if the emphasis on migrants' practices and tactics to demonstrate that migration politics are "a site of struggle" (Strange et al., 2017, p. 243) does not also contribute to a negative image of migrants as subversive of a rules-based international order. In addition, restrictive definitions of migrants run the risk of excluding other "privileged migrants" (Benson and O'Reilly, 2009), like professionals living abroad, recipients of golden visas, and retirees who migrate. Other approaches provide methodologies that even begin from migrants'

[2] Nader (1980) proposed the concept of "vertical slice" to, for instance, map the various actors, government agencies, policies, corporations, and associations to understand how power and governance of problems are organized (see also Shore et al., 2011). In using the term perpendicular sampling, I aimed to avoid implying that some vertical hierarchical organization exists.

perspectives, such as focusing on their acts and claims of citizenship (e.g., Isin, 2013), or where research adopt a more interventionist stance (e.g., Olivieri, 2023).

As Laura Nader stated in her 1972 article on studying up, "we aren't dealing with an either/or proposition; we need simply to realize when it is useful or crucial in terms of the problem to extend the domain of study up, down, or sideways" (Nader, 1972, p. 8). Since then, the sites and domains of ethnographically inspired research have expanded in many directions. Ethnographers have expanded their work to include tracing the work of scientists in laboratories (for example, Latour and Woolgar, 1986; Gusterson, 1996) and policymakers in governance processes (Shore et al., 2011). Furthermore, these developments coincided with researchers' recognition of other forms of non-human agency. For example, Glouftsios (2021) theorized that the agency of EU information systems, as well as the labor of IT and security professionals to maintain those systems, shapes mobility governance throughout the Schengen area. More researchers are now looking into how devices used in security practices affect political agency due to the realization that technological properties shape practices and are intertwined with effects that shape the world (for example, Amicelle et al., 2015).

Despite these achievements, the diversity of places and actors involved and the specialized and closed nature of border and security work can create barriers for researchers. A report by the "Advancing Alternative Migration Governance" project, for example, describes how the development of EU information systems "has been engineered in specialized and closed forums, such as expert workshops, task forces, technical studies, pilots, or advisory groups and technological platforms steering not just policies, but also the formulation of research and development priorities of funding programmes" (Jeandesboz, 2020, p.10). Moreover, according to the report's authors, the influence of less visible actors, such as global actors who build, fund, and thus profit from border infrastructure construction, needs to be studied more. To support this effort, in what follows, I propose two heuristics that can serve as valuable analytical tools. The first heuristic, based on "interpretative flexibility," allows us to identify significant moments when social groups challenge or transform identification technologies and practices. The second heuristic, employing "gateway moments," helps us understand the compromises and adaptations involved in building identification infrastructures and deploying technologies in diverse contexts. Together, these heuristics provide a robust framework to analyze the development and impact of identification systems, considering the role of under-the-radar actors and their long-term implications for identification practices and infrastructures.

6.3 Multi-temporal sampling, or tracing the genealogies of data infrastructures

The reach of sociotechnologies of identification lies not only in their ability to operate across numerous sites and bring together diverse actors but also in how the technologies develop over time and integrate into infrastructures to have long-lasting effects (see

also Ribes and Finholt, 2009; Karasti et al., 2010). While researchers tend to see rational processes of designing and implementing systems by system builders with well-defined goals, they tend to overlook the contingency in those processes over time. A closer examination of, for example, the well-known Second Generation Schengen Information System (SISII) reveals how the system's creation was nearly derailed by "delays, an escalating budget, political crises, and criticisms of the new system's potential impact on fundamental rights" (Parkin, 2011, p. 1, see also Figure 6.2). In building the SISII, the "instability of system requirements," which includes the European Commission's ignorance of how Member States' end-users use the system, was frequently cited as a cause for delays, according to a report by the European Court of Auditors (ECA, 2014, p. 13). Consequently, research on the sociotechnologies of identification should correct the misconception that designing and building systems are purely rational processes. Instead, researchers must recognize that such systems result from negotiations, adaptations over multiple timescales, and interactions between actors from different organizations (Gasson, 2006; Pollock and Williams, 2009).

Tracing how knowledge and technologies for matching identities emerged and circulated thus necessitates uncovering choices and contingencies in the design of information systems over time. Longitudinal studies, employed to understand the complexity of technological development over time, can adequately account for contingencies, though their feasibility in practice can be hindered by the expansive scope and time commitment they demand. As we know, social constructivists, for one, have long emphasized that the growth and spread of new technologies do not adhere to any simple linear models (e.g., Pinch and Bijker, 1984; Hughes, 1983b). Instead, various factors influence the trajectory, resulting in distinct sociotechnologies for identification. There are likely many forking paths that could have produced different technological outcomes, and avoiding teleological accounts is paramount. For example, Cowan (1985)'s research on refrigerator development showed that the choice of a compressor with its humming sound was not predetermined, as other cooling technologies like gas-based systems were also feasible alternatives. Another example can be found in the research conducted by Pollock and William (2009), which examined the integration of enterprise software systems. The study demonstrated how these integrations are complex and subject to change, without adhering to predictable patterns, but evolving within intricate dynamics between customers and suppliers.

The practical implementation of such approaches is, however, hindered by the extensive scope and time commitment they demand. Yet, by using methodological criteria, it becomes possible to overcome these challenges and identify pivotal moments in the lifecycle of a software. Literature offers limited methodological criteria for recognizing such moments in a software lifecycle. For instance, Hyysalo et al. (2019) suggest identifying "moments and sites in which the various focal actors in the ecology interlink and affect each other and the evolving technology." In general, STS has historically used technological controversies and breakdowns as points in time for understanding how technology

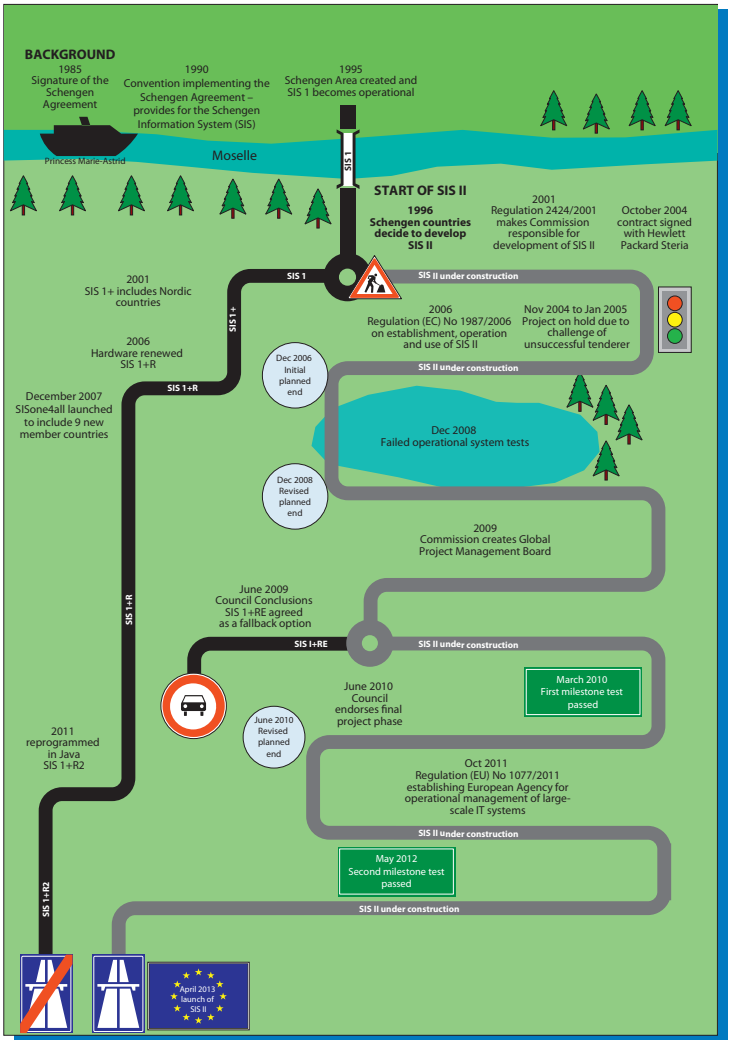


Figure 6.2: Chronology of the SISII (ECA 2014).

functions and the meanings ascribed by actors who are present or who claim to speak for others (Callon, 1984; Marres, 2007).

How can we meaningfully – if not systematically – collect data from various stages in the sociotechnical development of identification technologies? This chapter proposes two heuristics for detecting contingent moments in their developmental trajectory. First, tracing the moments when the meanings of technologies change can help to explain the emergence and establishment of standardized software. Second, tracing the moments when technology connects to other systems can help us understand the unfolding of large-scale identification infrastructures.

6.3.1 Interpretative flexibility and the making of a standard software

Social constructivist accounts of technological innovation provide the theoretical foundation for the first analytical heuristic to identify contingent moments in developing standardized identification systems. Constructivist approaches, such as the Social Construction of Technology (SCOT) and Social Shaping of Technology (SST), have criticized linear and deterministic models of innovation and technological development. Instead, these scholarships have shown how technological development is a long-term and open-ended process in which change can be disorderly and protracted (Bijker and Law, 1992; Bijker et al., 2012; MacKenzie and Wajcman, 1999).

A fundamental premise of constructivist approaches is that technologies aim to solve difficult-to-solve problems with multiple solutions due to competing demands and requirements. Thus, the adaptability of technological designs shapes and is shaped by the interpretation and interest of specific (groups of) actors. To demonstrate this "interpretative flexibility" of artefacts, the basic template for conducting a SCOT analysis calls for first identifying "relevant social groups" (Pinch and Bijker, 1984). Social groups are empirically introduced when a group of actors assigns a particular interpretation or meaning to a technological artefact. As such, a SCOT analysis is interested in how different interpretations of social groups give different problems to be solved by technology. The second step is to analyze how interpretative flexibility decreases through the process of "closure," in which the number of alternative solutions narrows and "stabilizes" into one or more artefacts (sic.). It is important to note that the original SCOT approach has some issues due to its ambivalence to structural contexts and that power imbalances can render some actors invisible (Klein and Kleinman, 2002). However, analyzing changes and closures in the interpretative flexibility of artefacts can be a valuable heuristic for identifying contingent moments of change in the lifecycle of identification sociotechnologies.

Information systems are typically (re)designed to be "generic" and "travel" across organizational contexts (Pollock and Williams, 2009). Such genericity can be seen as corresponding to SCOT's interpretive flexibility. Multiple meanings can be attributed to software, diverse uses are contemplated, and heterogenous problems can be solved with its mediation. Over time, however, interpretive flexibility leaves space for stabilization. Similarly, using the metaphor of a biography, Pollock and Williams (2009) demonstrated

how software suppliers must balance requirements as the software matures and accumulates functionalities through its history. For instance, they found that software may be more adjusted to particular user requirements early in the development process. Later, when vendors want to transfer their software to new customers, they must identify overlaps between the sites' needs. This moment of closure is labeled as "process alignment work" by the authors (p. 174). Power disparities between the supplier and large/small customers may skew the requirements in particular directions, eventually giving shape to best practices and standards.

Interpretive flexibility characterizes also security information systems. Scholars have noted that systems storing personal identity data can easily be used for new and derived purposes aside from their original objectives (Monahan and Palmer, 2009). This type of "function creep" has been referred to for the Eurodac biometric identification system (Ajana, 2013). The system's original purpose was to assist the Dublin system in preventing people from requesting asylum in multiple Member States. However, and as an attestation of the connection between migration and crime control (also known as "crimmigration," Stumpf (2006)), this scope has gradually expanded by allowing police authorities to query the database (Broeders, 2011; Amelung, 2021). As a result, it is essential to consider how diverse security organizations and their systems are interconnected and how this results in the emergence of new and contingent meanings.

The attention to such contingent moments can enable us to discover the "biography" of data matching systems and their standardization by considering tensions between "local" and "global" aspects of software, as well as the links between technical and organizational changes (Pollock and Williams, 2009). For example, one of the recommendations related to the SISII system problems mentioned above was that the Commission "ensure that there is effective global coordination when a project requires the development of different but dependent systems by different stakeholders" (ECA, 2014, p. 07). [3] Therefore, the report's recommendation for how to deal with the issue of end-user's differing perspectives was to establish a new organizational structure that could align the various actors, from Member States to international contracting firms.

Even a single European security artefact can have different meanings and be used differently by diverse states. Soysüren and Nedelcu (2022), for example, compared the deployment of the Dublin III regulation (for the Eurodac system) between a founding EU member (France) and an associated country (Switzerland). The researchers found that France took a more skeptical and decentralized approach to use the Dublin system for deportation, while Switzerland eagerly adopted the Dublin system and implemented it in a highly centralized manner. In this case, interpretative flexibility characterizes a single European security instrument, which can have different meanings and be used differently by diverse countries. The spatial and temporal reach of sociotechnologies of

[3] The report further notes that for the development of SISII a "Global Project Management Board" was established at a late stage in the project to "to draw more fully on the experience of end-users in member countries" (ECA, 2014, p. 37).

identification does not imply that these technologies have necessarily stabilized; instead, the systems may still be implemented in various ways, depending on the context.

Utilizing the SCOT and Biography of Artifacts and Practices (BOAP) approaches to analyze identification software provides several methodological advantages. Firstly, these approaches enable the examination of how the design of identification technology can reach points of closure, wherein customers perceive their problems to be resolved. Secondly, these approaches shed light on the local and global tensions inherent in the development of identification software. They highlight how software systems are designed to be generic and adaptable, capable of traveling across various domains, including security. By studying these tensions, researchers gain insights into the compromises and negotiations involved in adapting identification technologies to different contexts while maintaining their functionality and interoperability. Overall, the SCOT and BOAP approaches offer valuable methodological tools for comprehensively analyzing identification software, uncovering its design processes, and understanding its broader implications in diverse settings.

6.3.2 Gateways to infrastructures of identification

Infrastructure scholars have argued that (data) infrastructures can only "grow" and build up from pre-existing systems, practices, and communities rather than being purposefully constructed (for example, Edwards et al., 2007; Karasti et al., 2016; Monteiro et al., 2013; Star and Ruhleder, 1996). Hence, studies like those on creating the infrastructure to connect scientific communities (for example, Star and Ruhleder, 1996, and Edwards et al. (2007)) demonstrated how challenging it is to build large-scale data infrastructure deliberately. In connecting various IT systems, data infrastructures assemble "a combination of standard and custom technology components from different suppliers, selected and adapted to the user's context and purposes" (Pollock and Williams, 2009, p. 286). At the same time, it is debatable when the assemblage of various elements qualifies as infrastructure. Hence, in light of this argument, Star and Ruhleder (1996) posed the question, "When is an Infrastructure?" According to their argument, the concept of infrastructure is fundamentally relational, as it only becomes infrastructure through its relationship with organized practices.

With several detached systems, it is frequently uncertain which one will succeed or whether other technological and social compromises are necessary to allow systems to work together. Moments where several systems compete are pivotal for infrastructure development, as previously incompatible systems may be able to work and communicate with one another. Edwards et al. (2007) have referred to the phenomenon of making contending systems compatible as a "gateway problem" (p. 34). We can find a paradigmatic example of a gateway technology to solve such problems in the historical development of electricity infrastructure: the innovation of the rotary converter, which made it possible to have co-existing forms of electric power (David and Bunn, 1988; Edwards et al., 2007). The converter qualifies as a gateway technology because it enables compatibility between

competing delivery systems, such as alternating current (AC) and direct current (DC). Modern-day equivalents of such a gateway technology are the international travel plug adapters which enable us to charge our electronic devices in different parts of the world without worrying about the various voltages and plug types used.

More generally, Edwards et al. (2009) define a "gateway phase" as a period "in which technical, political, legal, and/or social innovations link previously separate, heterogeneous systems to form more powerful and far-reaching networks" (Edwards et al., 2009, p. 369). According to a definition provided by David and Bunn (1988), what gateway technologies, in effect, do is to "make it technically feasible to utilize two or more components/subsystems as compatible complements or compatible substitutes in an integrated system of production." Such sociotechnical arrangements, they say, would "permit multiple systems to be used as if they were a single integrated system" (p. 367). Information and communication technologies, for instance, heavily rely on gateway technologies, such as protocol converters that link telecommunications networks with various network protocols. Similarly, Hanseth (2001) uses the term gateway to refer to "elements linking together different networks which are running different protocols" (p. 72). Additionally, Hanseth argues that gateways can be just as critical to the success of large-scale network and infrastructure projects as better-known data standards.

In contrast to standards, gateways have gained less scholarly attention. Moreover, gateways are sometimes considered "as a consequence of a failed design effort as they are imperfect translators between the networks they are linking" (Hanseth, 2001, p. 71). However, as Hanseth rightly notes, gateways can be crucial building blocks for connecting heterogeneous networks into larger-scale infrastructures. He gives the example of health care data, which may be standardized within countries. However, standardizing such data for cross-border data exchange has proven unattainable. A different strategy would be to develop "gateways to enable the (limited, but slowly increasing) transnational information exchange" (Hanseth, 2001, p. 88). Building gateways to facilitate communication between heterogeneous systems is often more manageable than settling on a single standard.

The case of the *EU Digital COVID Certificate Gateway* As an example, look at how the European Union (EU) responded to the Covid-19 pandemic by creating the "EU Digital COVID Certificate Gateway" to authenticate digital COVID certificate signatures across EU member states (European Commission, 2021a, 2022). The European Commission established this gateway in 2021 as a means "through which all certificate signatures can be verified across the EU" (European Commission, 2021b). The EU's member states would have had difficulty agreeing on establishing a central health certificate database during the urgency of a pandemic. Since no personal data would be exchanged via the EU gateway, the system did "not require the setting up and maintenance of a database of health certificates at EU level" (European Commission, 2021b). This choice was also significant for Member States because it allowed them to "retain flexibility in how they link the issuing and verification

> of their certificates to their national systems so long as they meet [the] common standards" (sic.). In those moments of urgency, "most Member States [had] decided to launch a contact tracing" (European Commission, 2020a). Nevertheless, through "decentralised systems," those 20 or so apps could be made "interoperable through the gateway service" (sic.). As a result, a sophisticated contact-tracing infrastructure quickly developed as EU member states (and others) was able to link their national applications while maintaining their national back-ends and data standards. At the same time, this EU contact tracing system revealed variations in how the Member States applied the rules in their domestic context, such as how much time to consider a vaccine's viability before expiration (Calder, 2022). The EU gateway exemplifies gateway technologies' critical but underappreciated role in establishing and maintaining networks within larger-scale infrastructures. It also shows how gateways can be either short-lived or long-lived because the EU Gateway is already offline at the time of writing.

According to Egyedi (2001), there are different kinds of gateways with varying degrees of standardization and, thus, varying degrees of flexibility. Gateways, as per her typology, can be dedicated, generic, or meta-generic. In her view, a dedicated gateway is designed to link only predetermined subsystems and is not or only minimally standardized. She regards the AC/DC rotary converter, for instance, as a specific gateway for converting those two types of currents. On the other hand, generic gateways are standardized and thus can connect an undetermined number of subsystems. We can think of the EU Digital COVID Certificate Gateway (see box) as an example of a generic gateway because it established a common standard that any EU or non-EU country could adopt (European External Action Service, 2021). For example, South Korea, a non-EU country, established a connection with the EU gateway in July 2022 to allow "certificates of vaccination issued in South Korea to be valid in EU countries, and vice versa" (Kim, 2022). Lastly, the best way to understand meta-generic gateways is through examples such as the OSI reference model (ISO/IEC, 1994), which specifies a foundation for computing system communications. These reference models serve as frameworks for developing specific generic standards, rather than defining them. This typology of gateway technologies will help us understand how standardized and adaptable the gateways are in linking heterogeneous identification systems into more extensive networks.

This chapter proposes operationalizing the gateway technology concept as a second heuristic for identifying moments where systems and infrastructures intersect. The term "gateway moment" will be used broadly to refer to instances in which different systems and communities of practice are linked together into larger infrastructures using gateway-like technologies. Such gateway moments are thought to reveal structural constraints that must be reconciled to connect new components in the emergence of identification infrastructures.[4]

[4] Of course, these junctions between new components and existing infrastructures are also prone to failure

The following section uses these theoretical concepts as heuristics to identify points in the lifecycle of a system for matching people's identity data that can provide insight into the evolution of practices and technologies for identifying people in the context of migration, borders, and security.

6.4 Methodology

This section draws on the fieldwork data collected at a software vendor for matching people's identity data in the context of border security and migration management. This section builds on Chapter 5's findings on the specific deployment of the software at The Netherlands' government immigration agency. In addition, the focus on other software deployments in EU and Member State identification systems sheds light on different stages in developing and using the ELISE software. As a result, the study illustrates the diverse set of actors involved in practices of identifying and circulating data about people on the move at the European border (Pelizza, 2019). As detailed in Chapter 3, I joined the company "WCC Group" (WCC) to investigate the design, use, and evolution of a software product dealing with data matching and deduplication. Since I was a temporary member of the ID team, I could visit the company's headquarters in Utrecht (The Netherlands), review all necessary paperwork, conduct one-on-one interviews with relevant company and personnel, and sit in on some of the team's group meetings.

In the course of the research, seven interviews were conducted with individuals from the company WCC, spanning from July 2020 to July 2021. The interviews aimed to illuminate events in the history of their identity-matching software system. I asked people with different profiles about their connections with current and potential customers in the security and identity market. Based on their profiles, we can divide these participants into two clusters. The first cluster comprises WCC's "ID Team" members who hold consultant, pre-sales, and solutions manager positions. The second group consisted of the more technically minded; among them were a senior software developer and a user experience designer. Participants described their knowledge of building and deploying the company's software in six semi-structured interviews, each lasting approximately an hour. In addition, I also conducted observation of several extensive meetings among WCC staff as part of my fieldwork. These meetings served as briefings on various aspects of WCC's solutions and provided valuable insights into topics similar to those explored in the interviews. These observations were documented through detailed field notes.

The interview protocol (included in the Annex) comprised a series of initial questions to understand the interviewee's role at WCC and their insights regarding the challenges and solutions in identity matching. The interviews commenced by inquiring about the interviewee's position and function within the organization, with questions adapted in

(Edwards et al., 2009). Such failures have, for instance, been well documented in e-government and information systems literature more broadly, where failures have been a long-standing concern due to their high stakes and use of public money (Pelizza and Hoppe, 2018).

a semi-structured manner based on the individual's profile and experiences. Participants who had prior involvement with the EU-VIS or MITRE Challenge projects (see below) were presented with tailored questions, as these projects were perceived as potentially pivotal moments in the software's development, the dissemination of data matching expertise, and its securitization. Interestingly, the findings from these inquiries challenged my initial hypothesis, revealing instances where name matching expertise did not consistently circulate as anticipated. The interviews delved into the complexities of matching identity data across diverse organizations and geographical locales, including EU Member States and national or international institutions, and addressed the crucial role of achieving interoperability in identity data. These inquiries, which focused on the software's integration into broader sociotechnical networks, were instrumental in pinpointing and examining gateway moments in the evolution of the data matching software.

In particular, the protocol also encompassed questions designed to extract insights from the interviewee's extensive experience, such as the significance of various data categories in identity matching and the extent customers clearly understand their data matching needs. This line of questioning helped reveal how the company adapted to its customers' specific contexts and requirements, potentially influencing the design of the software to meet these specific needs. Moreover, it shed light on how customers embraced and integrated the proposed data matching software. Consequently, these questions played a pivotal role in uncovering moments of interpretative flexibility within the software, particularly in how customers configured data matching rules — whether they adhered to defaults, followed suggested configurations, or required extensive customization. Consequently, this line of questioning explored the role of the software's defaults and configurability in disseminating data matching expertise. This understanding was instrumental in analyzing the deployments of EU-VIS and IND and discerning their differences in terms of configurations.

Furthermore, the interview protocol probed into the generification of software for identity matching, exploring its adaptability across different domains, including employment and security. These questions were designed to uncover instances of interpretative flexibility within the data matching software as it navigated various domains. More specifically, the questions aimed to gain insight into the software's evolution as it extended its reach into security contexts, notably those within law enforcement and migration management. Furthermore, the interview protocol delved into the evaluations surrounding WCC solutions, examining the concept of "vendor-neutrality" and exploring potential challenges arising from proprietary formats or algorithms. It sought to understand how software solutions like ELISE accommodated diverse data formats. This facet was considered pivotal in comprehending shifts in interpretative flexibility, notably because it involved the incorporation of biometric formats, offering valuable insights into the software's securitization.

The data analysis aimed was to establish links between various data fragments that

document the software's historical development, pinpoint moments of contingency, and construct a narrative that, although fragmented, remains meaningful. Throughout this endeavor, I pursued threads that connected the diverse actors and entities that played roles in the software's development. The company's partnership with Accenture was identified through a notable thread of jointly undertaken projects. To create a structured narrative, I initially organized the data fragments around two overarching themes. First, I traced the software's evolution as it ventured into the identity and security domain, aiming to unveil moments of interpretative flexibility and the contingencies surrounding this transition. Secondly, the investigation extended into the software's assimilation within the EU-VIS and IND systems. By scrutinizing the configurations tailored for these two distinct systems, I identified these instances as gateway moments capable of shedding light on more expansive transformations within the realm of identification.

6.5 Tracing fields of identification through the evolution of software for matching data: Interpretative flexibility moments

This section delves into the biography of WCC's ELISE data matching system, exploring the evolutionary trajectories of data matching within transnational and commercialized security infrastructures. The exploration is divided into two parts based on the analysis using the two heuristic approaches to achieve this as methodological alternatives to conventional longitudinal research. The first part uses the "interpretative flexibility" heuristic to pinpoint moments where social groups have challenged, reshaped, or restricted the meanings associated with the ELISE data matching system. This first heuristic will focus our attention on the company's foundational roots while highlighting the complexities of adapting and generalizing the data matching software for diverse contexts and users. This adaptability ultimately paves the way for ELISE to assume a significant role within international identification systems. The second part uses the "gateway moments" heuristic to direct attention towards instances when ELISE was integrated to connect heterogeneous identification systems. This second heuristic will focus our attention on the role of data matching in bridging disparate identification systems and contributing to forming more extensive infrastructural networks.

6.5.1 Pioneering data matching in the dot-com era

The WCC company's early days reveal a surprising amount of interpretative flexibility regarding what the data matching software should accomplish and for whom it should be helpful. When the company was first conceived in 1996, its founders saw the software primarily as a generic database technology for matching various "things." In conversations with multiple media sources (such as Betlem, 2011), and as shared with me by company personnel, one of the co-founders and former CEO, Peter Went, recalls the genesis of the

product idea as follows. Mr. Went recounts that this concept initially emerged from his experience of encountering unsatisfactory outcomes while searching for a house. Furthermore, he was informed of similar challenges faced by friends during their online job hunts. The primary issue Mr. Went identified was that search results drastically declined when highly specific search criteria were employed. Recognizing this limitation, he saw the need for more advanced search engines capable of meeting users' expectations.

The company's website from 1998, preserved in the Wayback Machine,[5] provides insight into WCC's perspective on their "flexible approach to searching" and the rationale behind why they considered it superior to "traditional searching techniques." This archived webpage offers a valuable glimpse into the company's mindset during this era and sheds light on their approach to data matching and search technologies. Here is the website excerpt that sheds light on why WCC views this approach as a superior alternative (emphasis in original):

> There are a number of drawbacks to using traditional searching techniques. Firstly, traditional searching techniques simply go through a database — in a more or less intelligent manner — looking for a combination of keywords. This combination of keywords is either found or is not found. This type of searching is known as *binary* or *hard searching*. In practice, often a more "soft" and flexible approach to searching is desired. The *fuzzy searching* concept offers more flexibility than traditional (hard) searching. With fuzzy searching techniques, it is possible to define *ranges* in which each search criterion can lie, rather than specifying exact values for each criterion.
>
> Another disadvantage of traditional searching methods is that they perform what is called *one-sided searching*. This means that a search is performed from the viewpoint of one side only. In practice, often a *two-sided search* is wanted that considers the preferences of the supplying side as well as the demanding side.

This excerpt illustrates WCC's alternative interpretation of search technology, setting it apart from what it calls "traditional searching techniques" that necessitate strict adherence to all specified criteria and follow a "one-sided searching" model. In contrast, WCC proposed an alternative approach characterized by fuzzy matching, which advocates the utilization of permissible ranges for each search criterion. Moreover, it introduces a "two-sided search" concept that considers matches between a "supplying side" and a "demanding side." To illustrate, in the context of housing searches, the supplying side may comprise available houses listed on a website, while the demanding side represents individuals seeking a house. Fuzzy matching facilitates the identification of matches that align with specific criteria, such as proximity or budget, without rigidly

[5] https://web.archive.org/web/19981212033959/http://www.wcc.nl:80/s

adhering to precise criteria. This interpretation is embedded in the product's design features, as described on the company's 1998 website:

> ELISE is especially designed for matching purposes. After relevant information has been entered into the system, ELISE is used to find possible matches. ELISE finds matches by calculating *match scores* to quantify the degree of mutual interest between both parties. By definition, a match score will lie between 0% and 100%, with higher match scores indicating greater mutual interest between the parties involved. The one hundred best scoring matches are shown to the user for further (manual or electronic) processing.
>
> Due to the highly specialised nature of ELISE, it finds matches very quickly and very accurately at the same time. This is in contrast with non-dedicated systems that are usually inflexible and slow compared to ELISE. ELISE was designed with optimal speed and flexibility in mind and uses a proprietary database to meet the high-speed requirements set by its match engine(s). It is impossible to obtain these high speeds (thousands of transactions per second) with standard relational database products.

Traditional search methods, as referred to WCC, primarily rely on what are known Boolean expressions, such as is a house less than or equal to the desired price specified in a search, resulting in binary outcomes: either a match or no match. In contrast, WCC's software introduces a probabilistic data matching approach. Under this approach, search results are ranked based on the likelihood of a match, reflecting a notion of "mutual interest between the parties involved."

But who are these parties? Their definition is characterized by interpretive flexibility. According to the 1998 webpage, WCC envisions that "ELISE can basically be used in any markets where products are being offered and demanded." The page went on to specify that WCC was actively addressing the following markets with ELISE: the employment market, real estate, cars, and dating. Moreover, other search systems are depicted as sluggish and lacking in flexibility. This highlights the issue that WCC aimed to address, which is the limitations of conventional relational databases when it comes to efficient data retrieval. These relational databases excel in data storage but tend to struggle with quick data retrieval due to their technical architecture.[6] In response to this challenge, WCC's product employed alternative database technology, which allowed for faster data retrieval and enhanced search performance.

WCC's alternative interpretation of search technology presents a multifaceted approach to the problem. Firstly, it identified the need for a more flexible search and matching engine capable of handling multiple criteria with varying degrees of importance.

[6] To achieve this, WCC ELISE employed in-memory databases instead of the conventional disk-based databases. Additionally, for an accessible overview of the shifts in database technologies, see Dourish (2014).

This approach viewed search as a two-sided matching problem rather than a one-sided search. Secondly, it identified the inefficiencies of search methods based on relational database management systems, deemed too sluggish for the rapidly growing internet-driven economy. Thirdly, it positioned its solution as versatile and universally applicable, designed to streamline search processes across diverse industries where products were both offered and sought after. This perspective is further elucidated in a 2007 blog post by Mr. Went on the company's website, where he touches upon another dimension of the evolving internet landscape. Specifically, he emphasizes the transformation in reliance on search experts, underscoring the shifting dynamics in how people conduct searches and access information in the digital realm[7]:

> [M]arket-focused search engines can sometimes lower the number of total hits to zero. This is because their database technology is too rigid. For example, a user searching cars.com for their dream car, with the exact options they want, under a certain number of miles, within a certain distance from their home, and at a particular price is probably going to receive a "Sorry, but no results were found using the search criteria you entered" message. At that point, the user is forced to adjust the different criteria to see what the limiting factors may be. [...]
>
> Matching technology is rapidly developing a devoted following among staffing agencies, dating services, travel industry, real estate industry, automotive sales, etc. These industries are all built around complex searches that require a complicated database and an industry expert to perform the search. They are quickly finding that our matching technology lessens the dependence on the human expert and provides much more accurate, meaningful search results.

The quoted passage emphasizes the limitations inherent in conventional search methods tailored to specific markets. It suggests that these methods often yield unsatisfactory results when stringent criteria are applied, necessitating users to possess the expertise to fine-tune their searches. In contrast, WCC's proposed data matching solution is presented as a versatile alternative that can be applied across many industries. This approach aims to reduce reliance on human expertise, marking a departure from "market-focused search engines" where search accuracy hinges on the involvement of human experts, such as intermediaries in the travel or real estate sectors. Instead, WCC's approach is portrayed as market-independent and consumer-empowering. This shift reflects a broader transformation coinciding with the rise of the Internet and emerging IT technologies, characterized by removing intermediaries and empowering consumers. During the dot-com and new economy era, technology was expected to

[7] https://web.archive.org/web/20070301063433/http://www.wcc-group.com/page.aspx?page=pagecontent&id=4171069

reshape industries and consumer behavior by leveraging the value of information.[8] For example, technological innovations were anticipated to revolutionize how people searched for new homes, job opportunities, or planned vacations (Benjamin and Wigand, 1995). Consequently, pioneering technologies like advanced search engines were seen as catalysts for redefining or replacing the roles previously held by essential intermediaries, such as travel agents and real estate agencies (Wigand, 2020).

When examining WCC's founders as the initial social group, it becomes apparent that their interpretation of the challenges faced by organizations at that time can be summarized as follows. In order for goods and services to be effectively discovered by customers within the context of an emerging Internet-based economy, it was crucial for organizations to implement flexible search mechanisms. Technically, this problem definition consequently led to the development of the fuzzy data match system as a generic and high performance solution that could be universally applied across various domains. By conceptualizing the problem in this way, the founders aimed to address the overarching need for efficient discovery and enable organizations to adapt to the evolving digital landscape. This approach sought to provide a versatile and adaptable solution that could facilitate the connection between supply and demand across diverse sectors of the economy. The question is whether this interpretation was successfully put into practice or rather challenged by social groups.

6.5.2 Narrowing markets, narrowing design flexibility

Although customers were quite diverse at the time, we could consider them another social group. As such, it would seem that customers accepted WCC's problem definition and the technical solution. For example, based on interview data and the customers mentioned in old WCC marketing materials, we can understand that WCC had customers in diverse domains. At the time, WCC customers used the software solution to match house seekers with suitable houses, job seekers with relevant jobs, wine lovers with wines that suit their tastes, and tourists with their ideal holiday booking. However, despite the variety of industries and sectors served by WCC, not all amounted to a sizable market. Therefore, as the following interviewee recalls, WCC gradually reduced the software solution's interpretative flexibility and redefined the problem by focusing on fewer, more commercially successful domains, such as public employment:

> So, yes, we [WCC] were very broad. The first customer, a major customer, was the Dutch employment agency UVV. And that made us think. Because all those other customers were small amounts. And the UVV was a significant customer, and that convinced management at the time that it was a great match. And the main reason for that was—and we are still uniquely ourselves in that regard even compared to the open source competition you

[8] In a newspaper article Mr. Went was quoted as follows "During the internet bubble, almost every job site [in the Netherlands] was a customer [of WCC]. It was an opportunistic world."

> see now—the bidirectional matching. So ELISE can not only include your own search criteria but also what the other party wants. [...] So what the job needs and what the employee is looking for does matter. And that is then matched with each other and that is what ELISE can do very well. So, that's the reason we entered the labor market. And that has now been completely expanded into much more than just matching wishes with supply, and we are now also solving all kinds of preconditions. (Interview with WCC senior manager, May 31, 2021)

WCC shifted its focus towards a more concentrated set of specific markets while retaining the fundamental features of the original ELISE data matching software, including its distinctive bidirectional and fuzzy matching capabilities. This strategic shift was grounded in market size and the applicability of the interpretation and technological design to these markets. As the interviewee underscores, the bidirectional search design, based on the concept of supply and demand, found particular resonance in certain domains and emerged as a distinctive selling point for this technological feature. Bidirectional matching worked particularly well for public and private employment services that link jobseekers to suitable job opportunities and vice versa. In this context, it leverages fuzzy matching algorithms to gauge compatibility between two sets of data records: job descriptions and job applicants' preferences.

From a technical perspective, employing these functionalities necessitates the organization adopting the data matching system to perform a data mapping exercise. This entails taking the organization's data, often residing in a Relational Database Management System (RDBMS), and aligning it with ELISE's Object Model. The object model transforms data from databases, which are typically organized in the relational model of an RDBMS as a collection of tables, each consisting of rows and columns, into objects with associated properties. In doing so, the data can be precisely characterized in accordance with the supply and demand attributes integral to WCC's data matching model, facilitated through a programming interface. Subsequently, the data originating from the organization is synchronized with the ELISE database using a tool known as the ELISE Data Replicator. This synchronization process ensures that the data is up-to-date and consistent between the organization's systems, where data is stored, and the ELISE database, which facilitates the searching and matching. Together, these steps make the organization's data available for the ELISE data matching system.[9]

Customers, as a social group, expressed their requests in confidential interactions with vendors and clients, but evidence of these requests can be found in other documents and reports. For instance, the 2003 annual report introduces the ELISE Data Replicator as "a powerful tool to automatically synchronize ELISE with the most complex database designs." This description implies that customers needed a means to map their unique database structures to the ELISE data matching system, likely for scenarios that WCC

[9] According to the 2003 Annual Report the Object Model and Replicator were introduced in that year (WCC, 2003)

did not always anticipate. Consequently, the Data Replicator can be interpreted as a response to accommodate the diversity in database designs sought by customers. Similarly, the introduction of the ELISE Data Model can be attributed to customers' demands for an effective mapping solution to the data matching engine. The stability of these two components suggests that social groups have come to perceive those issues as resolved.

At this juncture, the software design had reached notable points of closure. The ELISE object model and replicator remained core components of the ELISE solution, even though, as we will soon discover, markets were undergoing significant shifts. Despite the fact that the range of applications for the searching and matching tool had narrowed, moving away from encompassing diverse contexts like e-commerce and increasingly focusing on employment and identity domains, the core data matching system itself retained a notably "generic" and context-independent nature. Rather than diversifying the core technology, the company embarked on the creation of more dedicated, context-specific platforms that were built upon this "generic" ELISE system. Despite the varying degrees of success experienced by these context-specific applications, the foundational principles of the data matching technology, which revolved around mapping and replicating data within the ELISE object model for matching purposes, remained rather stable.

6.5.3 Expanding data matching horizons in the post-9/11 landscape

As the interpretative flexibility of the ELISE system began to diminish, the 2002 Annual Report highlights a concurrent expansion into the domain of law enforcement. This report delineates WCC's target markets as Employment, Crime Fighting, Travel, and Other (WCC, 2002). Notably, previously listed markets such as "dating, real estate, used cars, pharmacy" were reclassified under the broader category of "Other" and were no longer actively pursued. The addition of "Crime Fighting," later renamed "Law Enforcement" in 2003 (WCC, 2003), to the company's markets is described in the report as follows:

> WCC has entered this new market [Law Enforcement] in 2002 inspired by the focus worldwide on crime fighting and anti-terrorism after the 9-11 attacks in 2001. WCC has developed, in collaboration with one of Europe's leading Forensic Science institute, an application for matching (crime scene) DNA strings with existing DNA profiles. The mutual expectations with respect to this market are high, because in our opinion DNA and DNA matching significantly enhance the results of forensic research and crime prevention. (p. 9)

The 2002 annual report notes how this feature arose out of a pilot project of student's thesis:

> WCC also met serious interest in a relatively new area, DNA-matching. A request from a student for an internship led to building a prototype for DNA

> matching as subject of his thesis. The prototype that was build shows impressive results and has attracted a, sought after, potential launching customer. Early 2003 the final prototype version should enable us to show the powerful solution ELISE has to offer to the crime fighting industry. (p. 19)

From the excerpt we can deduce that, following a renewed focus on crime fighting and anti-terrorism efforts following the tragic events of September 11, 2001, new social groups such as the student and the European "leading Forensic Science Institute" started to re-interpret the problems to be solved by data matching: matching DNA profiles found at crime scenes with DNA profiles stored within a database. By reintroducing interpretative flexibility, WCC had to adapt ELISE's data matching capabilities.

These new meanings can be linked to re-interpretations of two aspects of the original ELISE system's design. Firstly, DNA matching deviates from the bidirectional supply and demand model that characterized the software's original design. Forensic investigations involve a unidirectional process, focusing on matching crime scene DNA strings to existing DNA profiles rather than the reverse. From a technical standpoint, ELISE could be readily adapted to maintain its bidirectional search model while only executing the matching in one direction. Secondly, DNA samples can be represented as data strings through a process that translates the DNA code into sequences of characters, much like textual or numerical data strings. This representation aligns with the principles of data matching, which involve comparing one data string with another to identify patterns, similarities, or matches. From a technical standpoint, this compatibility likely facilitated the seamless integration of DNA matching into ELISE.

The integration of DNA matching into ELISE's data matching capabilities can be viewed as a larger reintroduction of design flexibility within the data matching system. The adaptations reflect a response to reinterpretations in matching various types of data for identification purposes. This shift is evident in a 2005 webpage where ELISE's matching capabilities were reimagined to encompass the "recognition of objects" (WCC, 2005). The website describes this feature as follows:

> ELISE enables recognition (matching) of objects in massive amounts of data in a sub second response time. In fact it does not matter what type of data is used; finger prints, pictures of faces, DNA structures, ELISE is able to find the best matching profiles based on data containing millions of profiles and thousands of characteristics per profile.

The transformation in the design of ELISE is notable because it extended matching capabilities to new forms of data and reverted to a uni-directional search. First, these new forms of data handle what is referred to on the web page as Binary Large Objects (blobs), signifying a departure from solely matching textual data. Now, ELISE could match not only text-based information but also images of faces, fingerprints, or diverse biometric profiles. Data from this array of sources could be effectively modeled and integrated into the data matching system. This expansion was presented as a development

that opened the door to practical, cross-disciplinary solutions. The company's website highlighted a myriad of potential applications, including crime matching, DNA analysis, fingerprint matching, disaster victim identification, stolen art recovery, car theft tracking, missing children locating, financial misdemeanors prevention (such as credit card fraud), anti-corruption efforts, and combating child pornography.

Second, when dealing with such identity data, data matching tends to revert to a unidirectional process, as the concept of supply and demand mapping no longer applies. A person's identity record typically has no requirements to match, unlike a job announcement in a database that specifies who should apply. Still, the company could solve both problems using the same data matching engine, despite different problem definitions between the employment services domain and the identity and security domain. The company could translate and reconcile such data matching across contexts, even with these different problem definitions of the customers as relevant social groups. Because unidirectional search is just a tweaked version of bidirectional matching, there is no technical incompatibility between the two designs. As the software can function without the user being aware of this difference, the software design has reached a point of closure as customers perceive their problems to be resolved.

In the post-9/11 era, with a heightened global emphasis on crime fighting and counter-terrorism efforts, data matching underwent a significant reinterpretation within a context that demanded interpretative flexibility. Rather than merely applying data matching to markets where products were offered and demanded, it was re-conceived as a potent tool for addressing critical challenges in policing and security. This shift in meaning necessitated the system's adaptation to handle various forms of data beyond the typical text-based matching found in HR and staffing or e-commerce. Specifically, it required the capacity to work with binary data, notably biometric images and profiles. These binary data forms were made compatible with the ELISE data matching, for example, by converting them into text data, which the system could process for matching purposes. For instance, a fingerprint scan can be digitally processed to generate a biometric template, a collection of extracted characteristics that could be stored and employed for matching.

6.5.4 Cultivating identity matching and international professional networks

The further expansion of WCC ELISE into identity matching was closely intertwined with collaborative efforts involving another social group: external partners. These partnerships encompassed various entities, including system integrators and software partners responsible for crafting tailored solutions for identity-related markets and technology partners providing crucial identity matching components, such as biometric matching capabilities. Furthermore, partnerships in data matching solutions in other markets facilitated WCC's expansion into security and identity markets. One illustrative example is the company's venture into the security and identity sectors, closely linked with collaborative efforts in data matching solutions for employment services. In the early 2000s,

WCC joined forces with Accenture, a global IT services and consulting powerhouse, to introduce a data matching system for a web platform for the German public employment service. Peter Went, who served as WCC's CEO at the time, underscored the achievements of this collaboration in a 2006 interview published in "Database Magazine." He specifically highlighted how this collaborative effort propelled the company into the field of identity matching:

> WCC entered that world [identity matching] through a successful trajectory with Accenture at the Employment Service in Germany. The response was so positive that Accenture decided to hire WCC for a huge project that the consultancy won in 2004 with US Visit, the US border security company. "They search there, as every traveler to America knows, by face and fingerprint. Ideally suited for our ranking technology, because there are no perfect Boolean-true matches with biometric data." [Peter Went] (Rippen, 2006, p. 37, translated from Dutch)

The 2003 WCC annual report described that an Accenture-led consortium secured a contract to develop and maintain a "Virtual Job Market" portal for the German Department of Labor, leading to a substantial ELISE license agreement with WCC. Remarkably, this virtual job portal was launched in under a year and garnered praise for its robustness and high-performance capabilities, efficiently handling significant loads of data. Recognizing the success of this partnership, the 2003 report notes that WCC and Accenture formalized their collaboration by entering into a "global alliance agreement," described as follows:

> Because of this success in Germany and the added value for the clients of such a combination, WCC and Accenture formalized the cooperation into a global alliance agreement. This alliance agreement emphasizes WCC's credibility for large projects all over the world. (p. 18)

If we consider Accenture as a social group, the partnership suggests that ELISE emerged as a trusted tool for effectively handling large-scale data matching projects globally, positioning ELISE as a dependable data matching solution for endeavors worldwide. Following the fruitful collaboration on the German employment service project and establishing the global alliance agreement, WCC and Accenture embarked on several substantial projects, notably also in identity matching, encompassing biometric and non-biometric data. In 2004, Accenture invited WCC to participate in a contract awarded to the consulting firm for the United States Visitor and Immigrant Status Indicator Technology (US-VISIT) system (Rippen, 2006). Due to confidentiality constraints, I lack access to specific details, and it remains uncertain whether the ELISE system was indeed employed in this context. Nonetheless, it is worth noting that, as reported in a newspaper article, Accenture played a pivotal role in introducing WCC to the field of biometric analysis during this time.

In 2012, the European Commission selected a consortium of companies, including Accenture, Morpho, and HP, to maintain the EU Visa Information and Biometric Matching Systems, with WCC serving as a subcontractor tasked with providing the search engine for alphanumeric data (Accenture, 2012). This collaboration continued, with WCC remaining a subcontractor to Accenture, furnishing the search and match solution for biographical and biometric data within the UNHCR's Identity Management System in 2015 (Accenture, 2015). These collaborative initiatives highlight how, for Accenture as a social group, WCC became a dependable vendor of data matching solutions, contributing significantly to the system's evolution through its recurrent deployment across various international projects. However, this expanded partnership also necessitated addressing the challenges associated with increased scale of data to match, as exemplified by the release of ELISE version 5 during the German public employment service project (WCC, 2003). This release highlights enhanced performance, scalability, and fault tolerance, reflecting the growing requirements stemming from the increased size and complexity of data to match in these projects.

In the post-9/11 world, characterized by a renewed emphasis on anti-terrorism and border security, there was a rapidly expanding market for biometric technologies tailored for data matching in these contexts.[10] Simultaneously, it is essential to recognize the prevailing uncertainty before and immediately after the 9/11 attacks. A glance at the 2001 Market Review of Biometric Technology Today reveals the industry's tumultuous state (Biometric Technology Today, 2002). The challenging U.S. economy significantly influenced numerous biometric companies even before the attacks. However, the aftermath of 9/11 brought about a remarkable shift in the biometrics industry. While share prices in other sectors plummeted due to the attacks, some biometric companies experienced meteoric rises as investors anticipated heightened demand for high-security products (see also, Amoore, 2006; Lyon, 2003). This period marked a significant turning point in the industry, with increased interest in biometrics despite economic challenges. Examining WCC's evolution and its product ELISE provides a lens through which we can observe the evolving alliances and transnational networks that emerged in this burgeoning market for data matching technology, especially in the realm of security, where novel interpretations of data matching and design solutions rapidly co-evolved.

6.5.5 Embracing multi-modal matching and pursuing interoperability

During the transition towards data matching for security purposes, there was a contingent moment marked by interpretative flexibility and a redefinition of data matching. This redefinition expanded the scope to encompass data matching from various sources and biometric modalities. In this context, "biometric modality" refers to distinct categories of biometric data used in biometric systems, including fingerprints, facial fea-

[10] In a newspaper article notes that "WCC founder Peter Went (49) admits honestly that the attacks on the WTC have created new opportunities for his company." (Betlem, 2011).

tures, iris patterns, voice, and DNA. The concept of "multi-modal matching" emerged, signifying the simultaneous utilization of multiple biometric modalities for identification purposes, often complemented by matching with biographic data. This shift in the design and functionality of data matching systems was closely tied to establishing partnerships with external collaborators because extracting features from raw biometric data often relied on proprietary technologies provided by third-party sources. Furthermore, this transformation aligned data matching with the critical task of supporting counterterrorism efforts and crime prevention by enabling the matching of information from diverse sources, such as from different government agencies.

These transformations are exemplified in a 2009 WCC position paper titled "Homeland Security Presidential Directive 24 (HSPD-24): A layered approach to accurate real time identification." This paper describes how the ELISE software could be used to comply with HSPD-24, a framework for interagency cooperation and interoperability of biographic and biometric data as part of US counterterrorism efforts and screening processes against terrorism watchlists. Here is the purpose, as defined in the directive titled "NSPD-59 / HSPD-24 on biometrics for identification and screening to enhance national security (Bush, 2008):

> This directive establishes a framework to ensure that Federal executive departments and agencies (agencies) use mutually compatible methods and procedures in the collection, storage, use, analysis, and sharing of biometric and associated biographic and contextual information of individuals in a lawful and appropriate manner, while respecting their information privacy and other legal rights under United States law.

The excerpts below from WCC's position paper demonstrates the company's response to the evolving redefinition of data matching, as the paper "explores the ramifications of HSPD-24 and explores its implications for the matching software that supports these processes, with a close look at how WCC's ELISE ID supports the layered approach" (WCC, 2009a):

> HSPD-24 recognizes that technological progress and real-world implementations have substantially advanced in recent years, but also that a lack of biometrics standardization and the existence of conflicting mission security rules limit data-sharing among federal agencies. It further acknowledges that biometrics is only one of several layers of identifying data, and that a layered approach instead of a single mechanism — is needed to improve the executive branch's ability to identify and screen for persons who may pose a national security threat. (WCC, 2009b, p.1)

> While HSPD-24 does not provide any definition of a layered approach, it is understood that it refers to successively applying any or all available biographic, biometric, and contextual identifying data in order to arrive at an

> informed and accurate decision about a person's identity. This process is commonly known as identity matching, but until now identity matching solutions were primarily single-layered or siloed approaches that used a single biometric modality or a single factor such as a name to perform the identification. (WCC, 2009b, p. 3)

The provided excerpts shed light on several facets of the directive and the evolving definitions of data matching. Firstly, there is a redefinition of data matching that involves pooling data from diverse government agencies to leverage existing information to identify known and suspected terrorists. This approach hinges on interagency cooperation and interoperability to enhance the efficiency of terrorist screening processes, yet it faces challenges associated with data-sharing constraints between these government entities. Secondly, there is a redefinition of data matching that emphasizes adopting a multidimensional data approach for identifying security threats. This leads to the proposal of a layered identification strategy that combines various forms of data, including biometric and biographic data, alongside other factors. The challenge in this context lies in the usability of these data forms for matching purposes, particularly given that biometrics often rely on proprietary technologies.

The second excerpt from the position paper highlights an intriguing observation: "HSPD-24 does not provide any definition of a layered approach" (WCC, 2009b, p. 3). This statement underscores the inherent need for entities like WCC to partake in the interpretation of government directives, including their objectives, challenges, and intended outcomes. In response to this interpretative flexibility, WCC introduces a technical design referred to as a multi-modal fusion solution, which boasts the ability to achieve "high accuracy with a large number of criteria by fusing individual match scores" (p. 8). In short, this design seeks to amalgamate or "fuse" match scores obtained from distinct sources, including various biographic data and biometric modalities, each assessed using different algorithms. The intention behind this approach is to address existing challenges and enhance accuracy when compared to relying solely on a single biometric modality. For example, when an individual's data aligns with fingerprint and iris readings, the likelihood of accurately identifying that person increases, especially compared to depending solely on fingerprint matches or even in cases where biographic data may differ considerably.

The directive highlights another pressing challenge: the need for standardized and interoperable biometric technologies. Many available biometric technologies utilize proprietary algorithms to generate profiles from raw biometric data. This diversity of formats and proprietary nature complicated the integration of different biometric modalities, hindering the development of a cohesive and interoperable identification system. In response to these emerging challenges, WCC introduced a new "vendor-neutral and future-proof" software architecture designed to allow seamless integration of new biometric standards "as soon as they are ratified and deployed" (WCC, 2009b, p. 7). This architecture was engineered to address the challenge by enabling the plug-in of other

vendors' biometrics through Software Development Kits (SDKs), facilitating the computation and fusion of match scores. This evolution in the design and functionality of data matching systems closely ties to establishing partnerships with other technology companies,[11] highlighting the interplay between technology providers in this dynamic field.

The possibility of new social groups forming and reintroducing interpretive flexibility means the closure and reduction of design flexibility are temporary. Accordingly, the issues and solutions of identification were once again open to interpretive flexibility in the post-9/11 era. The United States government, as a social group, re-problematized policy problems of identifying people in the context of security as a problem to be solved with new technical solutions. Hence, the significance of biometrics is growing, along with the need for data interoperability between various agencies as technological solutions to identify potential threats. The development of biometric and multimodal technical solutions evolved alongside the changing demands of identification practices in exchanges between government and business actors. In response to new problematizations, WCC proactively incorporated new features into the ELISE software solution to accommodate both biographic and biometric data.

6.5.6 Navigating the MITRE Challenge and the complexities of name matching

In 2011, there was a reintroduction of interpretative flexibility in the field of data matching, particularly concerning new problem definitions related to matching biographic data, especially names. The MITRE Corporation, a non-profit research organization serving the US federal government, orchestrated an international competition to identify the most effective solution for "multicultural name matching" (Miller et al., 2012-05-23/2012-05-25). The "MITRE Challenge," as it was termed, aimed to expedite technologies capable of addressing tasks involving identifying individuals referred to by various name versions or spellings. While the potential applications were considered diverse, the initial announcement highlighted MITRE's assistance to "the Department of Homeland Security and other national security sponsors with their identity matching needs," which can be assumed to include, for instance, watchlist matching. The following excerpt provides insights into the event's initial description (The MITRE Corporation, 2011):

> Challenge #1 entails multicultural name matching—a technology that's a key component of identity matching and involves measuring the similarity of database records that refer to people. Uses for this technology include verifying eligibility for Social Security or medical benefits, identifying and reunifying families in disaster-relief operations, vetting persons against a travel watchlist, and merging or eliminating duplicate records in databases.

[11] A webpage from 2019 on WCC's website featured a roster of technology partners specializing in biometrics, including: SecuGen, NEC, Toshiba, Iris ID, Cognitec, Warwick Warp, and Genkey (WCC, 2019).

The inception of this idea drew inspiration from the "Netflix prize," a competition hosted by the emerging streaming service at the time. In the Netflix competition, participants were encouraged to submit technical solutions capable of surpassing Netflix's movie recommendation system, utilizing a designated dataset. Similarly, MITRE devised a web platform where contestants could attempt to improve upon a baseline algorithm for name matching. The process was designed to be open and inclusive, allowing anyone to submit their responses to foster innovative and diverse solutions to the problem of matching names. The following excerpt illuminates how the initial announcement described the process (The MITRE Corporation, 2011):

> Anyone can join the Challenge—academic institutions, commercial companies, government laboratories, and individuals. Participants must match a query file and an index file, each containing a list of names, against one another to produce a list of scored matches for each query name. Registered teams receive a dataset and task guidelines, submit responses, and receive immediate feedback on their performance.
>
> The names of the best performing teams will be posted on a continuously updated leaderboard. The team that yields a reproducible result that demonstrates the greatest performance improvement over the baseline algorithm (Mean Average Precision) will be declared the winner and will have an opportunity to present at a MITRE/government technical exchange meeting.

The MITRE team provided two comprehensive datasets containing names, encompassing variations from diverse cultures and languages typically encountered in population registers. The challenge for participants revolved around matching names from a relatively compact "query list" with those from a substantially larger "index list" (both with names in Latin script, so potentially already transliterated). Contestants used these datasets to test their technical solutions aimed at identifying name matches between the query and index lists, with the goal of surpassing the performance of the baseline algorithm. Subsequently, participants would upload their outcomes to a dedicated online platform, affording them the opportunity to monitor their progress on a publicly visible leaderboard. The top-performing team or teams would have the privilege of attending a meeting hosted by MITRE alongside government representatives.

MITRE and its sponsors, US government agencies, organized the challenge as part of their broader initiative to "evaluate commercial off-the-shelf and government identity matching and resolution tools to help maximize [MITRE's] sponsors' effective use of the technology." For WCC, this competition presented a valuable opportunity to showcase their ELISE software solution and potentially attain the invitation to present before prominent US government agencies and before "the three letter agencies in America" (Interview 2021-05-31). These agencies include the Federal Bureau of Investigation (FBI), the Drug Enforcement Agency (DEA) and the Central Intelligence Agency (CIA).

By submitting a solution powered by a specially configured ELISE, as we will delve into shortly, WCC demonstrated exceptional performance in meeting the competition's challenge, earning them recognition as one of the esteemed "Top Tier Vendors" (PRNewswire, 2011) and securing the coveted meeting.

WCC's participation in the MITRE challenge necessitated a meticulous reconfiguration of the ELISE system to align with the novel task of matching names, which was not a primary feature of ELISE at the time. While ELISE could perform name matching to some extent, it had not been honed for this specific purpose. During my fieldwork with WCC, I had the privilege of speaking with one of the senior developers involved in crafting the solution for this challenge. The following are paraphrases from this conversation on how they addressed the challenge:

> At a certain point, the MITRE Challenge came along, which we also used to find out how to improve name matching for ELISE. Because that challenge was specifically concentrated on names. It wasn't about finding identities, but really just names. And then we thought to see how far we can get with ELISE. Our system was surprisingly good at first, but not optimal. So we decided to start a dedicated project to look at what we could incorporate in order to improve the name matching in ELISE and get a better score for the MITRE Challenge.

For WCC, the MITRE challenge represented more than just an opportunity to showcase their capabilities; it was a chance to evolve and enhance their data matching capabilities, particularly in the domain of name matching. While the core ELISE system was functional, it required fine-tuning to perform well for this specific task. However, the challenge was compounded by the fact that MITRE's scoring system provided a final score for submissions but lacked detailed feedback on potential areas for improvement. This meant that the development team had to engage in a process of trial and error, making various adjustments to optimize their solution and achieve a higher score.

> The score after submission did not give insight into what were good results and what were bad results. So we really just tweaked our submission here and there. We tested different adjustments and checked their impact on our ranking. So this made it possible to understand what mechanisms had a negative effect and which has a positive effect. So it was really a kind of puzzle, not a very scientific project. But in the process, we were able to put all that knowledge into the ELISE.

The experience gained through submitting solutions to the MITRE challenge was influential in developing new capabilities for name matching within ELISE. This name matching encompassed intricate adjustments within ELISE, such as fine-tuning the weights to determine the significance of factors like typos or name switches in the matching process. Moreover, the solution submitted by WCC extended its reach beyond

the confines of ELISE, incorporating external name databases as in the name matching process:

> Most of the knowledge comes from external source, name databases as we call them in ELISE. Those databases store facts about which name or name parts map on each other with a certain percentage. And of course we also gained insight into how important, for example, typo corrections on a name or phonetic variants are. Or how to deal with the name mix-ups, such as when a first name was put as a last name or vice versa.

WCC's submission for the MITRE challenge entailed the integration of "name databases" into ELISE.[12] This incorporation of name databases proved necessary due to the complexity of the MITRE datasets, which featured diverse name types, including variants of East Asian and Arabic names. Matching such names and variants can pose computational challenges that require a specialized approach to ensure accurate matching. Such name databases are provided by specialized organizations, who create and sell repositories of name variations based on linguistic insights as opposed to purely computational methods. Among these integrations was, for instance, the "Database of Arabic Names (DAN)" created by the organization called "CJKI Dictionary Institute," who describe the database on their website as follows: "DAN covers over 6.5 million entries and consists of Arabic personal names — both given names and surnames — and name variants mapped to the original Arabic script with a large variety of supplementary information."[13] CJKI licenses these name databases, offering the possibility of procuring licenses that enable their integration into a diverse range of software programs, including WCC's ELISE system.

The challenges posed by the MITRE competition, which treated name matching as a distinct and complex data matching problem, brought forth renewed interpretative flexibility for WCC's ELISE system. The system had to be adjusted and enhanced to address the diverse names of different cultures and regions. This undertaking highlighted the fusion of diverse knowledge domains, offering valuable insights into the sociotechnical networks involved in developing data matching knowledge and expertise. For instance, name databases like the one provided by CJKI are frequently leveraged within security contexts,[14] underscoring the broader connections between these developments and the evolving security landscape.

[12] A more encompassing term for these databases is "lexical databases," a concept commonly used in natural language processing. Lexical databases comprise collections of interconnected words and their relationships within natural language.

[13] https://www.cjk.org/data/arabic/proper/database-arabic-names/

[14] According to the website, the DAN database, for example, "plays an important role in helping software developers, especially of security applications related to anti-money laundering and terror watchlists, as well as natural language processing tools, enhance their technology by enabling named entity recognition and extraction, machine translation, variant normalization, and information retrieval of Arabic names" (CJK Dictionary Institute, 2018).

Furthermore, the challenge shaped WCC's strategy in presenting ELISE as a specialized solution for "multi-cultural name matching." A case in point is the evolution in ELISE's advertised capabilities.[15] For instance, the 2010 version of the WCC homepage lists one of ELISE's capabilities as "Name & address matching." However, by 2012, the website incorporates references to the MITRE challenge and offers a product information sheet related to "Multi-Cultural Name Matching." Similarly, NetOwl NameMatcher, the winner of the MITRE challenge, also underlines how their product "Achieves high accuracy as evidenced by winning the MITRE Multicultural Name Matching Challenge." At the time of writing, NetOwl's product continues to focus on name matching, emphasizing "multi-ethnicity name matching challenges" and "multilingual and cross-lingual matching."[16],[17] In this context, the MITRE Corporation introduced an "obligatory passage point" (Callon, 1984), where achieving a commendable score in the MITRE challenge has been employed to establish the credibility of name matching technologies. This underscores the significance of the challenge as an influential juncture in the development and validation of data matching technologies, leading to both technical advancements and heightened interconnections between research, governments, and industry.

6.5.7 Evolving landscapes of data matching

This exploration of ELISE's interpretative flexibility provided a window into the ever-evolving landscape of problem definitions and design solutions for matching identity data. Initially conceived as a versatile technology applicable to a wide array of markets on the internet, ELISE found its niche within diverse social groups with unique challenges, notably in job matching. As ELISE's core design solidified, its interpretative and design flexibility diminished. Over time, WCC shifted its focus to two primary data matching markets: public and private employment services on one front and the identity and security industry on the other. The latter pivot was catalyzed by geopolitical shifts and heightened global security concerns, where data matching technology was increasingly viewed as a valuable asset for law enforcement and border security. This transition brought about newfound design flexibility as the data matching system expanded to encompass biometric data matching and addressed specific identity matching intricacies, exemplified by name matching. By examining the various instances of interpretative flexibility within the ELISE system's evolution, it becomes evident that its original intent was never

[15] A 2010 version of the WCC homepage lists as one capability of ELISE as "Name & address matching." In 2012, this capability was still the same but there was a note about the MITRE challenge and a download related to "Multi-Cultural Name Matching" https://web.archive.org/web/20120322012217/http://www.wcc-group.com/. In 2016, "Built-in Name Matching" is stated as a specific feature described as follows: "Many years of research & development have lead to advanced algorithms that make the matching process much more reliable as it includes multi-cultural name conventions and much more." https://web.archive.org/web/20160531183140/https://www.wcc-group.com/identity-matching/border-management

[16] https://www.netowl.com/name-matching-software

[17] Another company Rosette also has dedicated Name Matching which includes "cross-script and cross-lingual matching" https://www.rosette.com/name-matching-algorithms/.

centered on identity matching. Instead, this transformation occurred gradually and contingently, influenced by many parties and factors over time. Of particular significance is the shift away from supporting a wide array of social groups utilizing ELISE for various purposes like housing and car searches, with the predominant focus shifting towards public and private employment, as well as identity and security contexts.

The exploration of interpretative flexibility has made it possible to highlight points at which governmental and private entities collaborated, co-producing and collectively shaping the challenges and solutions within the realm of identity data matching. On the one hand, this analysis has illuminated the formation of international professional networks, which secured contracts for developing identification systems in various national and international contexts. Consequently, WCC's product found itself repeatedly deployed across diverse international landscapes. On the other hand, the focus on interpretative flexibility underscored the development of data matching methodologies and technologies as a contingent and dynamic process. For instance, the HSPD directive and the MITRE challenge exemplified redefinitions of identity data matching issues, reintroducing interpretative flexibility. WCC's responses to these new actors, linked to the U.S. government and its security agenda, included proposing a plug-in architecture for biometric standards and devising innovative approaches to name matching. These moments highlight the reintroductions of interpretative flexibility and system adaptations, driven by evolving demands and problematizations within the domain of security and identity data matching.

WCC's recent strides in the identity and security sector have primarily revolved around the utilization of the ELISE system in two core domains: passenger screening and civil registrations. However, this narrative did not incorporate these recent developments due to their distinctive nature. Although built upon the ELISE system, they are effectively separate software applications, warranting individual analysis. In contrast to the past, when WCC primarily offered the ELISE data matching system as a back-end solution, the company now crafts comprehensive application packages encompassing back-end and front-end components. The design and functionality of the ELISE data matching system, as outlined in this section, have reached a point of design closure as these new applications are constructed upon the foundation of the current ELISE system.

Methodologically, this multi-temporal sampling approach, employing interpretative flexibility as a heuristic, has provided a valuable lens through which to understand how the ELISE system's design was contingent upon the specific circumstances and actors involved in its development. Instead of viewing the data matching system as a predetermined outcome within the realm of identity data matching, particularly in contexts such as migration and border control, we have unveiled how its securitization was profoundly influenced by specific choices made by actors and the evolving sociotechnical landscape. As an alternative to longitudinal research, this sampling method draws attention to analytically relevant moments that reveal the dynamic construction of technology and the

intricate interplay between social groups, technological artefacts, and evolving problems within the field of identity data matching.

6.6 Gateway moments

Examining moments of interpretative flexibility has allowed us to comprehend the dynamic evolution of the ELISE system, tracking its shifts in design flexibility as it ventured into novel markets and encountered fresh challenges, particularly in the context of identity data matching. However, this analytical lens is less apt to investigate the system integration processes, wherein data matching is used for interconnecting system components with more extensive infrastructures. Our second heuristic, centered on gateway moments, offers a complementary approach to identify contingent moments in the long-term evolution of the data matching software. This perspective will illuminate the nuanced interplay between technology and the broader networks it becomes embedded within, which are not visible by only looking at shifts in interpretative flexibility.

6.6.1 The VIS evolutions project and the problems of backwards compatibility

The previous section described how WCC's ELISE system was integrated into the EU Visa Information System (VIS), underscoring the company's growing prominence within international professional networks. When the European Commission selected a consortium comprising Accenture, Morpho, and HP to maintain the EU's visa information and biometric matching systems, WCC's ELISE system was chosen to power the VIS's searching and matching capabilities. The significance of this selection becomes apparent when considering the scale of the VIS. As detailed in the "Report on the technical function of the Visa Information System (VIS)," (eu-LISA, 2020) the central VIS system handled 17 million visa applications in 2019 alone, registering extensive personal data from non-EU citizens. According to the report, this data was subjected to approximately 25 million alphanumeric searches, illustrating its crucial role in identification within the VIS. On the one hand, their inclusion into the consortium exemplified WCC's reputation as a dependable partner for delivering efficient and effective data matching solutions that could meet the demands of data matching at this scale.

On the other hand, examining ELISE's role within the EU Visa Information System (VIS) through the lens of a gateway moment reveals a different perspective, particularly the challenges associated with interconnecting diverse systems. Notably, the consortium responsible for maintaining the EU Visa Information System was awarded a comprehensive contract that encompassed supporting "the exchange of visa data across border management authorities by ensuring the processing capacity of the system and the availability of high levels of search and matching capabilities required for visa applications." (Accenture, 2012). However, as systems evolve and interfaces adapt, interoperability and backward compatibility become critical concerns, especially in transnational

data sharing and complex infrastructures.

While the Visa Information System (VIS) is now widely available, its initial iterations were rolled out to different regions over time. The increased system usage necessitated launching a project to expand the system's capabilities called "VIS Evolutions." The then newly established European Agency for the Operational Management of Large Scale IT Systems in the Area of Freedom, Security, and Justice (eu-LISA) was overseeing the development of a new VIS system through a consortium of companies. The goal was to obtain a "completely new VIS system in terms of infrastructure, software versions, and search engine" (eu-LISA, 2016, p. 8). Interestingly, the project's objectives included changing "the search engine to improve its performance." (eu-LISA, 2013, p. 8). This way, WCC was part of the consortium as a "subcontractor for a maintenance contract" (Field notes 24-07-2020) for building this upgraded IT infrastructure. The software WCC supplied would provide an improved technical component to search and match based on alphanumeric data. It is common to gloss over these aspects of the VIS's history. However, understanding these evolutions of the VIS reflects how the identification infrastructure expands and grows from its existing technological basis. Following this growth and deployments in countries worldwide can highlight how system builders grapple with technical difficulties, develop solutions, and reach compromises. These processes, in turn, can reveal key moments in the long-term developments in data matching and identification.

The VIS Evolutions project upgrade was built upon existing systems, necessitating careful consideration to ensure seamless operation for Member States already integrated into the central VIS system. Technical specifications, established initially for the VIS system's search functionalities, had to be meticulously followed to maintain compatibility with the established integrations between the central EU VIS system and the VIS systems of the Member States (often referred to as "backwards compatibility). During a meeting, I had the opportunity to discuss the integration of ELISE into the EU-VIS system with a senior developer who was actively involved in the project. They provided a description of the process, which I have reconstructed and paraphrased below (Field notes, July 24, 2020):

> A lot of customers usually already have a different system, so they want to stick to the existing matching rules of that system. Even though that may not be the optimal way to match. And that is kind of how EU-VIS works. [...] The EU-VIS system is not actually using our name matching solution. Because in that EU-VIS deal, we are only subcontractors on a maintenance deal. There was an existing VIS system of which the hardware and was updated so, basically, the pre-exiting functionalities needed to be upheld. For example, for certain data fields, they wanted to be able to manually determine to use this kind of typo correction, or that kind of phonetization.

They provided the EU-VIS project as a recurring example where the ELISE matching engine is integrated with existing systems, emphasizing the need to avoid extensive

changes and maintain existing matching rules. This approach was particularly critical in projects like the EU-VIS, where preserving pre-existing functionalities was essential. Consequently, novel and specialized features, such as advanced name matching algorithms, could not be readily introduced. They elaborated on the process, explaining that WCC had to ensure that match results aligned with the expectations set in the tests, making significant deviations from these specifications challenging.

> So, on the one hand, there were new things that were added to the EU-VIS, from functional wishes of the EU member states. But on the other hand, there was also a given test set, stating how this query should yield that results, etc. And if we diverged from those expected outputs, then we really had to defend why we wanted to deviate from it or had to deviate from it. [...] So some deviations just had really technical reasons, because ELISE couldn't do certain things in the same way. And other deviations were because we simply said that it should not be solved in the way that was specified. For example, one of those use cases was to match Moscow with Moskva using typo correction. While we said that you should not match that way, to match place name variants via edit distance metrics. [...] But so, eventually there was actually nothing from the name databases or things like that in it added to the EU-VIS.

They highlighted that certain data matching practices were being utilized due to these constraints that might not align with WCC's preferred best practices. An example cited was matching place names using solely algorithmic methods, like calculating the edit distance between two strings. They emphasized that the company would recommend utilizing lexical databases for more accurate results. However, they acknowledged the complexities of implementing such changes within the EU-VIS system, given its widespread use across all member states, making extensive alterations a significant challenge.

> The matching is configured in the EU-VIS API. So a member state can query the EU-VIS database where they can indicate in the request if they, for example, want to do a typo match or not, a fuzzy match or not, an exact match on the first name, but not by last name, etc. [...] In effect, it is the member state application that then determines how the match is performed at EU-VIS. [...] And yes, they might miss a lot of functionalities of ELISE and they could get much better results than they're probably getting right now. [...] Of course, the problem is you would actually have to deal with all the 20+ member states when the EU-VIS system would change, because they all would have to adjust their system accordingly. Even if the API might remain the same, but with slightly different results, they would have to justify such a functional change to all the other parties and get approvals to implement it.

Examining the integration of ELISE into the EU-VIS system as a gateway moment reveals a complex interplay between various systems, shedding light on aspects like path

dependencies and contingency and unveiling insights that might otherwise remain undetectable. The EU-VIS architecture allows member states to configure their systems based on their preferred matching criteria, drawing from the capabilities provided by the EU VIS central system API. While one might expect the EU-VIS to encompass all the specialized name matching expertise developed previously, as seen in the MITRE Challenge discussed earlier, the reality is more intricate. The integration process was subject to constraints such as backwards compatibility and strictly pre-defined use cases. These constraints limited the implementation and utilization of ELISE's advanced matching functionalities, as the upgrade needed to align with the legacy system of the previous VIS iteration. The VIS Evolution project is a compelling illustration of how path dependencies within the development of upgraded systems do not always facilitate the integration of advanced identity data matching features.

6.6.2 INDiGO and traveling data matching knowledge

In contrast to the EU-VIS case, the integration of ELISE into the system of the Immigration and Naturalization Service of The Netherlands tailored data matching knowledge can circulate across organizations. This was particularly evident through discussions and interviews with WCC personnel involved in the project, in which it became clear that the IND project allowed for a much more organizationally specific configuration, enabling the implementation of advanced matching functionalities tailored to the specific requirements of the IND's context. By delving into historical meeting minutes and technical documentation, I gained insights into how WCC and IND collaborated to fine-tune various search criteria, including configuring the weighting of factors like last name matching in calculating match scores. The flexibility in configuring data matching capabilities in IND systems is in contrast to the more rigid use cases and testing procedures encountered in EU-VIS because of the many member states connecting to it. Consequently, it created an avenue to integrate advanced name matching functionalities, which were not available within the constraints of the EU-VIS framework.

During fieldwork, I had the opportunity to analyze a collection of documents, among which was a noteworthy presentation in 2013 titled "ELISE: New Features." This presentation aimed to showcase the enhanced capabilities of the upcoming version of ELISE, which the Immigration and Naturalization Service (IND) could leverage for its search operations. The presentation began with an overview of recent company updates, highlighting projects like the EU-VIS and the recognition of their rank in the MITRE multicultural name challenge. Subsequent slides, aptly titled "new name matching features," delved into various new and upgraded name matching functionalities. Furthermore, the slides noted new integrated algorithms for different biometric modalities and the supported vendors.

For the new name matching features slides, features related to the "transcription and transliteration of Arabic and Asian names" are of particular interest. They encompass name variations in original and Roman scripts. One slide specified "licensing" related

to a "name matching module" and "name databases," which can understood as relating to the need for purchasing an additional license to use, for instance, the CJK name databases (described in the previous section) in ELISE. Additionally, a table comparing matching features across different ELISE versions was presented, including the ability to match based on also-known-as information provided by third parties, with an illustrative example demonstrating how "Ahmed the Tall" should match with "Sheikh Ahmed Salim Swedan" and vice versa. New name matching features, like the ones mentioned, can be attributed to, among others, WCC's active participation in the MITRE Challenge. The subsequent integration of these features into the ELISE system through system upgrades, as highlighted in the presentation, demonstrates how name matching knowledge and technologies developed in one context was adopted and circulated to an organization like the IND.

The presentation also highlighted an "EU-VIS specific feature" integrated into ELISE, referred to as "partial," with an example demonstrating that "Wij should match Wijfels." Notably, this feature specific to the EU-VIS context, was not confused with another similar ELISE feature labeled "partial fuzzy," illustrated by the example "Wij should match Wilders." This distinction underscores that there was a need for a more stringent partial matching mechanism tailored explicitly for EU-VIS. So, the distinction between these two features lies in the level of precision they apply when matching names. The "partial" feature is strict and requires an exact portion of the name to match, while "partial fuzzy" is more flexible and allows for a broader similarity in names. The presentation further included a comprehensive table comparing various matching features across different ELISE versions, providing insights into the evolving software's capabilities and its adaptability to diverse organizational requirements.

The integration of ELISE into the IND systems as a gateway moment thus demonstrates how identity matching capabilities can transcend organizational boundaries. The software updates are demonstrative of the relation between the generification of the data matching system and the development of new customer-specific data matching functionalities (compare with Pollock et al., 2016). Over time, the ELISE "core," as it is known at WCC, has accumulated a wide variety of matching features developed for different contexts. Therefore, this core also possesses the potential for configurations to accommodate domain-specific solutions, including those tailored for identity and security or public employment services. In this way, we can conceivably conceptualize ELISE as a reusable system that "transports" identification expertise across organizations through its identity- and name-matching features, as evidenced by the name matching features employed at the IND. When WCC releases a new ELISE version, customers have the option to upgrade their existing software to access new features. Therefore, this process is not entirely deterministic either, as evidenced by the EU-VIS project, where specific name matching features were deliberately omitted to address concerns related to backward compatibility.

6.6.3 Data matching and standardizing data models

In exploring our final gateway moment, we delve into the role of data matching systems as triggers for standardizing data and data models. As elucidated earlier, WCC's ELISE system often integrates with existing systems that provide the necessary data. In this process, the incoming data is mapped to conform to ELISE's data model and replicated to facilitate efficient and effective data matching. During an interview, a senior WCC staff member shared insights on bridging the gap between old and new systems, or making diverse systems interoperable. Drawing on their experiences, they note this commonly involves addressing disparities in data models' values and categories across systems:

> An example that I always give is: when personal data was stored, the hair color field contained free text. So sometimes brown, sometimes light brown, sometimes "brn," sometimes as abbreviated or light. (…) So that was free text. To go from there to a pick list, a drop-down list — what they wanted in the new system — that's quite complicated. And that is something you have to do when you connect systems with each other. Now in this case it had to because the [organization] went from a legacy to a new system. But sometimes when you talk about interoperability, the winged words of the EU as well, then you have to be able to compare these kinds of data with each other. For example, a VIS system that contains something and a nationally different system that contains exactly the same data, but the field names are different. Or the notation is just a little different. Even then, it should be possible to match those data. So, there will need to be standard models. The EU is trying to achieve this with UMF, the Universal Message Format. America has a number of standards that are also included in our product; NIEM is one of them. And in addition to having standards, you also need to be able to match smartly, and that won't be easy. That's quite complex." (Interview with WCC senior manager, May 31, 2021)

In the interview quote, the informant alludes to two approaches to dealing with the heterogeneity of data infrastructures. One approach is to establish uniformity in the values used. The informant's first example shows how the values of a free-form field for hair color in a law enforcement database were standardized to determine categorical values. Another method is to adopt common standards to reduce the various formats used by different systems and organizations. The informant's second example focuses on the interoperability framework for EU information systems in the area of justice and home affairs. The introduction of a new data standard, the Universal Message Format (UMF), that would allow for consistent data exchange is a crucial component to achieving data interoperability, as the relevant EU legislation also specifies:

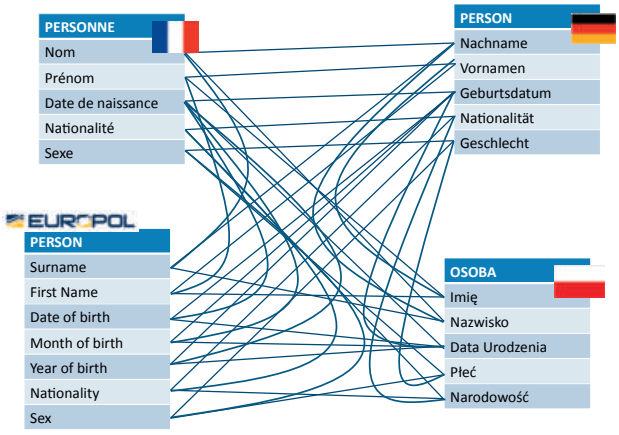
> The universal message format (UMF) should serve as a standard for structured, cross-border information exchange between information systems,

> authorities or organisations in the field of Justice and Home Affairs. The UMF should define a common vocabulary and logical structures for commonly exchanged information with the objective to facilitate interoperability by enabling the creation and reading of the contents of exchanges in a consistent and semantically equivalent manner. (European Union, 2019a, p. 22)

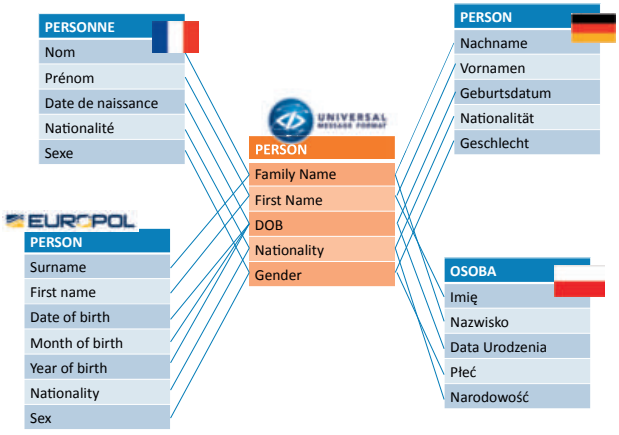
During the fieldwork, it became clear that WCC was familiar with this UMF data standard due to their previous work with the Finnish national police (WCC, 2020). The Finnish national police was one of the EU member states participating in the pilot project to automatically consult from their own national systems to the Europol watch lists via an interface called the QUEST (Querying Europol System) which provides query results in the UMF format (European Commission, 2020b; Kangas, 2019). Hence, as a supplier of their data matching system for the Finnish police for WCC was able to gain "valuable expertise in using UMF" (WCC, 2020). As a gateway moment, this illustrates how data matching in the Finnish police context included linking with Europol databases and the utilization of the UMF data model, which, in turn, demanded an additional mapping process between the UMF and ELISE object model.

UMF, as elucidated in a 2014 information sheet from the European Union Agency for Law Enforcement Cooperation (Europol, 2014), is characterized as a layer for facilitating cross-border data exchange: "It must be emphasised that UMF is not the internal structure of systems/databases (you are not required to change your national systems, legislation or processes!) but rather an XML-based data format acting as a layer between them to be used whenever structured messages cross national borders." (Europol, 2014). The UMF is conceived as a versatile multi-plug adapter connecting the concepts within different agencies' internal data models to those within the UMF's "reference model" (See also Figure 6.3). The first page of the Europol UMF brochure similarly describes the problem of law enforcement databases having similar data but in different formats using examples such as plug-and-socket standardization. As such, UMF can be used to connect systems, such as between EU member states and Europol, while keeping those systems' internal database structures intact.

Considering UMF's intended use as a common standard in the broader EU field of Justice and Home Affairs, questions arise about the mapping process between the internal data models of various agencies and this data model originally designed for law enforcement purposes. According to the brochure's definition of UMF version 1, it is "a standard or agreement on what the structure of the most important law enforcement concepts when they are exchanged across borders should be" (Europol, 2014, p.3). Subsequent updates to the model, such as the UMF version 3 project, continue to emphasize the goal of "enhancing information exchange among law enforcement authorities" (European Commission, 2020b). Mapping data from other systems to the UMF data model entails aligning with the ontologies commonly used in law enforcement. A senior WCC solutions manager, well-versed in UMF, described the general UMF data model and its



Direct mapping



Mapping with UMF

Figure 6.3: This diagram from Europol (2014) shows how concepts from national databases are mapped to UMF.

connections to law enforcement as follows:

> So POLE, persons objects locations and events. That model was used by police before computers. So, that's the model that any police forces, anywhere [in the world], use to categorize and classify crimes. So, if there is a crime, there are persons: the victim and the suspect. There are objects like weapon, like, now it's getting even different because of technology. Now, the objects are getting to be more a means of communication, any means of communication: the Internet, a mobile cellphone. Those are all objects. If there is a car, that's an object. Licence plate, vehicle, aeroplane, boat, so that's an object. [...] The location is not only the location of the event. The location can also include the addresses of the people involved in the crime. Or addresses where they used to go to. Any means of address, or regions, or even a journey. So, a route between two points. For example, there may be a crime where they used a car and they escaped from the crime scene at point A and then they hid in point B, from point A to B. And the event is the offence. What's the offensive action. That's the event. [...] So that model is what they are now trying to use in the systems, as a data model. And what the European Union did is to follow this model — but they didn't announce this anywhere. But following this model, they developed a standard for the format and the exchange of the data related to law enforcement. And this standard format is named UMF, Universal Message Format. (Interview with WCC solutions manager, 30 July, 2020)

Examining the mappings between diverse systems' data models as gateway moments reveals the emergence of new sociotechnical connections. The fundamental concept behind this process is mapping various data types from heterogeneous systems to make them suitable for searching and matching operations. However, in the case of EU member states' systems and their interaction with EU agency systems through the UMF, it is clear that these connections carry specific historical contexts and origins. As the interviewee underscores, the UMF is rooted in the well-established POLE model historically used by law enforcement for crime categorization and classification. This POLE model encompasses identifying individuals involved in criminal activities, the objects related to the crimes, locations connected to the incidents, and the criminal offences.

The contingency in this context remains unresolved, as adopting the UMF as a standard format for system interoperability in the Justice and Home Affairs field is still an ongoing process. Significantly, this adoption extends beyond law enforcement to encompass systems related to migration and asylum. The UMF will serve "to describe and label the identity, travel document and biometric data" within interoperability components, as laid down in the implementing decision (European Commission, 2023). These applications encompass specifying search queries and responses (as seen in the European Search Portal), acting as a common format for connecting identity data across JHA

databases (the Common Identity Repository) and facilitating the detection and connection of identity data between these databases (the Multiple Identity Repository). When viewed through the perspective of gateway moments, it becomes evident that a contingency exists in this alignment with ontologies originating from law enforcement. This alignment, while facilitating data matching, also carries the potential to contribute to the securitization of cross-border mobility.

The analysis of these gateway moments underscores their significance in recovering less visible facets of data matching systems. One such facet pertains to the capacity of identity matching expertise to circulate between organizations and when such circulations do not occur. The integration of the ELISE software into the IND systems, including the subsequent software updates, illustrated how specific features, notably in name matching, can traverse organizational boundaries. The incorporation of specific name matching functionalities within ELISE provides an example of the dissemination of data matching expertise across networks of organizations and entities. These include name databases, which are employed by various organizations for diverse purposes, exemplify the extensive network through which data matching knowledge is exchanged and shared. Conversely, the case of the VIS Evolutions project highlights that this transfer of knowledge does not always materialize, for example, due to factors such as backward compatibility constraints. Therefore, the multi-temporal sampling method of gateway moments enables us to identify the contingency inherent in these moments.

Methodologically, this multi-temporal sampling approach, employing gateway moments as a heuristic, has provided a lens to understand how the ELISE system's integration was contingent upon the specific circumstances and actors involved in its development. The difference between the IND and EU-VIS cases highlights that the transfer of data matching expertise is contingent upon various factors. The gateway moments of the UMF and data model mapping demonstrate how data matching does not require full data interoperability. Utilizing gateways facilitates a more open and adaptable approach to data matching across heterogeneous systems while introducing the potential for contingent standardization. This standardization is not assured, but rather dependent on specific organizational choices. In practice, organizations may merely map their data, retaining it as is in their databases, or choose to adapt it more extensively to align with the requirements of the integrated systems. As Hanseth (2001) suggests, gateways, while sometimes seen as a consequence of failed design and standardization due to their role as imperfect translators between linked systems' data models, can be practical tools in linking heterogeneous systems.

6.7 Conclusions on tracing the evolution of a data matching system: Insights into shifting landscapes of data matching and identification

This chapter has proposed a methodological approach known as "multi-temporal sampling" as an alternative to traditional longitudinal research, offering a means to examine

various contingent moments in the evolution of a data matching system. By employing this method, we were able to construct a biography of WCC's ELISE as a technological artifact. This approach aligns with the methodological strategies discussed in Chapter 3, where data matching serves not only as the object of investigation but also as a valuable methodological resource. This dual role allowed us to shed light on the intricate and contingent process behind the development of WCC's ELISE system, which continuously adapted to the evolving demands and challenges within the field of data matching, undergoing phases of openness and closure in its design. Moreover, our exploration served as a resource to address the research question of this chapter: : "*How do knowledge and technology for matching identity data circulate and traverse various organizations?*" (RQ3).

By tracing the interpretive flexibility of identity data matching software, this chapter has highlighted the work of otherwise rarely featured actors in the circulation of knowledge and technologies for matching identity data. The analysis illuminated, for example, how the formation of international professional networks, exemplified by the partnership with Accenture, spurred WCC to explore biometric data matching technology with greater depth. Additionally, the chapter highlighted the influence of a US government directive, which introduced new challenges for dealing with the heterogeneity of biometric technologies and the complexities of inter-organizational data matching. These emerging issues served as catalysts for re-opening design flexibility, prompting WCC to develop techniques for fusing biometric and biographic data matching and a plug-and-play architecture capable of accommodating diverse data formats and proprietary algorithms. Moreover, the MITRE challenge presented yet another novel problem in the form of multicultural name matching, driving international actors to engage in a competitive quest to devise innovative name matching solutions.

The examination of gateway moments has illuminated a critical aspect of the circulation of data matching knowledge and technology across organizations, one that does not unfold deterministically but is profoundly contingent upon specific contexts. The integration of ELISE into the EU-VIS system exemplified a stringent implementation approach, given the intricate web of diverse member state systems interconnected within the framework of the EU-VIS data infrastructure. This setup revealed the constraints imposed on utilizing specialized name matching technology, as its incorporation would have engendered backward compatibility issues with pre-existing member states' systems. Conversely, the integration of ELISE into the IND systems displayed a higher degree of adaptability in data matching. This flexibility was epitomized by software updates, illustrating how newly developed name matching functionalities, forged in disparate contexts like the MITRE challenge and EU-VIS, could seamlessly become part of the ELISE core, which, in turn, facilitates their capacity to circulate to other organizations such as the IND. The instance of UMF and data model mapping demonstrates how data matching among heterogeneous systems can introduce the prospect of contingent standardization. Here, organizations can decide whether to simply map their existing data, keeping it in its original form within their databases, or to opt for a more extensive

adaptation that aligns with the integrated system's specific requirements for matching identity data.

Moreover, our exploration served as a resource to address dissertation's main research question. *"How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?"*

The moments of interpretative flexibility afforded insight into the ever-evolving realm of problem definitions and design solutions for identity data matching. Originally conceived as a versatile technology with applications across diverse internet markets, ELISE underwent a significant transformation, redirecting its focus towards the identity and security sector. This strategic shift was primarily driven by shifting geopolitical dynamics and escalating security concerns, wherein data matching technology emerged as a crucial asset for law enforcement and border security agencies. This evolution ushered in newfound design flexibility, with the data matching system expanding to encompass biometric data matching and addressing specific name matching associated with identity verification in the context of security agendas. Moreover, these moments underscore the commercial nature of these technologies, not only in terms of WCC's product offerings but also in its extensive network of implementation partners, value-added resellers, and technology collaborators, including biometric vendors and external name matching databases through licensing agreements.

The gateway moments, on the other hand, highlighted opportunities and limitations surrounding data matching within transnational, commercialized security infrastructures. An illustrative case was the integration of ELISE into EU-VIS, revealing a data infrastructure where member states have greater control over defining their matching criteria through the API offered by the central system. However, from WCC's standpoint, this approach led to suboptimal data matching outcomes. Another noteworthy observation was integrating data matching into pre-existing and legacy systems, representing a potential standardization moment. New data models like the ELISE data model can serve as instances of mapping exercises between diverse systems, enabling data matching. This concept was also exemplified in the Universal Message Format, a dedicated gateway technology providing a standard for countries and agencies to align their data model concepts, thereby introducing the prospect of contingent standardization. These gateway moments thus offer valuable insights into the contingent influences on data matching within the broader landscape of infrastructural networks.

As highlighted, the interpretative flexibility framework in the Social Construction of Technology (SCOT) has some limitations, particularly its tendency to overlook structural contexts and the potential for power imbalances that can render certain actors invisible (Klein and Kleinman, 2002). Similarly, the gateway moments approach primarily emphasizes technological components and system builders, potentially privileging specific actors while making those subject to data matching technology less visible. However, this focus on specific actors presented an opportunity in this case. Paradoxically, previ-

ous research on identification technologies, which often emphasized various sampling methods, did not delve deeply into the work of less conspicuous actors within the international network that develops, finances, and profits from identification in border security and migration management.

As Hughes (1983a) observed in his analysis of the construction of large technical systems, the substantial investments made by individuals, businesses, and other system builders wield substantial influence over the trajectory of technology. This influence engenders a phenomenon he called "technological momentum," which propels the development of technology along specific pathways as a consequence of their concerted efforts. Nevertheless, the methods employed in this chapter also underscore the significance of contingent moments, yielding specific effects and aiming to illustrate that the realm of identification technology and securitization is not inherently deterministic nor linear but rather shaped by contingent choices, thereby emphasizing that alternative courses of action have been and continue to be possible.

Chapter 7

Conclusions

This concluding chapter summarizes and synthesizes the dissertation's research on matching identity data in transnational commercialized security infrastructures. Throughout the dissertation, we explored the practices and technologies of matching identity data, aiming to understand the types of knowledge and assumptions inscribed in data models, the technologies organizations used to match identity data, and the circulation of data matching knowledge and technologies across organizations. This last chapter begins by reviewing the previous empirical chapters' research questions and key findings. Next, the importance of these findings will be interpreted, and their theoretical and practical implications will be discussed. Lastly, the chapter will reflect on the research process and outline potential directions for future research.

7.1 Restatement of the main research question and summary of the research

The dissertation sought to investigate the interplay between data matching, identification systems, and data infrastructure to understand the internationalization, commercialization, securitization, and infrastructuring of identification through the materiality and performativity of data matching practices and technologies. Specifically, the dissertation addressed the following overarching research question:

> How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?

Through this research question, the dissertation sought to contribute to a more performative understanding of data matching technology's role in shaping the meaning of data, practices, and the organizations that employ it. In Chapter 1, the dissertation's goal

of examining data matching in transnational data infrastructures was broken down into several research objectives that were taken up in the subsequent chapters. Firstly, Chapter 2 mapped the theoretical landscape related to internationalization, securitization, and infrastructuring of identification, deriving the research questions and hypotheses that guided our investigation into data matching in transnational commercialized security infrastructures. Chapter 3 developed a methodological framework wherein data matching serves as both a research topic and a resource to investigate the internationalization, commercialization, securitization, and infrastructuring of identification.

Chapter 4 introduced a novel method and software tool designed to analyze the underlying data models of information systems in population management. This approach was used to investigate authorities' imaginaries about people on the move through the connections between different data models. Chapter 5 examined the relationship between identity data matching technologies and routine identification practices, demonstrating the practical implications and challenges organizations encounter when matching identity data to repeatedly identify individuals within and across organizations. Lastly, Chapter 6 investigated the circulation of a commercial data matching system throughout its lifecycle. Specifically, the chapter investigated contingent moments in the problematizations and design of a data matching system — such as shifts in the meanings attributed to the software from a generic data matching engine to a security tool or its use as a bridge to connect disparate identification systems — influence the practices and technologies employed in transnational security infrastructures.

7.2 Overview of the findings

This section summarizes the key findings of the empirical investigations conducted in the dissertation. The findings are linked to the research objectives, questions, and the use of the methodological framework's infrastructural inversion strategy.

7.2.1 Chapter 4: The Ontology Explorer: A method to make visible data infrastructures for population management

Chapter 4 contributes to the research objective of introducing a new method and software tool to analyze the schemas underpinning information systems in population management. This method and tool, dubbed the "Ontology Explorer" (OE), help achieve this objective by enabling a systematic examination and comparison of non-homogeneous data models used in diverse information systems. The chapter utilizes the OE method, which draws inspiration from schema matching techniques and STS research on classifications and standards, to bring attention to the assumptions embedded in the data models used in information systems for population management.

The infrastructural inversion strategy employed in Chapter 4 compared data mod-

els to uncover authorities' imaginaries and expectations about people on the move embedded within the technical standards of data infrastructures. Chapter 4 thus addresses research question 1 (*Which types of knowledge and assumptions about people-on-the-move are inscribed in data models of national and transnational security infrastructures? What implications does this have for how organizations can search and match identity data?*), through an illustrative analysis of the types of knowledge and assumptions about people-on-the-move inscribed in data models of national and transnational authorities. The illustrative analysis demonstrated how the OE can be used to make visible the implicit knowledge and assumptions in data infrastructures for population management.

Chapter 4 aids in answering the main research question by revealing implicit assumptions and patterns within information systems design. The findings indicate different conceptualizations of identity data among authorities involved in population management, which could have implications for matching data across various database systems. For example, the Eurodac system primarily collects fingerprint data and does not include biographical details such as name, date of birth, and nationality. Therefore, the only means of matching data with Eurodac are biometric matching methods or linking data using the Eurodac number. Furthermore, Chapter 4 contributes to Processing Citizenship's WP1 by introducing the OE, which facilitates the computational analysis of diverse authorities' assumptions about individuals while serving as a foundational component for critical analyses when integrated with ethnographic observations of the practical utilization of information systems.

7.2.2 Chapter 5: From registration to re-identification: Exploring the interplay of data matching software in routine identification practices

Chapter 5 contributes to the dissertation's research objective of examining the relationships between identity data matching technologies and routine identification practices. It addresses Research Question 2 (*How do organizations that collect information about people-on-the-move search and match for identity data in their systems? How is data about people-on-the-move matched and linked across different agencies and organizations?*) through an empirical investigation, centering on the iterative processes of identifying applicants within the Dutch government migration and asylum agency and the role of a data matching system in these processes. Re-identification is introduced to conceptualize the multifaceted iterative identification procedures, including retrieving corresponding identity data from databases and determining whether multiple database records, potentially originating from different organizations, pertain to the same real-world individual.

Utilizing the second infrastructural inversion strategy, this chapter investigates the practices surrounding identity matching and linking within the agency and across other organizations. In doing so, the findings reveal a considerable diversity in re-identification practices within the IND, which manifests in two primary dimensions. Firstly, these practices vary regarding the information accessible to staff during the process. Secondly, they exhibit differences in the precision criteria required for suc-

cessful re-identification. Building upon these findings, the chapter introduces an interpretative framework that categorizes re-identification practices based on the specific requirements for interpreting search inputs and results. This framework yields a matrix, acknowledging re-identification as a multifaceted process rather than a singular activity. It encompasses a spectrum of iterative practices, spanning diverse scenarios such as direct applicant interactions, telephone conversations managed by staff, handling application forms received through postal services, and automated re-identification procedures.

The chapter's findings emphasize the various practices of re-identification that can be impeded by data friction, potentially resulting in failed re-identification. The analysis identified three prominent forms of data friction that may hinder applicant re-identification: friction between standardized identification and the differences in institutional practices, friction from variations in the precision and accuracy of identity data during its transformation across different mediums and use in formulating search queries, and friction arising from the opaque calculation of match results and the need for thorough interpretation and fine-tuning of search results. These forms of friction, in turn, prompted a closer examination of the costs arising from failed re-identification, as exemplified by the existence of duplicate records and the labor-intensive process of deduplication.

Chapter 5 addresses the main research question by examining the interaction between the re-identification practices of the IND and a commercially developed data matching system. The IND's use of the commercial data matching system influenced the agency's re-identification strategies due to the system's embedded expertise in matching identity data. Additionally, the chapter draws attention to the link between the deduplication process and transnational systems. The IND and partner organizations use data from major European Union information systems, connecting seemingly unrelated records within their databases. These findings underscore re-identification processes and associated technologies are not isolated; they are intricately intertwined with broader commercialized security infrastructures.

7.2.3 Chapter 6: Uncovering the long-term development of identification infrastructures: A multi-temporal perspective

Chapter 6 delved into the evolution of a data matching system and the construction of transnational security infrastructures. The chapter's findings contribute to the research objective of exploring the evolution of identification systems, identifying contingent moments, and examining how data matching expertise circulates among various organizations. The evolution of a commercial data matching system is thus used to investigate more extensive long-term developments of identification systems and transnational commercialized security infrastructures.

The chapter uses the third infrastructural inversion strategy of sociotechnical change to trace the evolution of data matching technology and design. The strategy is opera-

tionalized using "multi-temporal sampling," which the chapter uses as an alternative to conventional longitudinal studies for exploring the contingent processes shaped by and shaping identification systems and infrastructures. The chapter proposes two heuristic devices to identify contingent moments: the changing interpretive flexibility of the data matching system and the gateway moments when identification systems intersect with broader infrastructures. The first heuristic suggests that changes in the interpretive flexibility of the data matching system, as per SCOT, are analytically valuable moments that can provide insight into how the data matching system's design was contingent upon specific social groups' problematizations of data matching. The second heuristic proposes examining gateway moments, as per infrastructure studies, which involve making sociotechnical compromises when integrating separate systems into more extensive infrastructures. These two heuristics were applied to analyze the contingencies in the evolution of a commercial data matching system, thereby providing insight into the internationalization, commercialization, securitization, and infrastructuring of identification.

In this way, the chapter addresses Research Question 3 (*How do knowledge and technology for matching identity data circulate and travel across organizations?*) by exploring the networks of people and organizations involved in developing and disseminating data matching technologies. Concerning the main research question, the chapter's analysis offers insights into how globally honed technologies are adjusted to new contexts by uncovering the compromises and adaptations required when building identification infrastructures. Furthermore, the activities of software vendors, often overlooked actors, are revealed as they distribute and reuse data matching systems, thereby influencing the long-term development of identification practices and infrastructures.

Significant findings from Chapter 6 arose by looking at the moments of interpretative flexibility and analyzing moments when the system demonstrated gateway-like characteristics. Moments of interpretative flexibility made it possible to retrace how a private company became enrolled in security logic through shifts in identification problems and solutions. The software was initially created as a versatile tool for data matching across industries and was only later tailored to meet specific needs like border security and migration management. At different points, the interpretative flexibility of data matching was closed down and reintroduced with shifts in data matching and identification problematizations. The design of the technology needed to address specific requirements and challenges of identification in border security and migration management, as exemplified by the development of matching functions incorporating multicultural name matching and biometrics in response to new problematization of identifying known and suspected security threats. Additionally, focusing on these contingent moments made apparent the role of diverse social groups, such as the development of international professional networks to secure contracts for developing identification systems.

Gateway moments offered a complementary approach to identifying contingent moments in the long-term evolution of the data matching software by considering the intricate technical and logistical challenges of integrating different identification systems,

which are not visible by only looking at shifts in interpretative flexibility. As such, the chapter provided insight into the contingent circulation of identity matching expertise among organizations. The integration of the data matching system into IND systems showed the capacity to disseminate specific name matching functionalities across organizational boundaries. In contrast, such knowledge transfer does not invariably transpire, as seen in the use of the data matching system in the VIS Evolutions project, where factors like backward compatibility constraints limited the capacity for data matching knowledge to circulate.

Methodologically, the multi-temporal sampling approach employing gateway moments as a heuristic unveiled the contingency of the data matching system's integration into more extensive networks, revealing that the transfer of data matching knowledge relies on various elements. Gateway moments like the UMF and data model mapping elucidated that full data interoperability is not always a prerequisite for data matching. Instead, gateways can offer adaptability and openness, potentially leading to contingent standardization. However, this standardization hinges on organizations' specific choices regarding data mapping and adaptation to integrated systems.

Together, the moments of interpretative flexibility and gateway moments helped answer the main research question by demonstrating how data matching knowledge and expertise can but does not invariably circulate between different actors and organizations. These findings indicate how technologies and methods for matching identity data can spread and influence identification practices across various domains and settings. Furthermore, the chapter addresses Processing Citizenship's WP4 by illustrating data matching's role in facilitating semantic interoperability between EU and MS systems. Furthermore, the analysis of the data matching system in EU-VIS highlights the specific relations between EU, MS and commercial actors who needed to configure it while balancing new features and adhering to backward compatibility needs from EU member states' systems.

7.3 Discussion of theoretical and practical implications

This section discusses the theoretical implications of the research findings and how each chapter contributed to existing theoretical frameworks and concepts in its respective area of focus.

7.3.1 Chapter 4

In Chapter 4, the introduction of the Ontology Explorer (OE) method and tool expands on methodological innovations in infrastructure studies and digital methods to analyze data models underpinning information systems in population management. The chapter highlights the potential for innovative methodological contributions to make data infrastructures for population management visible by examining the semantic level of in-

formation systems' data models. Methodologically, the chapter uses correspondences and discrepancies between data models to reveal the interconnections between technical details and the politics of knowledge production. The chapter's illustrative analysis enhances our understanding of the data models used in migration and border control information systems and their role in revealing how authorities envision managing populations.

The chapter's practical and theoretical implications of the research findings are twofold. Firstly, the OE method introduces a means to account for resistance within information systems. Using the OE to analyze "thin" data models, researchers can identify differences and absences between systems, which can serve as the basis for critical analyses. This method allows for the detection of discrepancies between what is represented in the data models and the actual use of the systems, providing valuable insights into power dynamics and potential areas of intervention. Furthermore, analysis via the OE can be complemented by ethnographic observations to account for practices of resistance exerted by actual people.

Secondly, the OE method contributes to ongoing discussions on the politics of data and data-driven governance. While, for example, digital sociology often focuses on user-generated content, interventions into data need to examine existing structures, such as the underlying data models. The OE has the potential to present and analyze data models, which could allow for experimental forms of participation and speculation on alternate ways to represent data about populations. By engaging with the technicalities of information infrastructures rather than solely relying on textual content, the OE facilitates critical engagements with the collected categories of data. This could open up opportunities for actors typically treated as data subjects to become active participants in rethinking and reshaping data practices, leading to more inclusive and transformative approaches to data governance.

Overall, Chapter 4 highlights the potential of the OE method to uncover resistance and power dynamics within information systems and to foster more participatory and reflexive approaches to data governance. By focusing on the technical aspects of data models, the method offers a distinctive perspective that complements traditional qualitative approaches, providing researchers and practitioners with valuable tools for critical analysis and intervention in the politics of data.

7.3.2 Chapter 5

Chapter 5 extends existing discussions on identification by emphasizing the broader scope of what the chapter refers to as re-identification. While previous research has often focused on initial registration and biometric data, this chapter highlights the iterative processes of retrieving and matching identity data across time and space. The chapter thus addresses an area not covered in existing literature — the use of data matching technologies in routine bureaucratic re-identification. These chapter's findings expand our theoretical understanding of identification beyond its initial stages and

underscore the importance of considering the influence of data matching technologies in identifying people throughout bureaucratic procedures.

Chapter 5 presents several practical implications that emerge from the study of re-identification practices and the use of data matching at the IND. The chapter explores the practical utility of data matching technologies, which serve as tools to mitigate errors and ensure the precision of re-identification processes. On the one hand, organizations' utilization of these tools can reduce re-identification errors and enhance the quality of identification data. Conversely, the findings on using data matching in re-identification underscore the potential consequences of errors in data linkage or incorrect data entry into databases, such as duplicate identity data. Such errors can adversely affect individuals and organizations, necessitating additional efforts to verify and rectify inaccuracies.

Additionally, the integration of various systems can introduce friction and ambiguity into the search process. For example, when users must query multiple systems equipped with distinct data matching engines, discrepancies can emerge between search behaviors and query formulations. To mitigate these issues, it would be advantageous for tools to consider the search behaviors of other systems. This consideration could lead to aligning search processes or implementing interfaces capable of translating and optimizing queries for multiple systems simultaneously. By facilitating this integration and alignment, the search process can be streamlined, reducing user confusion and ultimately enhancing the efficiency of re-identification efforts.

Furthermore, the design of the data matching tools and their use within identification systems notably influences re-identification. In the context of the IND, the data matching tools do not account for the contextual nuances of a search, such as the user's role or the stage of the identification process. However, the research findings have underscored that users often possess distinct needs and preferences concerning search functionalities. For instance, certain users may necessitate more comprehensive information explaining why a specific match result was included in the search results. To accommodate these varied user requirements, developing novel application features to assist users in crafting queries and comprehending search outcomes could be advantageous. Moreover, fostering a more coherent integration between the backend search capabilities and the frontend user interface holds promise in enhancing both the user-friendliness and the overall capabilities of these search tools.

7.3.3 Chapter 6

In Chapter 6, the theoretical implications are twofold. Firstly, the chapter offers an alternative to conventional longitudinal studies for exploring the complex processes shaping and shaped by identification systems and infrastructures. By proposing two multi-temporal sampling heuristics — moments of interpretative flexibility and gateway moments —, the chapter demonstrates the use of considering the contingency in developing sociotechnologies of identification. This contribution adds to the ongoing theoretical discourse, as exemplified by the Biography of Artefacts and Practices Approach (Pollock

and Williams, 2009), regarding methods used to comprehend the development and use of software systems. Adopting a multi-temporal sampling approach in the field of IT in border and migration value broadens the analytical focus and opens up new avenues for inquiry. Adopting this approach makes it possible to discern the contingency inherent in the sociotechnical processes involved in the development and evolution of identification technologies, rather than simply viewing identification systems as a final product.

Secondly, the chapter further advances our comprehension of an often overlooked facet in the existing literature by delving into a commercial software vendor's development of data matching technology. Employing heuristics to identify moments of contingency in the evolution of data matching software, the analysis effectively retraces the sociotechnical transformations of the software. It underscores instances where the internationalization, commercialization, securitization, and the infrastructuring of identification become apparent. Consequently, this chapter's insights contribute to the body of knowledge surrounding the involvement of non-state and commercial entities in the datafication of migration and border control.

Chapter 6 identifies practical implications stemming from using the interpretative flexibility heuristic. This heuristic underscores the inherently dynamic nature of identification technologies and practices, challenging the notion that these technologies possess fixed or stable meanings. Instead, it accentuates how various social groups actively engage in the interpretation and contestation of these technologies. This social agency, in turn, significantly influences the trajectory and outcomes of identification technologies and practices. Recognizing the concept of interpretative flexibility invites rethinking the design of identification systems in a way that is more responsive to the diverse and evolving needs of the individuals and communities they affect.

The concept of gateway moments provides practical insights into the expansion of identification systems into more extensive infrastructures. It underscores the critical role of data matching technologies as connectors, bridging gaps and enabling the seamless integration of heterogeneous identification systems and practices. Although sometimes underestimated, these gateway-like technologies serve as crucial facilitators in establishing and maintaining networks within larger-scale infrastructures. Recognizing the significance of these gateways can inform the design and implementation of identification systems. Acknowledging the role of these gateways can pave the way for more robust and effective identification infrastructures, benefiting the individuals and organizations that rely on them.

The following section will reflect on the research process and discuss limitations encountered during the study. It will provide an opportunity to evaluate the methodology, data collection, and analysis techniques, discussing their strengths and weaknesses. Additionally, the section will address constraints or challenges faced throughout the research, such as data access limitations or the generalizability of findings. By reflecting on the research process and its inherent limitations, the dissertation aims to produce recommendations for future studies on identification technologies and transnational com-

mercialized security infrastructures.

7.4 Overview of the research process

The research process was guided by a methodological framework that leveraged data matching as both a topic of investigation and a methodological resource. Drawing inspiration from Bowker and Star's notion of "infrastructural inversions," three distinct methodological strategies were employed: comparing data models, analyzing data practices, and tracing sociotechnical change. These strategies were integrated into a methodological framework, allowing for a comprehensive examination of the practices and technologies involved in matching identity data as a research topic. Furthermore, by reversing the tendency of data matching to recede into the background, this methodological framework made it possible to use data matching as a resource to investigate the internationalization, commercialization, securitization, and infrastructuring of identification.

The first inversion strategy employed was the comparison of data models, which provided insights into the information collected by various authorities' information systems for population management. This work included collecting data models from EU and EU Member State authorities through desk research. Other researchers' contributions from the Processing Citizenship project also provided indirect traces of data models, such as screenshots of graphical user interfaces that provided evidence of categories of data collected in systems. Utilizing indirect traces of data models emerged as a strategic approach to tackle the challenges of obtaining data models for systems operating in the sensitive realm of migration and border control. In this context, where technical specifications related to data models are typically not publicly accessible, the indirect traces served as a valuable alternative. However, this approach introduced an additional layer of complexity. It necessitated the development of methods to effectively compare various documents about data models, a task made more intricate by the diverse file formats and languages in which these documents were presented.

Therefore, the Ontology Explorer method dealt directly with those challenges. It successfully enabled us to compare diverse data models by employing additional steps to code, harmonize, and group all documents, categories, and values. The novelty came from integrating insights from existing classification and infrastructure studies and discourse analysis into a systematic method for analyzing data models. Additionally, drawing from an analysis of data models using the OE, Pelizza and Van Rossem (2023) built further on this work. This 2023 article described the different "scripts of alterity" of security subjects, highlighting how data infrastructures shape power dynamics among individuals and states. Furthermore, the OE tool used for this research has been released as open-source software (Van Rossem, 2021), enabling the analysis of information systems' data models in various domains and contexts beyond this research's scope.

The second infrastructural inversion strategy delved into the analysis of data prac-

tices, aiming to unveil the intricacies of routine activities involved in searching for and matching identity data within diverse organizational contexts. To operationalize this strategy data, fieldwork and interviews were conducted at a software supplier specializing in commercial data matching tools and at one of the company's customers, the Netherlands' Immigration and Naturalization Service (IND). The fieldwork involved a series of meetings and interviews with the company's personnel, which took place online and on-site at their headquarters in Utrecht, The Netherlands. Additionally, documentation was provided to me to facilitate an in-depth exploration of the technical intricacies underpinning the design of the data matching software. This approach enabled a thorough exploration of the data practices embedded within these organizations and their implications in the context of identity matching.

The initial project proposal was initially conceived to undertake a comprehensive analysis and comparison of the utilization of the company's data matching software at both the IND and the EU Visa Information System (EU-VIS). This approach was rooted in the anticipation of encountering distinct challenges and dynamics due to the differing scales of use at these two entities. However, as the research progressed, it became evident that adaptations to the original research plan were necessary. Access to the IND proved to be more readily available, while the complexities surrounding confidentiality and access hindered extensive research into the EU-VIS project. Consequently, the research focus was redirected towards a more in-depth exploration of the software's practical application within the IND. This shift in approach was embraced by all parties involved, as it offered mutual benefits. The software company gained valuable insights into the real-world utilization of their product. At the same time, the IND agency had the opportunity to gain a deeper understanding of the intricacies and challenges associated with the search and match functionalities.

I conducted semi-structured interviews with the IND agency staff members to understand their day-to-day activities concerning data searching and identity matching in identifying applicants throughout the agency's complex bureaucratic procedures. Detailed descriptions of the interview process and the subsequent data analysis can be found in Chapter 3 and Chapter 5 of this dissertation. Following the interview phase, I compiled a report summarizing my findings, that was then shared with the software company and the IND agency. The responses I received from both parties were notably positive, with an invitation extended for me to present the report at a company-wide online meeting dedicated to disseminating knowledge and insights. It is worth highlighting that the research offered the company valuable insights into how its software is actually being used. The explanation for this aspect lies in the company's function as a software supplier, where their products are used by integrator partners to deploy across different organizations, resulting in a lack of visibility into end-users' interactions with their software solutions.

The third infrastructural inversion strategy focused on mapping the dynamics of sociotechnical change, exploring the dissemination of knowledge, technologies, and prac-

tices associated with data matching across organizations and over extended periods. This approach arose as an adaptation to the constraints encountered during the research, which initially aimed to analyze and compare the use of data matching technology in the IND and EU-VIS systems. The modified research plan instead delved into the historical progression of the software, unveiling the sociotechnical networks that played a role in the transnational circulation of identification technologies. The research explored the growth of commercial solutions within the identification technology domain and illuminated the interplay between data matching and shifting problematizations of identification. To illustrate, I delved into the company's involvement in an international competition sponsored by a prominent US research organization. This competition sought innovative name matching solutions, and by retracing the various parties involved, I unearthed connections between technological advancements, market dynamics, and the imperative need for security measures. This exploration facilitated an understanding of how data matching solutions have continuously evolved to address changing problems.

This research phase was also grounded in fieldwork and interviews conducted at the same software supplier, a creator of commercial data matching tools. Here, the central objective was to delve into the transformations within data matching and transnational data infrastructures, seeking to unveil contingencies in these systems' historical development and evolution. The interviewees were generally keen to share their experiences, expressing interest in contemplating broader issues beyond their immediate organizational roles. Nevertheless, some aspects of the early evolutions of the company remained obscure, often due to interviewees' lack of direct involvement during that period. To address this gap, I complemented the interviews with information from archival sources, such as dated web pages, news articles, and press releases. This multifaceted approach allowed for a comprehensive understanding of the subject matter by cross-referencing and triangulating the information, revealing valuable insights into the transformations, adaptations, and challenges.

This section has provided an overview of the research process employed in this study. Using the methodological framework inspired by "infrastructural inversions," the research used data matching as both a research topic and a resource. Three methodological strategies were used, including comparing data models, analyzing data practices, and tracing sociotechnical change. These strategies enabled the examination of the interplay between identification systems and the internationalization, commercialization, and securitization of identification. Through data collection methods such as desk research, fieldwork, and interviews, the research developed new perspectives on the information collected by organizations, the everyday data practices of matching identity data and re-identification, and the dynamics of transnational commercialized security infrastructures. The next section will further reflect on the findings, including unexpected discoveries and the limitations and challenges faced during the research.

7.5 Reflection on findings and limitations

While conducting research, I discovered some surprising findings indicating how a company's integrated software affects how customers use it for data matching and identity management. Specifically, these findings challenged my assumptions about the data matching tools, as the software's integration in systems such as the Netherlands' migration and asylum was relatively invisible to users. Also, I was surprised to discover that the software was adaptable and effective across different domains, which challenges traditional notions about the difficulty of adapting generic software to local circumstances. However, throughout the research, several limitations also arose that had a direct impact on the findings I was able to produce. In this section, I will examine the unexpected results, discuss the identified limitations, and reflect on the research outcomes.

7.5.1 Surprising and unexpected findings

Several surprising and unexpected findings emerged from Chapter 4 that aided our analysis of data categories within systems and their implications. The investigation found that different systems had significant differences in the categories of data they stored. For example, the specificity of information about individuals' names varied greatly. These differences were crucial for the analysis as they revealed authorities' conceptions and imaginaries about individuals. Categories such as alias names and violence indicators showcased how these systems could indicate suspicion or criminality. Furthermore, the analysis showed that only a limited number of shared data categories, such as birthdate, nationality, and biometric data, were consistently used among the multiple systems. These findings showed how identity data categories, seemingly technical details, help examine authorities' imaginaries about people on the move.

Chapter 5 presented unexpected insights regarding the integration of WCC company's data matching software into the Immigration and Naturalization Service (IND) systems. Notably, the software's design, which positioned the data matching software as a relatively independent component within the IND systems, created distinctive challenges. While this technical design was efficient in some respects, it did not always align with the specific needs of all users. One significant consequence of this design was the relative invisibility of the search process to most users, resulting in a lack of awareness regarding the underlying data matching mechanisms. The chapter's exploration of diverse data matching scenarios within the IND underscores the potential benefits of tighter integration between the software, the local context, and the system's graphical user interface. For instance, during interviews, it became evident that users often required clarification when search results included matches based on historical data, such as matching previous names before marriage. Contrary to expectations before the research, this design resulted in an invisible layer of data matching. However, this also led to a lack of clear understanding of search results and functionality for users.

Chapter 6 revealed unexpected findings concerning the role of commercial software

companies within the domain of identification. The research investigated the dynamics between the software company as a technology supplier and its associations with technology integrators, unearthing intricate professional networks focused on identification solutions. At the beginning of the research, it was unclear how the company got involved in data matching for identity and security. However, it was surprising that these relationships significantly influenced the company's positioning in the market and paved the way for it to explore new realms of identity and security. An additional surprising finding pertained to the circulation of data matching knowledge; contrary to initial expectations, the adaptable design of the data matching software allowed for customization to meet customers' specific requirements. This means that the expertise in data matching within the software was not always utilized and circulated between organizations. Overall, Chapter 6's examination challenged the conventional view of the company's data matching software as a fixed solution for identity matching in migration, border control, and security, unveiling the diverse contingencies shaping the software's evolution.

7.5.2 Limitations

During this research, several limitations were encountered that are essential to understand the scope and implications of the study. These limitations impacted the breadth and depth of the investigation, making it necessary to consider them when interpreting the research findings.

As part of the Processing Citizenship project, the initial plan, in collaboration with the Principal Investigator, was to set up my fieldwork within the eu-LISA EU agency to investigate identity data and data quality in the context of border security and migration management. However, this endeavor proved challenging due to significant access restrictions and the customary barriers encountered in qualitative secrecy research. Regrettably, we could not secure access to commence the proposed project at the designated fieldwork location, resulting in significant time and effort spent in cultivating contacts and formulating a comprehensive project proposal for a potential traineeship with the agency. Ultimately, the primary stumbling block pertained to disagreements over the ownership and utilization of the research data that would be generated. Consequently, we decided to establish new fieldwork research by collaborating with a data matching technology supplier for the agency.[1]

The company swiftly accepted the project proposal in Dec 2019–Jan 2020, but the COVID-19 pandemic introduced unexpected challenges to the research project with the

[1] As the researcher conducting this study, I acknowledge that my background played a meaningful role in assisting me in accessing the fieldwork. Being a white male with an academic background in computer science and experience in software development, I was able to speak the same language as my interlocutors, which facilitated communication and ensured a better understanding of their perspectives. Moreover, my knowledge of Dutch and being based in the Netherlands helped to establish trustworthiness with the participants. It is worth mentioning that the research was conducted under the EU-funded project Processing Citizenship, which added to its credibility and relevance.

data matching software company. Initially, the research had to be postponed due to COVID-19-related safety measures and travel restrictions in the Netherlands. When restrictions were eventually eased, my on-site visits to the company were impacted by fewer employees on-site, as limitations were imposed on the number of individuals allowed to work in the office, with many staff members continuing to work remotely. Subsequently, with the reintroduction of restrictions, I could not visit the company on-site. Consequently, the research had to adapt to an online format, reducing opportunities for in-person visits and direct observations during fieldwork. This absence of face-to-face interactions and observations may have hindered my grasp of the intricacies and ramifications associated with the usage of the software

The study had limitations due to the decision to concentrate on the EU Visa Information System and the Netherlands' migration and asylum agency systems rather than investigating the company's developments in advanced passenger information systems. This decision aligned with the broader goals of the Processing Citizenship project, but it restricted the inquiry to specific projects in the company's portfolio; some other relevant customers would also have necessitated security clearances. Additionally, obtaining relevant documentation about the use of data matching software in the EU-VIS system proved to be challenging due to the company's limited access to EU documentation and confidentiality practices by the technology integrator, which hindered my ability to delve deeply into the intricacies of the software implementation and its influence on identity data management within the EU context. Indeed, the research in Chapter 5 focused on the implementation within The Netherlands' immigration and asylum agency, with access to this specific context's technical details. In contrast, Chapter 6 shifted its focus towards exploring connections with EU-VIS and the broader transnational security infrastructures by tracing the software's sociotechnical evolution, drawing from different accessible data sources.

It is also important to note that the exclusion of migrants' experiences from this research was a result of the overall design of the Processing Citizenship project. Other researchers within the project focused more specifically on migrant experiences. For example, elements of this dissertation's Chapter 4, which delves into information systems' categorization of data, can be complemented by the work of Lorenzo Olivieri (2023). His research involved examining the processes of translating migrants' identities and personal stories into standardized categories and traits, utilizing innovative techniques that included working closely with migrants. Additionally, Pelizza and Van Rossem (2023) examine how migrants are inscribed as particular kind of users, referred to as "scripts of alterity." This article utilizes field observations from the project to analyze forms of resistance against these scripts. The Processing Citizenship project's overarching framework ensured the incorporation of various perspectives, including those of migrants, into the project's broader scope.

Despite the encountered limitations, this study undoubtedly contributes to our comprehension of how the integration of commercial identification software shapes

and is shaped by transnational security infrastructures. Recognizing these limitations also serves to highlight the complexities inherent in researching technology within the realm of security, including issues related to access and secrecy (de Goede et al., 2020). Moreover, it emphasizes the significance of being attuned to the contingencies and unforeseen developments that can unfold in research. Significantly, the shift towards collaborating with a less widely known data matching software supplier unveiled previously lesser-known dimensions of identification and identity management. The subsequent section will explore the ethical considerations that arose throughout this collaboration and research process.

7.6 Ethical considerations

Collaborating with a company in identity data and security inevitably prompted ethical considerations that revolved around the potential repercussions of their technology on individuals. These concerns extended to pondering how the findings and insights derived from the research might influence the company's ongoing development of its software solutions. The research's dual role, offering an in-depth analysis of the software's functionalities while simultaneously assisting the company in refining its technology, necessitated reflecting on the parties involved, including the company, the individuals impacted by the technology, and the broader societal implications of data matching in the securitization of migration and border control.

On the one hand, during my interviews with Immigration and Naturalization Service (IND) personnel, I observed the potential advantages of data matching in streamlining bureaucratic processes, benefiting both employees and applicants. The implementation of data matching technology assisted in the execution of daily tasks, offering support in various aspects of their work. However, my observations also revealed that particular system design choices could unintentionally impede efficient operations. For instance, I observed that there was no automatic check to determine whether an applicant existed within the agency's database. This seemingly minor technical oversight sometimes led to the creation of duplicate data entries, subsequently necessitating additional efforts to resolve these duplicates at a later stage. Nevertheless, rectifying this issue is not straightforward, as it requires a clear and reliable definition of what constitutes a duplicate record. Additionally, duplicate entries could raise suspicions of fraudulent activities on the part of applicants. These issues can often stem from previous data entry errors rather than malicious intent, underscoring the complexities of data matching in sensitive contexts like migration and identity management.

On the other hand, ethical considerations related to the development of data matching software arose in Chapter 6. One area of concern is the use of name matching functionalities within the software, which involves the utilization of name databases. These databases contain a wide range of names for identification purposes, including variations of Arabic and Asian names. The development of name matching technology, influ-

enced by linguistic and transliteration differences, can potentially flag names from these groups as potential matches disproportionately. This overrepresentation could lead to the perpetuation of stereotypes and marginalization of specific individuals. Therefore, evaluating the potential consequences and unintended effects of such technological solutions is essential. Unfortunately, such systems often operate as opaque "black boxes" that are not transparent to the public, which is a significant concern, especially when proprietary technologies of commercial entities are involved. Therefore, the "multi-temporal sampling" method in Chapter 6 also acts as a way to examine the characteristics of such systems by studying specific moments in their development.

7.7 Future research directions

There are several areas that require further research to fill the knowledge gaps, anticipate emerging technological developments, advance methodologies, encourage interdisciplinary collaborations, and explore practical applications of data matching and other identification technologies in transnational commercialized security infrastructures.

Chapter 4 emphasized the potential benefits of alternative methods and tools for analyzing identification systems, which highlight the need for further research in this area. To advance research on identification technology, it is recommended that future studies continue incorporating and analyzing the technical intricacies of identification systems. The methods used in Chapter 4 combined both qualitative and quantitative approaches to analyze one such aspect: data models. The Ontology Explorer (OE) method and tool developed and applied in this dissertation required interdisciplinary collaboration across various fields, including computer science, sociology, law, ethics, and policy. Future research should actively promote collaborations that engage researchers from diverse domains, as this can provide multifaceted perspectives and yield novel insights. The OE method and tool were instrumental in analyzing the similarities and differences among various authorities' data models. Future research could analyze the mechanisms that hinder or facilitate data interoperability, thereby enhancing our understanding of how data models affect identification in transnational infrastructures. Lastly, the OE offers flexibility that extends beyond the scope of analyzing data models in population management and security, enabling its application in other contexts as well.

Chapter 5 examined the interplay between data matching technologies and re-identification processes, revealing both the potential benefits and challenges of these technologies. Future research can delve deeper into the evolving technologies used in data matching practices, particularly considering the influence of emerging technologies like generative artificial intelligence and biometrics on identification. For instance, AI-driven data matching systems can present unique challenges for individuals who may be falsely identified, as explaining and comprehending these matches can introduce further complexity, potentially hindering people's potential for disputing false positive results. It is imperative to scrutinize these technologies' potential benefits and

associated risks, all while considering ethical and policy implications. Establishing robust governance mechanisms will be essential to safeguard individual rights and ensure accountability and transparency, such as how matching results are calculated and presented.

Chapter 6 provides a detailed analysis of the evolution of a data matching system, highlighting its dependence on various factors and the significant role played by different actors in shaping its development. Future research should continue exploring identification systems' evolution and practical applications. Such an approach can offer valuable insights into changes in the broader domain of identification, such as the role of commercial actors and shifts in securitization. Additionally, combining academic theories with practical experiences can show how specific identification technologies function across various domains, including healthcare, finance, and public administration. Collaborative efforts across multiple disciplines are essential, as experts from diverse fields can significantly contribute to our understanding of identification systems.

7.8 Final reflections and concluding remarks

This dissertation embarked on a journey driven by an initial curiosity about the interplay between data matching and linking processes and the blind spots faced by authorities. It aimed to decipher how authorities navigate the complexities of identifying individuals even when faced with incomplete data, aliases, and false identities. By mapping the theoretical landscape to untangling the intricacies of matching identity data in the domain of transnational commercialized security infrastructures, it laid the foundation for the research questions. Furthermore, it introduced a methodological framework to not only examine data matching but also harness data matching as a research resource, opening new avenues for investigating the challenges posed by identification processes across national borders, notably in the context of border security and migration control.

First, introducing a novel method and tool in this study facilitated a comprehensive exploration of the differences and commonalities embedded in national and transnational security infrastructures' data models, shedding light on the varying knowledge and assumptions about individuals on the move. Within the categories of data found in these models, one can discern authorities' imaginaries, revealing how they conceptualize and enact individuals in distinct ways. This conceptualization also has relevance for the realm of data matching, as the integration and interoperability of data hinge on the connections between these data models, playing a pivotal role in matching individuals' data across diverse data sources.

Secondly, the focus on data matching made it possible to shift attention beyond the initial registration and identification phases to re-identification practices across space and time. Re-identification was introduced as a concept that entails the continuous utilization and interconnection of data from diverse sources to determine if multiple sets of identity data correspond to a singular real-world individual. However, examining the it-

erative processes of re-identifying applicants throughout different stages of bureaucratic procedures highlighted that, while integrating data matching tools for re-identification can mitigate friction in re-identification, it can also introduce certain associated costs.

Thirdly, tracing the sociotechnical changes of a data matching system allowed us to detect the circulation of knowledge, technologies, and practices involved in data matching over time and across various organizations. The data matching system's use in identification for migration and border control was shown as not predetermined. Instead, we found that the system's securitization was influenced by specific choices made by key actors and the changing sociotechnical landscape of identification and data matching. Through this analysis, we encountered unexpected connections between software suppliers and their customers, highlighting the intricate networks that underlie identification infrastructures. These findings challenge simplistic narratives and underscore the necessity of adopting a more performative understanding of the sociotechnical dynamics that shape identification practices. By emphasizing contingency in the system's evolution, we found moments where the outcomes in development are not predetermined but rather influenced by the specific circumstances, factors, and decisions made by the individuals and entities involved, thereby expanding our comprehension of the processes underpinning identification technologies' development.

In conclusion, this dissertation has sought to unravel the intricate dynamics of practices and technologies for matching identity data in the sensitive domains of migration management and border control, elucidating how these processes both shape and are shaped by the broader, transnational commercialized security infrastructures. This dissertation's performative approach to data matching has underscored that the intersection of identification technologies, data matching, and securitization is far from deterministic or linear. Instead, it is shaped by contingent choices and sociotechnical dynamics, emphasizing that alternative courses of action have been and continue to be possible. In this ever-evolving landscape, the actors involved play a pivotal role in shaping the direction of identification technologies, ultimately influencing the complex interplay of identity, security, and migration in our contemporary world. In a world where borders, identities, and security technologies are becoming more intertwined, this research highlights the importance of maintaining a critical and adaptable approach when examining identification systems.

Appendix A

Supplement to Chapter 4

A.1 Definitions 1: Graphs, nodes, links

We recall that a graph is formally defined as a pair $G = (N, L)$. Where N is a set whose elements are the nodes (also called vertices or points), and L is a set of links (also called edges) which are ordered pairs of distinct nodes. L is a subset of the set of all possible links between nodes, where in our case a node cannot be associated to itself: $L \subseteq \{(x, y) | (x, y) \in N^2 \wedge x \neq y\}$. For a link (x, y), x and y are called the *endpoints* of a link. In our approach, links are *undirected* and a link represents an occurrence of x in y. As a shorthand we write the link between x and y as l_{xy}. For example, the link l_{xy} can represent the occurrence of a category 'date of birth' (node x) in the document group 'Eurodac' (node y). All categories present in the document group have corresponding nodes and links in the graph. In fact, we can treat each data model (i.e., document group) as a separate graph. The complete graph is then in effect the combination of different data models. For each data model as a separate graph G_i, the combined graph is the disjoint union of graphs: $G = \bigcup_{i \in I} G_i$.

A.2 Definitions 2: Attributes

In our graph model, nodes are objects composed of attributes that are used to keep metadata of nodes. These attributes are formulated using the notation $n.a$ for an attribute a of a node n. The most important metadata kept for a node are $n.name$ and $n.type$, where $name$ is the natural language label of the node. The attribute $type$ can only take a limited set of values:
$type \in \{category, categoryValue, codeGroup, document, documentGroup\}$.

A.3 Definitions 3: Graph drawing

A drawing of a graph $G = (N, L)$ is a collection of points in a two-dimensional space. Each point p_i with coordinates x and y is the position of the node n_i in the layout. Whenever there exists a link $(p_i, p_j) \in L$, a line is drawn between points p_i and p_j. The task of the layout algorithm is to find a positioning of points so that specific criteria are optimally met. Examples of commonly used criteria are: nodes should not overlap, neighbouring nodes should be grouped together, the number of crossing link should be minimised. Each algorithm and set of criteria has its own benefits and drawbacks.

A.4 Definitions 4: Degree & neighbourhood

For a node n_jn the degree is defined as the number of links a node has:
$deg(x) = |\{n_j : l_{ij} \in L\}|$. The set of linked nodes is called the *neighbourhood* of a node. The neighbourhood H_i for a node n_j is defined as: $H_i = \{n_j : l_{ij} \in L \vee l_{ji} \in L\}$.

A.5 Definitions 5: Betweenness centrality

The betweenness centrality of a node n is defined as $bc(n) = \sum_{s \neq n \neq t} \frac{\sigma_{st}(n)}{\sigma_{st}}$. Where σ_{st} is the total amount of shortest paths from node s to node t and $\sigma_{st}(n)$ is the amount of those paths that pass through n. A path is a sequence of nodes, where each pair of nodes in the sequence is linked. The shortest path is the path between two nodes s and t that traverses the smallest number nodes. The equation for betweenness centrality takes into account that there may be several possible paths from s to t, with only some passing through n.

A.6 Definitions 6: Presence

The presence of all categories in a document group node n_x is a set of all category nodes $Categories(x) = \{n_y \in N : (l_{xy} \in L \vee l_{yx} \in L) \wedge n_y.type = category\}$. The presence of a category n_x in a document group is the set of nodes of the type document group for which there exist a link between this category and the document group. Formally defined:
$Presence(x, documentGroup) = \{n_y \in N : l_{xy} \in Links \wedge y.type = docGroup\}$.

A.7 Definitions 7: Intersection and difference

The absence of categories between a $docGroup_1$ and $docGroup_2$ is the set of categories present in the second document group minus the set of categories present in the first. In our notation: $Absence(docGroup_1, docGroup_2) =$

$\{Categories(docGroup_2) \setminus Categories(docGroup_1)\}$. The categories that are common between those same two document groups are determined using the intersection of the sets of categories that are present in either: $CommonCodes(docGroup_1, docGroup_2) = docGroup_1 \cap docGroup_2$. This operation is not limited to two sets. The intersection between more sets can be notated as $\bigcap_{i=1}^{n} Presence(docGroup_i)$.

A.8 Table presence of code groups for authorities

Table A.1: Presence of code groups for EU (Eurodac, SIS, VIS), Greek (HRF), German (GRF), and their relative degree and betweenness centrality.

Degree	Code group	EU	Greek	German	Betweenness
3					
	asylum				
	biometrics				
	biometrics: photograph data				
	birth data				
	birth data: date of birth				
	birth data: place of birth				
	education				
	name: surname				
	nationality				
	occupation data				
	residence				
	sex & gender				
	travel document data				

Table A.1: Presence of code groups for EU (Eurodac, SIS, VIS), Greek (HRF), German (GRF), and their relative degree and betweenness centrality.

Degree	Code group	EU	Greek	German	Betweenness
2					
	additional info / comments	■	■		
	biometrics: fingerprint data	■		■	
	contact info		■	■	
	criminal offence data	■		■	
	education: extent		■	■	
	family status data		■	■	
	language		■	■	
	linking data: EU		■	■	
	name: earlier/other names	■		■	
	name: forename	■		■	
	occupation data: current	■		■	
	parents data	■	■		
	personal ties	■	■		
	personal ties: in EU	■	■		
	procedure data	■	■		
	religion		■	■	
	residence: previous		■	■	
	travel: relocation	■	■		
1					

Table A.1: Presence of code groups for EU (Eurodac, SIS, VIS), Greek (HRF), German (GRF), and their relative degree and betweenness centrality.

Degree	Code group	EU	Greek	German	Betweenness
	application data				
	asylum: rejection				
	citizenship				
	country of origin				
	date of application				
	date of entry				
	date of exit				
	ethnicity				
	integration				
	language: speaking				
	law enforcement				
	law enforcement: extradition				
	law enforcement: investigation				
	law enforcement: unauthorized entry and residence				
	linking data: MS				
	linking data: responsibility				
	name: variations				
	occupation data: past				
	operator data				

Table A.1: Presence of code groups for EU (Eurodac, SIS, VIS), Greek (HRF), German (GRF), and their relative degree and betweenness centrality.

Degree	Code group	EU	Greek	German	Betweenness
	registration status			■	
	residence: current			■	
	residency request data			■	
	restrictions: movement		■		
	stay data	■			
	temporary accomodation/housing		■		
	travel data	■			
	travel document: validity: expiration	■			
	travel document: visa-related data	■			
	travel: rejection or removal			■	
	vulnerability		■		

Appendix B

Supplement to Chapter 5

B.1 Interview protocol

- Kun je me kort in je eigen woorden uitleggen wat je rol binnen de IND is, wat je taken zijn/waren en hoe je de applicaties voor het zoeken naar personen hebt gebruikt?
- Welke informatie heb je nodig voordat je een zoekopdracht kan doen?
  - Zijn er bepaalde situaties waar je vooraf al weet of een zoekopdracht niet goed gaat lukken?
  - Wat zijn de minimum velden denk je die nodig zijn voor een goede/exacte match?
  - Helpt de applicatie je bij het formuleren van een zoekopdracht? Hoe zou dit volgens jou kunnen gebeuren?
- Kun je me uitleggen welke velden je invult voor het zoeken naar personen.
  - Zijn er volgens jou bepaalde velden die betere resultaten opleveren?
  - In de zoekvelden kun je soms ook gebruik maken van speciale symbolen zoals een sterretje (* wildcards) waardoor je de zoekopdracht flexibeler kan maken. Maak je hiervan gebruik? Indien ja, voor welke velden wel en welke niet?
  - Ben je er bewust van dat er ook bepaalde functionaliteiten in de zoekapplicatie zitten die bijvoorbeeld makkelijker moeten maken om te zoeken met verschillende alfabetten of verschillen in hoe namen fonetisch klinken enz.
- Kun je me uitleggen hoe je de juiste persoon vindt in de lijst van resultaten?
  - Is het altijd duidelijk waarom bepaalde resultaten in de lijst staan?
  - Als er zeer vergelijkbare resultaten zijn in de lijst, hoe identificeer je dan de juiste persoon?
  - Zijn er soms te veel resultaten in de lijst volgens jou?

- In de lijst van resultaten staat er normaal steeds een match score, weet jij hoe die berekend wordt?
  - Zijn de match scores die worden toegekend aan de resultaten nuttig volgens jou?
  - Hoe zouden deze match scores verbeterd kunnen worden?
- Vind je de lijst van resultaten betrouwbaar? Hoe kun je dit volgens jou zien?
  - Heb je soms duplicaten in de resultatenlijst gezien? Hoe zie je dit? En waarom denk je dat dit soms voorkomt?
  - Ik heb vernomen dat er soms ook een memo aan personen gehangen wordt voor extra informatie, kun je me uitleggen hoe dit juist gebruikt wordt?
- Ik heb begrepen dat sommige personen in de database relaties hebben met andere personen. Wat soort relaties zijn dit? En heeft dit een invloed op hoe je zoekopdrachten doet?
- Op welke databases worden de zoekopdrachten uitgevoerd volgens jou? Enkel de system binnen de IND, of ook van andere partners?
  - Zijn er verschillen in hoe personen in deze andere system staan?
- Hoe heb jij gebruik leren maken van de applicaties en zoekfunctionaliteiten. Was dit via ervaringsdeskundigen, collega's, of?

B.1.1 Inkomende poststukken vragen

- Kun je me uitleggen hoe je het verschil vindt of je moet zoeken op een al bestaande klant of een nieuwe klant zal moeten aanmaken.
  - Help de applicatie je hiermee? Hoe zou dit volgens jou beter kunnen?

B.1.2 Deduplicatie vragen

- Volgens welke criteria worden er volgens jou duplicaten gevonden?
- Waarom zijn er volgens jou duplicaten in de IND database en waarom is het belangrijk om deze te vinden en corrigeren?
- Hoe gebeurd de rapportage van duplicaten? Wat zijn de stappen die je onderneemt nadat je deze resultaten hebt gekregen?
  - Hoe ga je dubbele klanten samenvoegen? Treden er soms moeilijkheden op?
- Ik heb gehoord dat de historische waarden belangrijk zijn voor IND. In hoeverre is dit ook zo voor de deduplicatie?

Table B.1: Examples of search strategies

Code	Comment
add more data	Repeatedly adding more data in the search query to exclude some results.
build up the query	Assumes that too broad a query will lead to bad results and that it is therefore better to progressively add data.
common biographic data	From experiene a user may know that certain common name combinations will lead to too many results and they will need to adapt their query.
first search	Using specific categories for a first search, e.g. first and last name.
first search on personal	First searching on basic personal data such as nationality and data of birth, and then add more if needed.
make search as wide as possible	Trying to include many categories of data for the search query to broaden he search.
searching for a person through relations	Try to find a person by using other relations, for example checking the clients of a lawyer a person is connected to.
trial and error	Using different kinds of combinations of categories of data in a trial-and-error fashion.
trying different combinations	Using different kinds of combinations of variations of the categories data, e.g. manually trying out name variations.

B.2 Search strategies

Table B.1 presents the search strategies derived from the analysis of interview data. These strategies are grounded in the descriptions provided by individuals within the IND, outlining their methods for potential applicant re-identification.

Appendix C

Supplement to Chapter 6

C.1 Interview questions

- Can you briefly explain to me your position/function at WCC?
- Identity matching:
  - What are some challenges of matching identity data across different organizations, according to you?
  - What about matching identity data across different locales, for example, between EU Member States or national and international organizations or institutions?
  - What is the role of WCC solutions in making identity data interoperable?
- From your experience, are some kinds or categories of data more important for matching identity data?
- Can you tell me something about these needs of the customers. Do they always know what they want in how the data matching should work?
- What are some of the most important features of WCC solutions that you highlight to the customer? And vice versa, what features may the customer see as important when you talk to them?
- Generification of the software:
  - How can software such as identity matching work across different contexts/domains be made possible? For example, how can it be made to work in employment and security?
  - In the case of security, are there differences and challenges to doing such things as identity matching in domains such as policing vs. migration or travel?
- Tensions of customizable/generic software for the customer.

  - At the IND, I found that some of the challenges for users to work with the search tools are due to the way ELISE is deployed as a very generic service in the SOA, with few adaptations for the specific context of the IND.

- I have sometimes heard the WCC solution described as "vendor-neutral." What is your view on this issue?
  - Do you know of cases when proprietary formats or algorithms were an issue?
  - How can software such as ELISE work with different data formats?

Summary

This dissertation analyzes the interconnections between data matching technologies, identification practices, and transnational commercialized security infrastructures, particularly in relation to migration management and border control. The research was motivated by a curiosity about the intersection between identity data matching and the challenges authorities encounter when identifying individuals, especially the "blind spots" caused by incomplete data, aliases, and uncertainties. The dissertation addresses the following main research question: "How are practices and technologies for matching identity data in migration management and border control shaping and shaped by transnational commercialized security infrastructures?"

The dissertation begins by presenting an overview of the literature regarding the connections between data matching technology, which is used across various sectors, and its interrelationships with the internationalization, commercialization, securitization, and infrastructuring of identification infrastructure. This overview highlights a noticeable gap in the understanding of how data matching influences the meaning of the interconnected data and shapes relationships between organizations that use it. To address this gap, Chapter 3 proposes a methodological framework for using data matching as both a research topic and a resource for answering specific sub-questions related to specific aspects of data matching.

Chapter 4 emphasizes the significance of data models in information systems for categorizing individuals and establishing connections between different data models for accurate matching. The analysis of this aspect of data matching is made possible by introducing the "Ontology Explorer", which serves as a novel method for examining the knowledge and assumptions embedded within data models. By applying this method to analyze national and transnational data infrastructures for population management, this method is shown to reveal authorities' imaginaries on people-on-the-move. In this way, the method demonstrates the importance of data categories in data models, as they are crucial for data matching while also offering valuable insights into how authorities enact people in different ways.

Following that, the dissertation investigates how identity data matching is employed to re-identify applicants within a government migration and asylum agency in The Netherlands. Chapter 5 introduces the concept of re-identification, which involves the

ongoing utilization and integration of data from various sources to establish whether multiple sets of identity data pertain to a single individual. This chapter uses insights gathered from interviews with personnel from the agency to investigate the integration of data matching tools for re-identification. The chapter shows that striving to minimize data friction in re-identification through data matching can have unintended consequences and additional burdens for the agency's personnel.

Lastly, this dissertation examines the evolution of a commercial data matching system employed for identification and security, adopting a sociotechnical approach. Chapter 6 introduces heuristics that are then used to identify moments that emphasize the design contingencies of the data matching system. Through the examination of fieldwork data collected from the company that created the system, the chapter highlights the reciprocal influences between the system's design and the actors and entities involved. The system experienced adaptive and contingent changes from a generic data matching system to a specialized tool for identification and security because of such influences. In a broader sense, the chapter brings attention to the interrelationships among software suppliers, integrators, and customers, and the circulation and use of knowledge and technology for matching identity data across organizations.

Nederlandse samenvatting (Summary in Dutch)

Het vinden van blinde vlekken: Onderzoek naar identiteitsdatamatching in transnationale gecommercialiseerde veiligheidsinfrastructuur

Deze dissertatie onderzoekt het samenspel tussen datamatching technologieën, identificatiepraktijken en transnationale gecommercialiseerde veiligheidsinfrastructuur, met een specifieke focus op migratie- en grenscontrole. Het onderzoek werd gedreven door een nieuwsgierigheid naar de intersectie tussen datamatching van identiteiten en de zogenaamde "blinde vlekken" waarmee autoriteiten worden geconfronteerd bij het identificeren van individuen, zoals wanneer zij te maken hebben met onvolledige gegevens, aliassen en andere onzekerheden. De dissertatie behandelt de volgende hoofdonderzoeksvraag: "Hoe worden praktijken en technologieën voor het matchen van identiteitsgegevens in migratiemanagement en grenscontrole gevormd door en vormen zij op hun beurt transnationale gecommercialiseerde veiligheidsinfrastructuren?"

De dissertatie begint met een overzicht van de bestaande literatuur over de verbanden tussen datamatchingtechnologie, die in verschillende sectoren wordt gebruikt, en de relaties met de internationalisering, commercialisering, securitisatie en infrastructurering van identificatie. Dit overzicht benadrukt een merkbare kloof in het begrip van hoe datamatching de betekenis van de onderling verbonden gegevens beïnvloedt en de relaties tussen organisaties die het gebruiken vormgeeft. Om deze kloof te dichten, stelt Hoofdstuk 3 een methodologisch kader voor voor het gebruik van datamatching als zowel een onderzoeksonderwerp als een hulpmiddel voor het beantwoorden van specifieke deelvragen met betrekking tot specifieke aspecten van datamatching.

Hoofdstuk 4 benadrukt het belang van datamodellen in informatiesystemen voor het categoriseren van individuen en het tot stand brengen van verbindingen tussen verschillende datamodellen om gegevens te kunnen matchen. De analyse van dit aspect van datamatching wordt mogelijk gemaakt door de introductie van de "Ontology Explorer", die dient als een nieuwe methode voor het onderzoeken van de kennis en aannames die in datamodellen zijn ingebed. Door deze methode toe te passen om nationale en transnationale datainfrastructuren voor bevolkingsbeheer wordt het mogelijk om de denkbeelden van nationale en EU autoriteiten voor bevolkingsbeheer te onderzoeken. De categorieën van datamodellen komen zo naar voren als een belangrijke manier om

inzicht te krijgen in hoe verschillende autoriteiten mensen op verschillende manieren beschouwen en verwezenlijken.

Vervolgens onderzoekt de dissertatie hoe identiteitsdatamatching wordt gebruikt om aanvragers opnieuw te identificeren binnen de processen van een overheidsmigratie- en asielagentschap.Hoofdstuk 5 introduceert het concept van re-identificatie als het herhaaldelijk gebruiken en verbinden van gegevens uit verschillende bronnen om vast te stellen of meerdere sets identiteitsgegevens betrekking hebben op één individu. Het hoofdstuk gebruikt inzichten uit interviews met werknemers van de Immigratie- en Naturalisatiedienst (IND) om de integratie van datamatchingtools voor re-identificatie te onderzoeken. Daaruit blijkt dat hoewel het doel van datamatching vaak is om frictie in re-identificatie te minimaliseren, het gebruik van deze technologie kan leiden tot onvoorziene uitdagingen voor organisaties.

Tot slot onderzoekt deze dissertatie de historische ontwikkeling van een commercieel datamatchingsysteem dat wordt gebruikt voor identificatie- en veiligheidsdoeleinden vanuit een sociotechnisch perspectief. Hoofdstuk 6 introduceert heuristieken die vervolgens worden gebruikt de ontwerpcontingenties van het datamatchingsysteem te identificeren. Op basis van gegevens van interviews en veldwerk verzameld bij het bedrijf dat het systeem heeft gecreëerd, belicht het hoofdstuk de wederzijdse invloeden tussen het ontwerp van het systeem en de betrokken actoren en entiteiten. Het systeem onderging adaptieve en contingente veranderingen van een generiek datamatchingsysteem naar een gespecialiseerd hulpmiddel voor identificatie en beveiliging als gevolg van dergelijke invloeden. Het onderzoek belicht zo de interconnecties tussen softwareleveranciers, integratoren en klanten, evenals de circulatie en het gebruik van kennis en technologie voor het matchen van identiteitsgegevens tussen organisaties.

List of other publications

The contributions below were specified according to the 'CRediT system' (Contributor Roles Taxonomy).[1]

Pelizza A and Van Rossem W (2021) Sensing European alterity: An analogy between sensors and hotspots in transnational security networks. In: Klimburg-Witjes N, Pöchhacker N and Bowker GC (eds.) *Sensing in/Security: Sensors as Transnational Security Infrastructures*. Manchester, UK: Mattering Press. ISBN 978-1-912729-10-4, pp. 262–286.

- **Pelizza, Annalisa**: Conceptualization, Data curation, Funding acquisition, Formal Analysis, Investigation, Project administration, Supervision, Writing — original draft, Writing — review & editing.
- **Van Rossem, Wouter**: Data curation, Formal Analysis, Visualization, Writing — original draft (including section entitled "Separation of concerns as design criterion"), Writing — review & editing.

Van Rossem, Wouter. 2021. "Ontology-Explorer". Zenodo. https://doi.org/10.5281/zenodo.4899316.

Van Rossem W and Pelizza A (2022) The Ontology Explorer: A method to make visible data infrastructures for population management. *Big Data & Society* 9(1): 1–18. DOI:10.1177/20539517221104087.

- **Van Rossem, Wouter**: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing — original draft, Writing — review & editing.
- **Pelizza, Annalisa**: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing — original draft, Writing — review & editing.

[1] https://credit.niso.org/

Pelizza A and Van Rossem W (2023) Scripts of alterity: Mapping assumptions and limitations of the border security apparatus through classification schemas. *Science, Technology, & Human Values* 0(0): 1–33. DOI:10.1177/01622439231195955.

- **Pelizza, Annalisa**: Conceptualization, Data curation, Funding acquisition, Formal Analysis, Investigation, Project administration, Supervision, Writing — original draft, Writing — review & editing.
- **Van Rossem, Wouter**: Data curation, Formal Analysis, Software, Visualization, Writing — original draft (including sections related to comparative analysis using The Ontology Explorer), Writing — review & editing.

Bibliography

Abbott O, Jones P and Ralphs M (2015) Large-scale linkage for total populations in official statistics. In: *Methodological Developments in Data Linkage*. Chichester: John Wiley & Sons, Ltd. ISBN 978-1-119-07245-4, pp. 170–200. DOI: 10.1002/9781119072454.ch8.

About I, Brown J and Lonergan G (eds.) (2013a) *Identification and Registration Practices in Transnational Perspective*. St Antony's Series. London: Palgrave Macmillan. ISBN 978-1-349-34643-1. DOI: 10.1057/9781137367310.

About I, Brown J and Lonergan G (2013b) Introduction. In: *Identification and Registration Practices in Transnational Perspective*, St Antony's Series. London: Palgrave Macmillan. ISBN 978-1-349-34643-1, pp. 1–13. DOI: 10.1057/9781137367310.

Accenture (2010) Unique Identification Authority of India (UIDAI) Selects Accenture to Implement a Multimodal Biometric Solution for "Aadhaar" Program. URL https://web.archive.org/web/20220812203854/https://newsroom.accenture.com/article_display.cfm?article_id=5040.

Accenture (2012) European Commission selects consortium of Accenture, Morpho and HP to maintain EU visa information and biometric matching systems. URL https://web.archive.org/web/20201206154800/https://newsroom.accenture.com/subjects/client-winsnew-contracts/european-commission-chooses-consortium-of-accenture-morpho-and-hp-to-maintain-eu-visa-information-and-biometric-matching-systems.htm.

Accenture (2015) United Nations High Commissioner for Refugees and Accenture deliver global biometric identity management system to aid displaced persons. URL https://web.archive.org/web/20221203234022/https://newsroom.accenture.com/news/united-nations-high-commissioner-for-refugees-and-accenture-deliver-global-biometric-identity-management-system-to-aid-displaced-persons.htm.

Adler E (1997) Seizing the middle ground: Constructivism in world politics. *European Journal of International Relations* 3(3): 319–363. DOI: 10.1177/1354066197003003003.

Ajana B (2013) Asylum, identity management and biometric control. *Journal of Refugee Studies* 26(4): 576–595. DOI: 10.1093/jrs/fet030.

Akrich M (1992) The de-scription of technical objects. In: Bijker WE and Law J (eds.) *Shaping Technology/Building Society: Studies in Sociotechnical Change*, Inside Technology. Cambridge, Mass.: The MIT Press, pp. 205–224.

Akrich M and Latour B (1992) A summary of a convenient vocabulary for the semiotics of human and nonhuman assemblies. In: Bijker WE and Law J (eds.) *Shaping Technology/Building Society:*

Studies in Sociotechnical Change. Cambridge, Mass.: The MIT Press, pp. 259–264.

Amelung N (2021) "Crimmigration control" across borders: The convergence of migration and crime control through transnational biometric databases. *Historical Social Research* 46(3): 151–177. DOI: 10.12759/HSR.46.2021.3.151-177.

Amicelle A, Aradau C and Jeandesboz J (2015) Questioning security devices: Performativity, resistance, politics. *Security Dialogue* 46(4): 293–306. DOI: 10.1177/0967010615586964.

Amoore L (2006) Biometric borders: Governing mobilities in the war on terror. *Political Geography* 25(3): 336–351. DOI: 10.1016/j.polgeo.2006.02.001.

Amoore L (2013) *The Politics of Possibility: Risk and Security beyond Probability*. Duke University Press. ISBN 978-0-8223-7726-9.

Amoore L and de Goede M (2005) Governance, risk and dataveillance in the war on terror. *Crime, Law and Social Change* 43(2-3): 149–173. DOI: 10.1007/s10611-005-1717-8.

Aradau C (2004) Security and the democratic scene: Desecuritization and emancipation. *Journal of International Relations and Development* 7(4): 388–413. DOI: 10.1057/palgrave.jird.1800030.

Aradau C (2012) Infrastructure. In: Salter MB and Mutlu CE (eds.) *Research Methods in Critical Security Studies: An Introduction*. London & New York: Routledge. ISBN 978-0-203-10711-9, pp. 181–185.

Aradau C and Blanke T (2021) Algorithmic surveillance and the political life of error. *Journal for the History of Knowledge* 2(1): 1–13. DOI: 10.5334/jhk.42.

Aradau C and Blanke T (2022) *Algorithmic Reason: The New Government of Self and Other*. Oxford, UK: Oxford University Press. ISBN 978-0-19-285962-4. DOI: 10.1093/oso/9780192859624.001.0001.

Aradau C and Huysmans J (2014) Critical methods in International Relations: The politics of techniques, devices and acts. *European Journal of International Relations* 20(3): 596–619. DOI: 10.1177/1354066112474479.

Baird T (2017) Knowledge of practice: A multi-sited event ethnography of border security fairs in Europe and North America. *Security Dialogue* 48(3): 187–205. DOI: 10.1177/0967010617691656.

Balzacq T (2005) The three faces of securitization: Political agency, audience and context. *European Journal of International Relations* 11(2): 171–201. DOI: 10.1177/1354066105052960.

Balzacq T (2009) Constructivism and securitization studies. In: Dunn Cavelty M and Mauer V (eds.) *The Routledge Handbook of Security Studies*. London: Routledge. ISBN 978-0-203-86676-4, pp. 56–72. DOI: 10.4324/9780203866764.

Balzacq T, Basaran T, Bigo D, Guittet EP and Olson C (2010) Security practices. In: *Oxford Research Encyclopedia of International Studies*. Oxford: Wiley Blackwell.

Basis Technology (2012) Understanding Dari and Pashto names: A challenge to intelligence gathering in Afghanistan. URL https://web.archive.org/web/20230605072915/https://www.rosette.com/blog/understanding-dari-and-pashto-names-a-challenge-to-intelligence-gathering-in-afghanistan/.

Basis Technology (2021) Strengthening U.S. borders with intelligent name matching. URL https://web.archive.org/web/20211105135209/https://www.rosette.com/case-studies/us-customs

-border-protection/.

Bates J (2017) The politics of data friction. *Journal of Documentation* 74(2): 412–429. DOI: 10.1108/JD-05-2017-0080.

Batini C and Scannapieco M (2016) Object identification. In: Batini C and Scannapieco M (eds.) *Data and Information Quality: Dimensions, Principles and Techniques*, Data-Centric Systems and Applications. Cham: Springer International Publishing. ISBN 978-3-319-24106-7, pp. 177–215. DOI: 10.1007/978-3-319-24106-7_8.

Bellahsene Z, Bonifati A and Rahm E (eds.) (2011) *Schema Matching and Mapping*. Berlin, Heidelberg: Springer. ISBN 978-3-642-16518-4. DOI: 10.1007/978-3-642-16518-4.

Bellanova R and de Goede M (2022) The algorithmic regulation of security: An infrastructural perspective. *Regulation & Governance* 16(1): 102–118. DOI: 10.1111/rego.12338.

Bellanova R and Duez D (2012) A different view on the 'making' of European security: The EU Passenger Name Record system as a socio-technical assemblage. *European Foreign Affairs Review* 17(SI). DOI: 10.54648/eerr2012017.

Bellanova R and Glouftsios G (2022) Controlling the Schengen Information System (SIS II): The infrastructural politics of fragility and maintenance. *Geopolitics* 27(1): 160–184. DOI: 10.1080/14650045.2020.1830765.

Benjamin R (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity. ISBN 978-1-5095-2643-7.

Benjamin R and Wigand R (1995) Electronic markets and virtual value chains on the information superhighway. *MIT Sloan Management Review* URL https://sloanreview.mit.edu/article/electronic-markets-and-virtual-value-chains-on-the-information-superhighway/.

Benson M and O'Reilly K (2009) Migration and the Search for a Better Way of Life: A Critical Exploration of Lifestyle Migration. *The Sociological Review* 57(4): 608–625. DOI: 10.1111/j.1467-954X.2009.01864.x.

Beraldo D and Milan S (2019) From data politics to the contentious politics of data. *Big Data & Society* 6(2): 1–11. DOI: 10.1177/2053951719885967.

Bergsma B (2013) Systeem IND duurder en trager ingevoerd. *Algemeen Nederlands Persbureau ANP* URL https://web.archive.org/web/20201013120039/https://www.nu.nl/binnenland/3628805/systeem-ind-duurder-en-trager-ingevoerd.html.

Bertrand S (2018) Can the subaltern securitize? Postcolonial perspectives on securitization theory and its critics. *European Journal of International Security* 3(3): 281–299. DOI: 10.1017/eis.2018.3.

Betlem R (2011) Utrechtse datatechnologie moet terroristen buiten de VS houden. *Het Financieele Dagblad* .

Bigo D (2014) The (in)securitization practices of the three universes of EU border control: Military/Navy – border guards/police – database analysts. *Security Dialogue* 45(3): 209–225. DOI: 10.1177/0967010614530459.

Bigo D, Carrera S, Hayes B, Hernanz N and Jeandesboz J (2012) *Justice and Home Affairs Databases and a Smart Borders System at EU External Borders: An Evaluation of Current and Forthcoming Proposals*. Brussels: Centre for European Policy Studies. ISBN 978-94-6138-258-0. URL https://www.ce

ps.eu/ceps-publications/justice-and-home-affairs-databases-and-smart-borders-system-eu-external-borders/.

Bijker WE (1993) Do not despair: There is life after constructivism. *Science, Technology, & Human Values* 18(1): 113–138. DOI: 10.1177/016224399301800107.

Bijker WE, Hughes TP and Pinch T (eds.) (2012) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Anniversary ed edition. Cambridge, Mass: MIT Press. ISBN 978-0-262-51760-7.

Bijker WE and Law J (eds.) (1992) *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Inside Technology. Cambridge, Mass: MIT Press. ISBN 978-0-262-02338-2.

Biometric Technology Today (2002) 2001 market review: Uncertain times. *Biometric Technology Today* 10(1): 9–11. DOI: 10.1016/S0969-4765(02)00118-2.

Bloomfield BP and Vurdubakis T (1997) Visions of organization and organizations of vision: The representational practices of information systems development. *Accounting, Organizations and Society* 22(7): 639–668. DOI: 10.1016/S0361-3682(96)00024-4.

Borgman CL (2015) *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Mass.; London, Eng.: The MIT Press. ISBN 978-0-262-02856-1.

Bovens M and Zouridis S (2002) From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review* 62(2): 174–184. DOI: 10.1111/0033-3352.00168.

Bowker GC (1994) *Science on the Run: Information Management and Industrial Geophysics at Schlumberger, 1920-1940*. Inside Technology. Cambridge, Mass.: The MIT Press. ISBN 978-0-262-02367-2.

Bowker GC, Baker K, Millerand F and Ribes D (2009) Toward information infrastructure studies: Ways of knowing in a networked environment. In: Hunsinger J, Klastrup L and Allen M (eds.) *International Handbook of Internet Research*. Dordrecht: Springer Netherlands. ISBN 978-1-4020-9788-1, pp. 97–117. DOI: 10.1007/978-1-4020-9789-8_5.

Bowker GC and Star SL (1999) *Sorting Things out: Classification and Its Consequences*. Inside Technology. Cambridge, Mass.: The MIT press.

Breckenridge K and Szreter S (eds.) (2012) *Registration and Recognition: Documenting the Person in World History*. Oxford: Oxford University Press for the British Academy. ISBN 978-0-19-726531-4.

Broeders D (2011) A European 'border' surveillance system under construction. In: Dijstelbloem H and Meijer A (eds.) *Migration and the New Technological Borders of Europe*, Migration, Minorities and Citizenship. London: Palgrave Macmillan, pp. 40–67. DOI: 10.1057/9780230299382_3.

Broeders D and Dijstelbloem H (2016) The datafication of mobility and migration management: The mediating state and its consequences. In: van der Ploeg I and Pridmore J (eds.) *Digitizing Identities: Doing Identity in a Networked World*. London: Routledge, pp. 242–260.

Buffat A (2015) Street-level bureaucracy and e-government. *Public Management Review* 17(1): 149–161. DOI: 10.1080/14719037.2013.771699.

Bundesamt für Migration und Flüchtlinge (2020) Standard XAusländer. URL https://web.archive.org/web/20200621191446/https://www.bamf.de/DE/Themen/Digitalisierung/Xauslaender/

xauslaender-node.html.

Bundesverwaltungsamt (2021) Ausländerzentralregister. URL http://web.archive.org/web/20211104160414/https://www.bva.bund.de/DE/Das-BVA/Aufgaben/A/Auslaenderzentralregister/azr_node.html.

Burgess JP (ed.) (2010) *The Routledge Handbook of New Security Studies*. Routledge Handbooks. London; New York: Routledge. ISBN 978-0-415-48437-4.

Burns R and Wark G (2020) Where's the database in digital ethnography? Exploring database ethnography for open data research. *Qualitative Research* 20(5): 598–616. DOI: 10.1177/1468794119885040.

Busch PA and Henriksen HZ (2018) Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity* 23(1): 3–28. DOI: 10.3233/IP-170050.

Bush GW (2008) NSPD-59 / HSPD-24 on biometrics for identification and screening to enhance national security. URL http://web.archive.org/web/20221006170237/https://irp.fas.org/offdocs/nspd/nspd-59.html.

Buzan B, Wæver O and de Wilde J (1998) *Security: A New Framework for Analysis*. Boulder, Colorado: Lynne Rienner Pub. ISBN 978-1-55587-603-6.

Cakici B, Ruppert E and Scheel S (2020) Peopling Europe through data practices: Introduction to the special issue. *Science, Technology, & Human Values* 45(2): 199–211. DOI: 10.1177/0162243919897822.

Calder S (2022) EU brings in vaccine expiration date of 270 days for travellers. URL https://web.archive.org/web/20220307103943/https://www.independent.co.uk/travel/news-and-advice/eu-vaccine-expiration-date-travel-270-b2004777.html.

Callon M (1984) Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc bay. *The Sociological Review* 32(1_suppl): 196–233. DOI: 10.1111/j.1467-954X.1984.tb0011.

Caplan J and Torpey J (eds.) (2001) *Documenting Individual Identity: The Development of State Practices in the Modern World*. Princeton, N.J.; Oxford: Princeton University Press. ISBN 0-691-00911-2.

Caswell M (2012) Using classification to convict the Khmer Rouge. *Journal of Documentation* 68(2): 162–184. DOI: 10.1108/00220411211209177.

Christen P (2012) *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin; New York: Springer. ISBN 978-3-642-31164-2. DOI: 10.1007/978-3-642-31164-2.

CJK Dictionary Institute (2018) Database of Arabic names. URL http://web.archive.org/web/20230329222226/https://www.cjk.org/data/arabic/proper/database-arabic-names/.

Clarke AE (2005) *Situational Analysis: Grounded Theory after the Postmodern Turn*. Thousand Oaks, Calif.: Sage Publications. ISBN 978-0-7619-3055-6.

Clarke R (1994) Human identification in information systems: Management challenges and public policy issues. *Information Technology & People* 7(4): 6–37. DOI: 10.1108/09593849410076799.

Cole SA (2001) *Suspect Identities: A History of Fingerprinting and Criminal Identification*. Cambridge,

Mass.; London, Eng.: Harvard University Press. ISBN 978-0-674-00455-9.

Cole SA (2005) Review of imprint of the Raj: How fingerprinting was born in Colonial India. *Technology and Culture* 46(1): 252–253. DOI: 10.1353/tech.2005.0010.

Collins SB (2016) The space in the rules: Bureaucratic discretion in the administration of Ontario works. *Social Policy and Society* 15(2): 221–235. DOI: 10.1017/S1474746415000251.

Cornford J, Baines S and Wilson R (2013) Representing the family: How does the state 'think family'? *Policy & politics* 41(1): 1–18. DOI: 10.1332/030557312X645838.

Côté-Boucher K, Infantino F and Salter MB (2014) Border security as practice: An agenda for research. *Security Dialogue* 45(3): 195–208. DOI: 10.1177/0967010614533243.

Cowan RS (1985) How the refrigerator got its hum. In: *The Social Shaping of Technology*. Philadelphia: Open University Press. ISBN 978-0-335-19914-3, pp. 202–218.

Cresswell T (2010) Towards a politics of mobility. *Environment and Planning D: Society and Space* 28(1): 17–31. DOI: 10.1068/d11407.

Dalton C and Thatcher J (2014) What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space* URL http://web.archive.org/web/20230827144723/https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care.

David PA and Bunn JA (1988) The economics of gateway technologies and network evolution: Lessons from electricity supply history. *Information Economics and Policy* 3(2): 165–202. DOI: 10.1016/0167-6245(88)90024-8.

Davidshofer S, Jeandesboz J and Ragazzi F (2017) Technology and security practices: Situating the technological imperative. In: Basaran T, Bigo D, Guittet EP and Walker R (eds.) *International Political Sociology: Transversal Lines*, Routledge Studies in International Political Sociology. London & New York: Routledge. ISBN 978-1-138-91071-3.

De Genova N (ed.) (2017) *The Borders of "Europe": Autonomy of Migration, Tactics of Bordering*. Durham, NC: Duke University Press.

de Goede M, Bosma E and Pallister-Wilkins P (eds.) (2020) *Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork*. London & New York: Routledge. ISBN 978-0-367-02724-7.

de Goede M and Sullivan G (2016) The politics of security lists. *Environment and Planning D: Society and Space* 34(1): 67–88. DOI: 10.1177/0263775815599309.

D'Ignazio C and Klein LF (2020) *Data Feminism*. Strong Ideas Series. Cambridge, Mass.; London, Eng.: The MIT Press. ISBN 978-0-262-04400-4.

Dijstelbloem H (2021) *Borders as Infrastructure: The Technopolitics of Border Control*. Infrastructures. Cambridge, Mass.; London, Eng.: The MIT Press. ISBN 978-0-262-36638-0. DOI: 10.7551/mitpress/11926.001.0001.

Donko K, Doevenspeck M and Beisel U (2022) Migration control, the local economy and violence in the Burkina Faso and Niger borderland. *Journal of Borderlands Studies* 37(2): 235–251. DOI: 10.1080/08865655.2021.1997629.

Dourish P (2014) No SQL: The Shifting Materialities of Database Technology. *Computational Culture*

(4). URL http://web.archive.org/web/20230529102119/http://computationalculture.net/no-sql-the-shifting-materialities-of-database-technology/.

Dourish P (2017) *The Stuff of Bits: An Essay on the Materialities of Information*. Cambridge, Mass.: The MIT Press. ISBN 978-0-262-03620-7.

Dunn HL (1946) Record linkage. *American Journal of Public Health and the Nations Health* 36(12): 1412–1416. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1624512/.

ECA ECoA (2014) *Lessons from the European Commission's Development of the Second Generation Schengen Information System (SIS II)*, volume Special report No 03/2014. Luxembourg: Publications Office of the European Union. URL https://data.europa.eu/doi/10.2865/8113.

ECA ECoA (2020) *EU Information Systems Supporting Border Control: A Strong Tool, but More Focus Needed on Timely and Complete Data. Special Report No 20, 2019*. Luxembourg: Publications Office of the European Union. URL https://data.europa.eu/doi/10.2865/83092.

Edwards PN (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, Mass.; London, Eng: The MIT Press. ISBN 978-0-262-29071-5.

Edwards PN, Bowker GC, Jackson SJ and Williams R (2009) Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems* 10(5): 364–374. DOI: 10.17705/1jais.00200.

Edwards PN, Jackson SJ, Bowker GC and Knobel CP (2007) Understanding infrastructure: Dynamics, tensions, and design. Working Paper Final report of the workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures". URL http://deepblue.lib.umich.edu/handle/2027.42/49353.

Egyedi T (2001) Infrastructure flexibility created by standardized gateways: The cases of XML and the ISO container. *Knowledge, Technology & Policy* 14(3): 41–54. DOI: 10.1007/s12130-001-1015-4.

Emerson RM, Fretz RI and Shaw LL (2011) *Writing Ethnographic Fieldnotes*. Second edition. University of Chicago Press. ISBN 978-0-226-20686-8.

eu-LISA (2013) Report on the technical functioning of VIS, including the security thereof, pursuant to Article 50(3) of the VIS Regulation. Technical report.

eu-LISA (2016) *VIS Report Pursuant to Article 50(3) of Regulation (EC) No 767/2008: VIS Report Pursuant to Article 17(3) of Council Decision 2008/633/JHA. July 2016*. Luxembourg: Publications Office of the European Union. URL https://data.europa.eu/doi/10.2857/022699.

eu-LISA (2020) *Report on the Technical Function of the Visa Information System (VIS)*. Luxembourg: Publications Office of the European Union. URL https://data.europa.eu/doi/10.2857/66661.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press. ISBN 978-1-4668-8596-7.

European Commission (2016a) Commission Decision of 17 June 2016 setting up the High Level Expert Group on Information Systems and Interoperability. URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016D0715(01).

European Commission (2016b) Communication from the Commission to the European Parliament and the Council: Stronger and smarter information systems for borders and security. Technical Report COM/2016/0205 final. URL https://publications.europa.eu/en/publication-det

ail/-/publication/702fa423-fca2-11e5-b713-01aa75ed71a1/language-en/format-PDF.

European Commission (2017) Frequently asked questions - interoperability of EU information systems for security, border and migration management. Technical Report MEMO/17/5241, Strasbourg. URL https://ec.europa.eu/commission/presscorner/detail/en/MEMO_17_5241.

European Commission (2020a) Coronavirus: EU interoperability gateway. URL https://web.archive.org/web/20220831164020/https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1904.

European Commission (2020b) Factsheets Internal Security Fund - Police (ISFP). URL http://web.archive.org/web/20230920100118/https://www.sg.mai.gov.pt/FundosComunitarios/QFP20142020/Documents/Compilation_isfp_just.pdf.

European Commission (2021a) EU Digital COVID Certificate: EU Gateway goes live. URL https://web.archive.org/web/20221028104309/https://ec.europa.eu/commission/presscorner/detail/en/ip_21_2721.

European Commission (2021b) Questions and Answers – EU Digital COVID Certificate. URL http://web.archive.org/web/20221016124101/https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_2781.

European Commission (2022) eHealth and COVID-19. URL https://web.archive.org/web/20230110095405/http://health.ec.europa.eu/ehealth-digital-health-and-care/ehealth-and-covid-19_en.

European Commission (2023) COMMISSION IMPLEMENTING DECISION (EU) 2023/220 of 1 February 2023 laying down and developing the universal message format (UMF) standard pursuant to Regulation (EU) 2019/817 of the European Parliament and of the Council. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023D0220.

European External Action Service (2021) Non-EU countries welcome to join the EU Digital COVID certificate system. URL http://web.archive.org/web/20220816070933/https://www.eeas.europa.eu/eeas/non-eu-countries-welcome-join-eu-digital-covid-certificate-system_en.

European Union (2018) Regulation (EU) 2018/1726 of the European Parliament and of the Council of 14 November 2018 on the European Union Agency for the Operational Management of Large-Scale IT Systems in the Area of Freedom, Security and Justice (eu-LISA), and amending Regulation (EC) No 1987/2006 and Council Decision 2007/533/JHA and repealing Regulation (EU) No 1077/2011. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1726.

European Union (2019a) Regulation (EU) 2019/817 of the European Parliament and of the Council of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of borders and visa and amending Regulations (EC) No 767/2008, (EU) 2016/399, (EU) 2017/2226, (EU) 2018/1240, (EU) 2018/1726 and (EU) 2018/1861 of the European Parliament and of the Council and Council Decisions 2004/512/EC and 2008/633/JHA. URL http://data.europa.eu/eli/reg/2019/817/oj/eng.

European Union (2019b) Regulation (EU) 2019/818 of the European Parliament and of the Council of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of police and judicial cooperation, asylum and migration and amending

Regulations (EU) 2018/1726, (EU) 2018/1862 and (EU) 2019/816. URL http://data.europa.eu/eli/reg/2019/818/oj/eng.

European Union Agency for Fundamental Rights (2018) *Under Watchful Eyes: Biometrics, EU IT Systems and Fundamental Rights*. Luxembourg: Publications Office of the European Union. URL https://data.europa.eu/doi/10.2811/136698.

Europol (2014) *Universal Message Format: Faster, Cheaper, Better*. Luxembourg: Publications Office of the European Union. ISBN 978-92-95078-86-4. URL https://data.europa.eu/doi/10.2813/15318.

Euzenat J and Shvaiko P (2007) *Ontology Matching*. Berlin; New York: Springer. ISBN 978-3-540-49611-3. DOI: 10.1007/978-3-540-49612-0.

Fellegi IP and Sunter AB (1969) A theory for record linkage. *Journal of the American Statistical Association* 64(328): 1183–1210. DOI: 10.1080/01621459.1969.10501049.

Ferguson AG (2017) *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press. ISBN 978-1-4798-9282-2. DOI: 10.2307/j.ctt1pwtb27.

Flyverbom M and Murray J (2018) Datastructuring—Organizing and curating digital traces into action. *Big Data & Society* 5(2): 1–12. DOI: 10.1177/2053951718799114.

Follis KS (2017) Vision and transterritory: The borders of Europe. *Science, Technology, & Human Values* 42(6): 1003–1030. DOI: 10.1177/0162243917715106.

Fors-Owczynik KL and van der Ploeg I (2015) Migrants at/as risk: Identity verification and risk-assessment technologies in the Netherlands. In: van der Ploeg I and Pridmore J (eds.) *Digitizing Identities*, number 30 in Routledge Studies in Science, Technology and Society. New York; London: Routledge. ISBN 978-1-315-75640-0, pp. 261–281.

Franz M, Lopes CT, Huck G, Dong Y, Sumer O and Bader GD (2016) Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* 32(2): 309–311. DOI: 10.1093/bioinformatics/btv557.

Friese S (2014) *Qualitative Data Analysis with ATLAS.Ti*. Second edition. London: SAGE Publications Ltd. ISBN 978-1-4462-8203-8.

Gandy OH Jr (1989) The surveillance society: Information technology and bureaucratic social control. *Journal of Communication* 39(3): 61–76. DOI: 10.1111/j.1460-2466.1989.tb01040.x.

Garfinkel H (1964) Studies of the routine grounds of everyday activities. *Social Problems* 11(3): 225–250. DOI: 10.2307/798722.

Gargiulo E (2017) Monitoring or selecting? Security in Italy between surveillance, identification and categorisation. In: Orrù E, Porcedda MG and Weydner-Volkmann S (eds.) *Rethinking Surveillance and Control: Beyond the "Security versus Privacy" Debate*. Baden-Baden: Nomos. ISBN 978-3-8452-7809-4, pp. 195–215.

Gasson S (2006) A genealogical study of boundary-spanning IS design. *European Journal of Information Systems* 15(1): 26–41. DOI: 10.1057/palgrave.ejis.3000594.

Gazan R (2005) Imposing structures: Narrative analysis and the design of information systems. *Library & Information Science Research* 27(3): 346–362. DOI: 10.1016/j.lisr.2005.04.004.

Geiger SR and Ribes D (2011) Trace ethnography: Following coordination through documentary practices. In: *Proceedings of the 2011 44th Hawaii International Conference on System Sciences*. DOI: 10.1109/HICSS.2011.455.

Gitelman L (ed.) (2013) *"Raw Data" Is an Oxymoron*. Infrastructures. Cambridge, Mass.: The MIT press. ISBN 978-0-262-51828-4.

Glouftsios G (2018) Governing circulation through technology within EU border security practice-networks. *Mobilities* 13(2): 185–199. DOI: 10.1080/17450101.2017.1403774.

Glouftsios G (2019) Designing digital borders: The Visa Information System (VIS). In: Hoijtink M and Leese M (eds.) *Technology and Agency in International Relations*. London; New York: Routledge. ISBN 978-0-429-87175-7, pp. 164–187.

Glouftsios G (2021) Governing border security infrastructures: Maintaining large-scale information systems. *Security Dialogue* 52(5): 452–470. DOI: 10.1177/0967010620957230.

Glouftsios G and Leese M (2023) Epistemic fusion: Passenger Information Units and the making of international security. *Review of International Studies* 49(1): 125–142. DOI: 10.1017/S0260210522000365.

Grijpink J (1997) Chain-computerisation for interorganisational public policy implementation. *Information Infrastructure and Policy* 6(2): 81–93. URL https://content.iospress.com/articles/information-infrastructure-and-policy/iip082.

Grommé F and Ruppert E (2021) Imagining citizens as more than data subjects: A methodography of a collaborative design workshop on co-producing official statistics. *Science & Technology Studies* 34(3): 103–124. DOI: 10.23987/sts.89444.

Gusterson H (1996) *Nuclear Rites. a Weapons Laboratory at the End of the Cold War*. Berkeley and Los Angeles, California: University of California Press. ISBN 0-520-2 1373-4.

Gusterson H (1997) Studying up revisited. *PoLAR: Political and Legal Anthropology Review* 20: 114–119. URL https://heinonline.org/HOL/Page?handle=hein.journals/polar20&id=122&div=&collection=.

Hand M, Shove E and Southerton D (2005) Explaining showering: A discussion of the material, conventional, and temporal dimensions of practice. *Sociological Research Online* 10(2): 1–13. DOI: 10.5153/sro.1100.

Hanseth O (2001) Gateways — just as important as standards: How the internet won the "religious war" over standards in Scandinavia. *Knowledge, Technology & Policy* 14(3): 71–89. DOI: 10.1007/s12130-001-1017-2.

Hanseth O, Jacucci E, Grisot M and Aanestad M (2006) Reflexive standardization: Side effects and complexity in standard making. *MIS Quarterly* 30: 563–581. DOI: 10.2307/25148773.

Hanseth O and Monteiro E (1997) Inscribing behaviour in information infrastructure standards. *Accounting, Management and Information Technologies* 7(4): 183–211. DOI: 10.1016/S0959-8022(97)00008-8.

Hanseth O, Monteiro E and Hatling M (1996) Developing information infrastructure: The tension between standardization and flexibility. *Science, Technology, & Human Values* 21(4): 407–426. DOI: 10.1177/016224399602100402.

Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML and Goldstein H (2017) Challenges in administrative data linkage for research. *Big Data & Society* 4(2): 1–12. DOI: 10.1177/2053951717745678.

Higgs E (2013) Consuming identity and consuming the state in Britain since c.1750. In: About I, Brown J and Lonergan G (eds.) *Identification and Registration Practices in Transnational Perspective*, St Antony's Series. London: Palgrave Macmillan. ISBN 978-1-349-34643-1, pp. 164–182. DOI: 10.1057/9781137367310.

Hine C (2006) Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36(2): 269–298. DOI: 10.1177/0306312706054047.

Hobbing P (2010) Tracing terrorists: The European Union–Canada Agreement on Passenger Name Record (PNR) matters. In: Salter MB (ed.) *Mapping Transatlantic Security Relations*. London: Routledge. ISBN 978-0-203-85067-1.

Hoijtink M and Leese M (eds.) (2019) *Technology and Agency in International Relations*. London; New York: Routledge. ISBN 978-0-429-87175-7.

Hu Y and Nöllenburg M (2019) Graph visualization. In: Sakr S and Zomaya AY (eds.) *Encyclopedia of Big Data Technologies*. Cham: Springer International Publishing. ISBN 978-3-319-77525-8, pp. 904–912. DOI: 10.1007/978-3-319-77525-8_324.

Hughes TP (1983a) Chapter 4: Reverse salients and critical problems. In: *Networks of Power: Electrification in Western Society, 1880-1930*. Baltimore: The Johns Hopkins University Press. ISBN 978-0-8018-4614-4, pp. 79–105.

Hughes TP (1983b) *Networks of Power: Electrification in Western Society, 1880-1930*. Baltimore: The Johns Hopkins University Press. ISBN 978-0-8018-4614-4.

Hui A, Schatzki T and Shove E (2016) *The Nexus of Practices: Connections, Constellations, Practitioners*. London & New York: Routledge. ISBN 978-1-317-19939-7. DOI: 10.4324/9781315560816.

Hull MS (2012) Documents and bureaucracy. *Annual Review of Anthropology* 41(1): 251–267. DOI: 10.1146/annurev.anthro.012809.104953.

Huysmans J (1998) Security! What do you mean?: From concept to thick signifier. *European Journal of International Relations* 4(2): 226–255. DOI: 10.1177/1354066198004002004.

Huysmans J (2000) The European Union and the securitization of migration. *JCMS: Journal of Common Market Studies* 38(5): 751–777. DOI: 10.1111/1468-5965.00263.

Hyysalo S, Jensen TE and Oudshoorn N (eds.) (2016) *The New Production of Users: Changing Innovation Collectives and Involvement Strategies*. Number 42 in Routledge Studies in Innovation, Organization and Technology. New York: Routledge. ISBN 978-1-138-12456-1.

Hyysalo S, Pollock N and Williams RA (2019) Method matters in the social study of technology: Investigating the biographies of artifacts and practices. *Science & Technology Studies* 32(3): 2–25. DOI: 10.23987/sts.65532.

ICTU (2015) Informatievoorziening vreemdelingenketen. URL https://web.archive.org/web/20190525093722/https://www.ictu.nl/publicaties/informatievoorziening-vreemdelingenketen.

Iliadis A (2018) Algorithms, ontology, and social progress. *Global Media and Communication* 14(2): 219–230. DOI: 10.1177/1742766518776688.

Iliadis A and Russo F (2016) Critical data studies: An introduction. *Big Data & Society* 3(2): 1–7. DOI: 10.1177/2053951716674238.

Isin EF (2013) Claiming European Citizenship. In: Isin EF and Saward M (eds.) *Enacting European Citizenship*. Cambridge: Cambridge University Press, pp. 19–46.

ISO/IEC (1994) 7498-1:1994 Open Systems Interconnection — Basic Reference Model: The Basic Model. URL https://www.iso.org/standard/20269.html.

ISO/IEC (2008) Identification cards — Machine readable travel documents — Part 1: Machine readable passport.

Jackson MH, Poole MS and Kuhn T (2002) The social construction of technology in studies of the workplace. In: Lievevrouw LA and Livingstone S (eds.) *Handbook of New Media: Social Shaping and Consequences of ICTs*. London: SAGE Publications Ltd, pp. 236–253. DOI: 10.4135/9781848608245.n18.

Jeandesboz J (2016) Smartening border security in the European Union: An associational inquiry. *Security Dialogue* 47(4): 292–309. DOI: 10.1177/0967010616650226.

Jeandesboz J (2020) Final report on entry. Technical Report AdMiGov Deliverable D.1.4, Université libre de Bruxelles, Brussels. URL http://web.archive.org/web/20221202051249/https://admigov.eu/upload/Deliverable_D14_Jeandesboz_Final_Report_on_Entry.pdf.

Jones C, Valdivia A and Kilpatrick J (2022) Funds for fortress europe: Spending by Frontex and EU-LISA. URL http://web.archive.org/web/20220812111011/https://www.statewatch.org/analyses/2022/funds-for-fortress-europe-spending-by-frontex-and-eu-lisa/.

Jutte DP, Roos LL and Brownell MD (2011) Administrative record linkage as a tool for public health research. *Annual Review of Public Health* 32(1): 91–108. DOI: 10.1146/annurev-publhealth-031210-100700.

Kangas A (2019) UMF3+ Technological solution for better access to MS data held at Europol. URL http://web.archive.org/web/20230920101049/https://www.europarl.europa.eu/committees/it/fifth-meeting-of-the-joint-parliamentary/product-details/20190911EOT03961.

Karasti H, Baker KS and Millerand F (2010) Infrastructure time: Long-term matters in collaborative development. *Computer Supported Cooperative Work (CSCW)* 19(3): 377–415. DOI: 10.1007/s10606-010-9113-z.

Karasti H and Blomberg J (2018) Studying infrastructuring ethnographically. *Computer Supported Cooperative Work (CSCW)* 27(2): 233–265. DOI: 10.1007/s10606-017-9296-7.

Karasti H, Millerand F, Hine CM and Bowker GC (2016) Knowledge infrastructures: Part I. *Science & Technology Studies* 29(1): 2–12. DOI: 10.23987/sts.55406.

Kementsietsidis A (2009) Schema matching. In: Liu L and Özsu MT (eds.) *Encyclopedia of Database Systems*. Boston, MA: Springer. ISBN 978-0-387-39940-9, pp. 2494–2497. DOI: 10.1007/978-0-387-39940-9_962.

Kim A (2022) South Korea joins EU's digital COVID-19 certificate system. URL http://web.archive.org/web/20220817052820/http://www.koreaherald.com/view.php?ud=20220701000611.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE Publications Ltd. ISBN 978-1-4739-0947-2. DOI: 10.4135/9781473909472.

Kitchin R and Dodge M (2011) *Code/Space: Software and Everyday Life*. Software Studies. Cambridge, Mass.: The MIT Press. ISBN 978-0-262-04248-2. DOI: 10.7551/mitpress/9780262042482.001.0001.

Kitchin R and Lauriault TP (2018) Towards critical data studies: Charting and unpacking data assemblages and their work. In: Thatcher J, Eckert J and Shears A (eds.) *Thinking Big Data in Geography: New Regimes, New Research*. Lincoln and London: University of Nebraska Press. ISBN 978-1-4962-0537-7, pp. 3–20. DOI: 10.2307/j.ctt21h4z6m.

Klein HK and Kleinman DL (2002) The social construction of technology: Structural considerations. *Science, Technology, & Human Values* 27(1): 28–52. DOI: 10.1177/016224390202700102.

Kloppenburg S and van der Ploeg I (2020) Securing identities: Biometric technologies and the enactment of human bodily differences. *Science as Culture* 29(1): 57–76. DOI: 10.1080/09505431.2018.1519534.

KPMG IT Advisory (2011) Audit INDiGO: "Willen, kunnen en doen". Audit, The Hague, The Netherlands.

Kuster B and Tsianos VS (2016) How to liquefy a body on the move: Eurodac and the making of the European digital border. In: Bossong R and Carrapico H (eds.) *EU Borders and Shifting Internal Security: Technology, Externalization and Accountability*. Cham: Springer International Publishing. ISBN 978-3-319-17560-7, pp. 45–63. DOI: 10.1007/978-3-319-17560-7_3.

Lampland M and Star SL (eds.) (2009) *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Ithaca: Cornell University Press. ISBN 978-0-8014-7461-3.

Landsbergen D (2004) Screen level bureaucracy: Databases as public records. *Government Information Quarterly* 21(1): 24–50. DOI: 10.1016/j.giq.2003.12.009.

Larkin B (2013) The politics and poetics of infrastructure. *Annual Review of Anthropology* 42(1): 327–343. DOI: 10.1146/annurev-anthro-092412-155522.

Latour B (1986) Visualisation and cognition: Drawing things together. *Knowledge and society studies in the sociology of culture past and present* 6: 1–40. URL http://web.archive.org/web/20230325193816/http://www.bruno-latour.fr/node/293.

Latour B (1992) Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker WE and Law J (eds.) *Shaping Technology/Building Society: Studies in Sociotechnical Change*. Cambridge, Mass.: The MIT Press, pp. 225–258.

Latour B (1999) On recalling ANT. *The Sociological Review* 47(1_suppl): 15–25. DOI: 10.1111/j.1467-954X.1999.tb03480.x.

Latour B (2002) Gabriel Tarde and the end of the social. In: Joyce P (ed.) *The Social in Question: New Bearings in History and the Social Sciences*. London: Routledge. ISBN 0-203-99453-1, pp. 117–132.

Latour B (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Clarendon Lectures in Management Studies. Oxford & New York: Oxford University Press. ISBN 978-0-19-925604-4.

Latour B and Woolgar S (1986) *Laboratory Life: The Construction of Scientific Facts*. Second edition. Princeton, N.J.: Princeton University Press. ISBN 978-1-4008-2041-2.

Lauriault TP (2017) Ontologizing the city. In: Kitchin R, Lauriault TP and McArdle G (eds.) *Data and the City*. London: Routledge. ISBN 978-1-315-40738-8, pp. 171–186. DOI: 10.4324/9781315407388.

Law J (2006) Traduction / trahison: Notes on ANT. *Convergencia Revista de Ciencias Sociales* (42): 32–57. URL https://convergencia.uaemex.mx/article/view/1394.

Law J and Urry J (2004) Enacting the social. *Economy and Society* 33(3): 390–410. DOI: 10.1080/03085140420002257 16.

Lee ML, Clymer R and Peters K (2016) A naturalistic patient matching algorithm: Derivation and validation. *Health Informatics Journal* 22(4): 1030–1044. DOI: 10.1177/1460458215607080.

Leese M (2018) Standardizing security: The business case politics of borders. *Mobilities* 13(2): 261–275. DOI: 10.1080/17450101.2017.1403777.

Leese M (2022) Fixing state vision: Interoperability, biometrics, and identity management in the EU. *Geopolitics* 27(1): 113–133. DOI: 10.1080/14650045.2020.1830764.

Lemberg-Pedersen M, Rübner Hansen J and Halpern OJ (2020) The political economy of entry governance. Technical Report Advancing Alternative Migration (ADMIGOV) Deliverable 1.3, Aalborg University, Copenhagen. URL http://web.archive.org/web/20230705132811/https://admigov.eu/upload/Deliverable_D13_Lemberg-Pedersen_The_Political_Economy_of_Entry_Governance.pdf.

Leszczynski A and Zook M (2020) Viral data. *Big Data & Society* 7(2). DOI: 10.1177/2053951720971009.

Lipsky M (2010) *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. 30th anniversary expanded edition. New York: Russell Sage Foundation. ISBN 978-0-87154-544-2.

Loukissas YA (2019) *All Data Are Local: Thinking Critically in a Data-Driven Society*. Cambridge, Mass.: The MIT Press. ISBN 978-0-262-03966-6.

Lyon D (2003) *Surveillance after September 11*. Malden, Mass.: Polity. ISBN 978-0-7456-3181-3.

Lyon D (2009) *Identifying Citizens: ID Cards as Surveillance*. Cambridge, UK; Malden, MA: Polity. ISBN 978-0-7456-4155-3.

MacKenzie D and Wajcman J (eds.) (1999) *The Social Shaping of Technology*. Philadelphia: Open University Press. ISBN 978-0-335-19914-3.

MacKenzie DA and Wajcman J (1985) *The Social Shaping of Technology: How the Refrigerator Got Its Hum*. Milton Keynes: Open University Press. ISBN 978-0-335-15026-7.

Marcus GE (1995) Ethnography in/of the world system: The emergence of multi-sited ethnography. *Annual Review of Anthropology* 24(1): 95–117. DOI: 10.1146/annurev.an.24.100195.000523.

Marres N (2007) The issues deserve more credit: Pragmatist contributions to the study of public involvement in controversy. *Social Studies of Science* 37(5): 759–780. DOI: 10.1177/0306312706077367.

Marres N (2017) *Digital Sociology the Reinvention of Social Research*. Malden, MA: Polity. ISBN 978-0-7456-8482-6.

Marres N and Gerlitz C (2016) Interface methods: Renegotiating relations between digital social research, STS and sociology. *The Sociological Review* 64(1): 21–46. DOI: 10.1111/1467-954X.12314.

McCoy AB, Wright A, Kahn MG, Shapiro JS, Bernstam EV and Sittig DF (2013) Matching identifiers in electronic health records: Implications for duplicate records and patient safety. *BMJ Quality & Safety* 22(3): 219–224. DOI: 10.1136/bmjqs-2012-001419.

McDonald M (2008) Securitization and the construction of security. *European Journal of International Relations* 14(4): 563–587. DOI: 10.1177/1354066108097553.

M'charek A, Schramm K and Skinner D (2014) Topologies of race: Doing territory, population and identity in Europe. *Science, Technology, & Human Values* 39(4): 468–487. DOI: 10.1177/0162243913509493.

Meershoek A, Krumeich A and Vos R (2011) The construction of ethnic differences in work incapacity risks: Analysing ordering practices of physicians in the Netherlands. *Social Science & Medicine* 72(1): 15–22. DOI: 10.1016/j.socscimed.2010.10.022.

Mezzadra S and Neilson B (2013) *Border as Method, or, the Multiplication of Labor*. Durham: Duke University Press. ISBN 978-0-8223-5487-1.

Miller KJ, Richerson ES, McLeod S, Finley J and Schein A (2012-05-23/2012-05-25) International multicultural name matching competition: Design, execution, results, and lessons learned. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 3111–3117. URL http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Ministerie van Justitie en Veiligheid (2022) Protocol identificatie en labeling: gestandaardiseerde werkwijze voor de unieke identificatie en registratie in de migratieketen. Richtlijn Versie 12.1, Ministerie van Algemene Zaken. URL http://web.archive.org/web/20230906100658/https://open.overheid.nl/documenten/ronl-041514e0-ffef-448f-acfe-626fd7c8160e/pdf.

Ministerie van Justitie en Veiligheid (2023) Thema's en Architectuurprincipes - MIRA-Online. URL http://web.archive.org/web/20230802095011/https://www.miraonline.nl/index.php/Thema%27s.

Monahan T and Palmer NA (2009) The emerging politics of DHS fusion centers. *Security Dialogue* 40(6): 617–636. DOI: 10.1177/0967010609350314.

Monteiro E, Pollock N, Hanseth O and Williams R (2013) From artefacts to infrastructures. *Computer Supported Cooperative Work (CSCW)* 22(4): 575–607. DOI: 10.1007/s10606-012-9167-1.

Nader L (1972) Up the anthropologist: Perspectives gained from studying up. In: Hymes D (ed.) *Reinventing Anthropology*. New York: Pantheon Books, pp. 284–311. URL https://eric.ed.gov/?id=ED065375.

Nader L (1980) The vertical slice: Hierarchies and children. In: Britan GM and Cohen R (eds.) *Hierarchy & Society: Anthropological Perspectives on Bureaucracy*. Philadelphia: Institute for the Study of Human Issues, pp. 31–44. URL https://archive.org/details/hierarchysociety00brit_0/mode/2up.

Newcombe HB and Kennedy JM (1962) Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the ACM* 5(11): 563–566. DOI: 10.1145/368996.369026.

Newcombe HB, Kennedy JM, Axford SJ and James AP (1959) Automatic linkage of vital records. *Sci-*

ence 130(3381): 954–959. DOI: 10.1126/science.130.3381.954.

Newman M (2018) *Networks*. Oxford & New York: Oxford University Press. ISBN 978-0-19-880509-0. DOI: 10.1093/oso/9780198805090.001.0001.

Nicolini D (2016) Is small the only beautiful? Making sense of 'large phenomena' from a practice-based perspective. In: Hui A, Schatzki T and Shove E (eds.) *The Nexus of Practices Connections: Constellations, Practitioners*. London & New York: Routledge. ISBN 978-1-138-67515-5, pp. 98–113. DOI: 10.4324/9781315560816.

Olivieri L (2023) *Temporalities of Migration. Time, Data Infrastructures and Intervention*. Padova: Padova University Press. ISBN 978-88-6938-368-7.

Olwig KF, Grünenberg K, Møhl P and Simonsen A (2019) *The Biometric Border World: Technologies, Bodies and Identities on the Move*. London: Routledge. ISBN 978-0-367-80846-4. DOI: 10.4324/9780367808464.

Oosterbaan T (2012) *Architectuur Als Agenda: Een Theoretische En Empirische Analyse van de Rol van Frames Bij Architectuurontwikkeling Voor Keteninformatisering*. PhD Thesis, Erasmus University Rotterdam. URL http://hdl.handle.net/1765/31677.

Oudshoorn N (2012) How places matter: Telecare technologies and the changing spatial dimensions of healthcare. *Social Studies of Science* 42(1): 121–142. DOI: 10.1177/0306312711431817.

Pallitro R and Heyman J (2008) Theorizing cross-border mobility: Surveillance, security and identity. *Surveillance & Society* 5(3). DOI: 10.24908/ss.v5i3.3426.

Parkin J (2011) *The Difficult Road to the Schengen Information System II: The Legacy of 'laboratories' and the Cost for Fundamental Rights and the Rule of Law*. Brussels, Belgium: Centre for European Policy Studies (CEPS). ISBN 978-94-6138-088-3. URL https://www.ceps.eu/system/files/book/2011/06/INEX_PB_No_13_Parkin%20on%20SIS.pdf.

PC (2017/2023) Processing Citizenship: Digital registration of migrants as co-production of citizens, territory and Europe. DOI: 10.3030/714463.

Pelizza A (2010) From community to text and back: On semiotics and ANT as text-based methods for fleeting objects of study. *Tecnoscienza : Italian journal of science & technology studies* 1(2): 57–89. URL https://research.utwente.nl/en/publications/from-community-to-text-and-back-on-semiotics-and-ant-as-text-base.

Pelizza A (2016a) Developing the vectorial glance: Infrastructural inversion for the new agenda on governmental information systems. *Science, Technology and Human Values* 41(2): 298–321. DOI: 10.1177/0162243915597478.

Pelizza A (2016b) Disciplining change, displacing frictions: Two structural dimensions of digital circulation across land registry database integration. *Tecnoscienza. Italian Journal of Science & Technology Studies* 7(2): 35–60.

Pelizza A (2019) Processing alterity, enacting Europe: Migrant registration and identification as co-construction of individuals and polities. *Science, technology, & human values* 45(2): 262–288. DOI: 10.1177/0162243919827927.

Pelizza A (2021) Identification as translation: The art of choosing the right spokespersons at the securitized border. *Social Studies of Science* 51(4): 487–511. DOI: 10.1177/0306312720983932.

Pelizza A and Aradau C (2024) Scripts of security: Between contingency and obduracy. *Science, Technology, & Human Values* 0(0). DOI: 10.1177/01622439241258822.

Pelizza A and Hoppe R (2018) Birth of a failure: Consequences of framing ICT projects for the centralization of inter-departmental relations. *Administration & Society* 50(1): 101–130. DOI: 10.1177/0095399715598343.

Pelizza A and Loschi C (2023) Telling 'more complex stories' of European integration: How a sociotechnical perspective can help explain administrative continuity in the Common European Asylum System. *Journal of European Public Policy* : 1–22DOI: 10.1080/13501763.2023.2197945.

Pelizza A and Van Rossem W (2021) Sensing European alterity: An analogy between sensors and hotspots in transnational security networks. In: Klimburg-Witjes N, Pöchhacker N and Bowker GC (eds.) *Sensing in/Security: Sensors as Transnational Security Infrastructures*. Manchester, UK: Mattering Press. ISBN 978-1-912729-10-4, pp. 262–286. DOI: 10.28938/9781912729111.

Pelizza A and Van Rossem W (2023) Scripts of alterity: Mapping assumptions and limitations of the border security apparatus through classification schemas. *Science, Technology, & Human Values* 0(0): 1–33. DOI: 10.1177/01622439231195955.

Peoples C and Vaughan-Williams N (2021) *Critical Security Studies: An Introduction*. Third edition. Abingdon, Oxon; New York, NY: Routledge, Taylor & Francis Group. ISBN 978-0-429-27479-4.

Pinch T (2008) Technology and institutions: Living in a material world. *Theory and Society* 37(5): 461–483. DOI: 10.1007/s11186-008-9069-x.

Pinch T and Bijker WE (1984) The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science* 14(3): 399–441. DOI: 10.1177/030631284014003004.

Pollock N and Williams R (2009) *Software and Organisations: The Biography of the Enterprise-Wide System or How SAP Conquered the World*. Number 5 in Routledge Studies in Technology, Work and Organisations. London; New York: Routledge. ISBN 978-0-203-89194-0.

Pollock N and Williams R (2010) E-Infrastructures: How do we know and understand them? Strategic ethnography and the biography of artefacts. *Computer Supported Cooperative Work (CSCW)* 19(6): 521–556. DOI: 10.1007/s10606-010-9129-4.

Pollock N, Williams R and D'Adderio L (2016) Generification as a strategy how software producers configure products, manage user communities and segment markets. In: Hyysalo S, Jensen TE and Oudshoorn N (eds.) *The New Production of Users: Changing Innovation Collectives and Involvement Strategies*, number 42 in Routledge Studies in Innovation, Organization and Technology. New York: Routledge. ISBN 978-1-138-12456-1, pp. 160–189.

Pollozek S and Passoth JH (2019) Infrastructuring European migration and border control: The logistics of registration and identification at Moria hotspot. *Environment and Planning D: Society and Space* 37(4): 606–624. DOI: 10.1177/0263775819835819.

Priest D (2005) Wrongful imprisonment: Anatomy of a CIA mistake. *Washington Post* URL http://web.archive.org/web/20230417081601/https://www.washingtonpost.com/archive/politics/2005/12/04/wrongful-imprisonment-anatomy-of-a-cia-mistake/939bc95a-4031-4f83-a91

6-aaacc9acc8e7/.

PRNewswire (2011) WCC wins top tier vendor position at MITRE multi-cultural name matching challenge. *PR Newswire* URL https://web.archive.org/web/20111017050549/http://www.prnewswire.com/news-releases/wcc-wins-top-tier-vendor-position-at-mitre-multi-cultural-name-matching-challenge-131213309.html.

Quintel TA (2018) Connecting personal data of third country nationals: Interoperability of EU databases in the light of the CJEU's case law on data retention. URL http://hdl.handle.net/10993/35318.

Reckwitz A (2002) Toward a theory of social practices. *European Journal of Social Theory* 5(2): 243–263. DOI: 10.1177/13684310222225432.

Ribes D (2019) Materiality methodology, and some tricks of the trade in the study of data and specimens. In: Vertesi J and Ribes D (eds.) *DigitalSTS: A Field Guide for Science & Technology Studies*. Princeton, N.J.; Oxford: Princeton University Press. ISBN 978-0-691-19060-0, pp. 43–60. DOI: 10.1515/9780691190600.

Ribes D and Finholt TA (2009) The long now of infrastructure: Articulating tensions in development. *Journal of the Association for Information Systems* 10(5): 375–398. DOI: 10.17705/1jais.00199.

Riles A (ed.) (2006) *Documents: Artifacts of Modern Knowledge*. Ann Arbor: The University of Michigan Press. ISBN 978-0-472-09945-0.

Rinkinen J, Shove E and Smits M (2019) Cold chains in Hanoi and Bangkok: Changing systems of provision and practice. *Journal of Consumer Culture* 19(3): 379–397. DOI: 10.1177/1469540517717783.

Rippen R (2006) Sterke positie in HR en identity matching. *Database Magazine* 8: 36–38. URL http://web.archive.org/web/20230906103716/https://biplatform.nl/magazines/Aveq/111773.pdf.

Rogers R (2013) *Digital Methods*. Cambridge, MA: The MIT Press. ISBN 978-0-262-01883-8. DOI: 10.7551/mitpress/8718.001.0001.

Rogers R, Sánchez-Querubín N and Kil A (2015) *Issue Mapping for an Ageing Europe*. Amsterdam: Amsterdam University Press. ISBN 978-90-8964-716-0. DOI: 10.5117/9789089647160.

Rumsfeld DH (2002) Press conference by former US Secretary of Defence. URL http://web.archive.org/web/20220922103427/https://www.nato.int/docu/speech/2002/s020606g.htm.

Ruppert E (2013) Not just another database: The transactions that enact young offenders. *Computational culture* (3): 1–13. URL http://computationalculture.net/not-just-another-database-the-transactions-that-enact-young-offenders/.

Ruppert E (2014) Category. In: Lury C and Wakeford N (eds.) *Inventive Methods: The Happening of the Social*. Oxon, UK: Routledge. ISBN 978-0-415-72110-3, pp. 48–60.

Ruppert E, Isin E and Bigo D (2017) Data politics. *Big Data & Society* 4(2): 1–7. DOI: 10.1177/2053951717717749.

Ruppert E and Scheel S (eds.) (2021) *Data Practices: Making up a European People*. London: Goldsmiths Press. ISBN 978-1-912685-86-8.

Salter MB (2013) To make move and let stop: Mobility and the assemblage of circulation. *Mobilities*

8(1): 7–19. DOI: 10.1080/17450101.2012.747779.

Sauleau EA, Paumier JP and Buemi A (2005) Medical record linkage in health information systems by approximate string matching and clustering. *BMC Medical Informatics and Decision Making* 5(1): 32. DOI: 10.1186/1472-6947-5-32.

Schatzki T (2005) Introduction: Practice theory. In: Schatzki T, Knorr Cetina K and von Savigny E (eds.) *The Practice Turn in Contemporary Theory*. London; New York: Routledge. ISBN 978-0-203-97745-3, pp. 10–23.

Schatzki T (2010) Materiality and social life. *Nature and Culture* 5(2): 123–149. DOI: 10.3167/nc.2010.050202.

Scheel S (2019) *Autonomy of Migration? Appropriating Mobility within Biometric Border Regimes*. Abingdon, Oxon; New York, NY: Routledge. ISBN 978-1-315-26903-0.

Scheel S, Ruppert E and Ustek-Spilda F (2019) Enacting migration through data practices. *Environment and planning D: society and space* 37(4): 579–588. DOI: 10.1177/0263775819865791.

Scheers ML (2021) Biographic matching & UMF standards for EU interoperability. URL https://web.archive.org/web/20230211102336/https://www.wcc-group.com/company/post/2021/02/22/high-quality-biographic-matching-umf-standards-for-eu-interoperability/.

Schmitt E and Schmidt MS (2013) 2 U.S. agencies added Boston bomb suspect to watch lists. *The New York Times* URL https://web.archive.org/web/20210513103658/https://www.nytimes.com/2013/04/25/us/tamerlan-tsarnaev-bomb-suspect-was-on-watch-lists.html.

Scott JC (1998) *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven; London: Yale University Press. ISBN 978-0-300-07815-2.

Sengoopta C (2004) *Imprint of the Raj: How Fingerprinting Was Born in Colonial India*. London: Pan. ISBN 978-0-330-49140-2.

Shankar K, Hakken D and Østerlund C (2017) Rethinking documents. In: Felt U, Fouché R, Miller CA and Smith-Doerr L (eds.) *The Handbook of Science and Technology Studies*, fourth edition. Cambridge, Mass.; London, Eng.: The MIT Press. ISBN 978-0-262-03568-2, pp. 59–85.

Shore C, Wright S and Però D (eds.) (2011) *Policy Worlds: Anthropology and the Analysis of Contemporary Power*. EASA Series. New York: Berghahn Books. ISBN 978-0-85745-116-3.

Shove E (2016) Matters of practice. In: Hui A, Schatzki T and Shove E (eds.) *The Nexus of Practices: Connections, Constellations, Practitioners*. London & New York: Routledge. ISBN 978-1-138-67515-5, pp. 155–168. DOI: 10.4324/9781315560816.

Shove E, Pantzar M and Watson M (2012) *The Dynamics of Social Practice: Everyday Life and How It Changes*. Thousand Oaks, CA: SAGE Publications Ltd. ISBN 978-0-85702-042-0.

Shove E, Watson M and Spurling N (2015) Conceptualizing connections: Energy demand, infrastructures and social practices. *European Journal of Social Theory* 18(3): 274–287. DOI: 10.1177/1368431015579964.

Silvast A and Virtanen MJ (2023) On theory-methods packages in Science and Technology Studies. *Science, Technology, & Human Values* 48(1): 167–189. DOI: 10.1177/01622439211040241.

Skinner D (2018) Race, racism and identification in the era of technosecurity. *Science as Culture* 29(1):

77–99. DOI: 10.1080/09505431.2018.1523887.

Snellen I (2002) Electronic governance: Implications for citizens, politicians and public servants. *International Review of Administrative Sciences* 68(2): 183–198. DOI: 10.1177/0020852302682002.

Soysüren I and Nedelcu M (2022) European instruments for the deportation of foreigners and their uses by France and Switzerland: The application of the Dublin III Regulation and Eurodac. *Journal of Ethnic and Migration Studies* 48(8): 1927–1943. DOI: 10.1080/1369183X.2020.1796278.

Sparke MB (2006) A neoliberal nexus: Economy, security and the biopolitics of citizenship on the border. *Political Geography* 25(2): 151–180. DOI: 10.1016/j.polgeo.2005.10.002.

Squire V (ed.) (2010) *The Contested Politics of Mobility: Borderzones and Irregularity*. London: Routledge. ISBN 978-0-203-83982-9. DOI: 10.4324/9780203839829.

Star SL (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3): 377–391. DOI: 10.1177/00027649921955326.

Star SL and Ruhleder K (1996) Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1): 111–134. DOI: 10.1287/isre.7.1.111.

Steinbock DJ (2005/2006) Data matching, data mining, and due process. *Georgia Law Review* 40: 1. URL https://heinonline.org/HOL/Page?handle=hein.journals/geolr40&id=21&div=&collection=.

Strange M, Squire V and Lundberg A (2017) Irregular migration struggles and active subjects of trans-border politics: New research strategies for interrogating the agency of the marginalised. *Politics* 37(3): 243–253. DOI: 10.1177/0263395717715856.

Stumpf J (2006) The crimmigration crisis: Immigrants, crime, and sovereign power. *American University Law Review* 56(2): 367–419.

Suchman L (1993) Do categories have politics? *Computer Supported Cooperative Work (CSCW)* 2(3): 177–190. DOI: 10.1007/BF00749015.

Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. Learning in Doing: Social, Cognitive and Computational Perspectives, second edition edition. Cambridge: Cambridge University Press. ISBN 978-0-521-85891-5. DOI: 10.1017/CBO9780511808418.

Suchman L (2020) Algorithmic warfare and the reinvention of accuracy. *Critical Studies on Security* 8(2): 175–187. DOI: 10.1080/21624887.2020.1760587.

Suchman L, Follis K and Weber J (2017) Tracking and targeting: Sociotechnologies of (in)security. *Science, Technology, & Human Values* 42(6): 983–1002. DOI: 10.1177/0162243917731524.

Sweeney S (2008) The ambiguous origins of the archival principle of "provenance". *Libraries & the Cultural Record* 43(2): 193–213. URL https://www.jstor.org/stable/25549475.

SWIFT (2018) Simplify the complex world of sanctions screening: What do screening activities cover? Technical report. URL http://web.archive.org/web/20220226233517/https://www.swift.com/news-events/news/helping-simplify-complex-world-sanctions-screening.

SWIFT (2021) Name Screening: Fulfil your customer due diligence, maintain accurate customer risk profiles and mitigate business and reputational risks. Factsheet. URL http://web.archive.org/web/20230415104752/https://www.swift.com/swift-resource/250436/download.

Talburt JR (2013) Special issue on entity resolution overview: The criticality of entity resolution in data and information quality. *Journal of Data and Information Quality* 4(2): 6:1–6:2. DOI: 10.1145/2435221.2435222.

Taureck R (2006) Securitization theory and securitization studies. *Journal of International Relations and Development* 9(1): 53–61. DOI: 10.1057/palgrave.jird.1800072.

Thales Group (2021) DHS's Automated Biometric Identification System IDENT – the heart of biometric visitor identification in the USA. URL http://web.archive.org/web/20230509140409/https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/customer-cases/ident-automated-biometric-identification-system.

The MITRE Corporation (2011) MITRE invites technology innovators to take the challenge. URL https://web.archive.org/web/20200721132651/https://www.mitre.org/publications/project-stories/mitre-invites-technology-innovators-to-take-the-challenge.

Thomer AK and Wickett KM (2020) Relational data paradigms: What do we learn by taking the materiality of databases seriously? *Big Data & Society* 7(1): 1–16. DOI: 10.1177/2053951720934838.

Timmermans S and Epstein S (2010) A world of standards but not a standard world: Toward a sociology of standards and standardization. *Annual Review of Sociology* 36: 69–89. DOI: 10.1146/annurev.soc.012809.102629.

Toet D (2009) Indigo-systeem van IND wint architectuurprijs. *Computable.nl* URL https://web.archive.org/web/20201013115613/https://www.computable.nl/artikel/nieuws/crm/3174763/2333360/indigosysteem-van-ind-wint-architectuurprijs.html.

Torpey JC (2018) *The Invention of the Passport: Surveillance, Citizenship and the State*. Cambridge Studies in Law and Society, second edition. New York, NY; Cambridge, UK: Cambridge University Press. ISBN 978-1-108-46294-5. DOI: 10.1017/9781108664271.

Trauttmansdorff P (2022) The fabrication of a necessary policy fiction: The interoperability 'solution' for biometric borders. *Critical Policy Studies* DOI: 10.1080/19460171.2022.2147851.

Trauttmansdorff P and Felt U (2023) Between infrastructural experimentation and collective imagination: The digital transformation of the EU border regime. *Science, Technology, & Human Values* 48(3): 635–662. DOI: 10.1177/01622439211057523.

Tsianos V and Karakayali S (2010) Transnational migration and the emergence of the European border regime: An ethnographic analysis. *European Journal of Social Theory* 13(3): 373–387. DOI: 10.1177/1368431010371761.

Valdivia A, Aradau C, Blanke T and Perret S (2022) Neither opaque nor transparent: A transdisciplinary methodology to investigate datafication at the EU borders. *Big Data & Society* 9(2): 1–17. DOI: 10.1177/20539517221124586.

van der Ploeg I (1999) The illegal body: 'Eurodac' and the politics of biometric identification. *Ethics and Information Technology* 1(4): 295–302. DOI: 10.1023/A:1010064613240.

van Keulen M (2012) Managing uncertainty: The road towards better data interoperability. *IT - Information Technology* 54(3): 138–146. DOI: 10.1524/itit.2012.0674.

van Keulen M (2018) Probabilistic Data Integration. In: Sakr S and Zomaya A (eds.) *Encyclopedia of*

Big Data Technologies. Cham: Springer International Publishing. ISBN 978-3-319-63962-8, pp. 1–9. DOI: 10.1007/978-3-319-63962-8_18-1.

Van Rossem W (2021) Ontology-explorer. Zenodo. DOI: 10.5281/zenodo.4899316.

Van Rossem W and Pelizza A (2022) The Ontology Explorer: A method to make visible data infrastructures for population management. *Big Data & Society* 9(1): 1–18. DOI: 10.1177/20539517221104087.

Venturini T, Jacomy M and Jensen P (2021) What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts. *Big Data & Society* 8(1): 1–16. DOI: 10.1177/20539517211018488.

Wæver O, Buzan B and Carlton D (eds.) (1993) *Identity, Migration and the New Security Agenda in Europe*. New York: St. Martin's Press. ISBN 978-1-85567-041-9.

Walt SM (2017) Realism and security. *Oxford Research Encyclopedia of International Studies* DOI: 10.1093/acrefore/9780190846626.013.286.

Walters W (2014) Drone strikes, dingpolitik and beyond: Furthering the debate on materiality and security. *Security Dialogue* 45(2): 101–118. DOI: 10.1177/0967010613519162.

WCC (2002) Annual report 2002. Technical report, Went Computing Consultancy Group B.V., Amstelveen, The Netherlands. URL https://web.archive.org/web/20050121083505if_/http://www.wcc.nl/doc/WCCGROUP_2002.pdf.

WCC (2003) Annual report 2003. Technical report, Went Computing Consultancy Group B.V., Amstelveen, The Netherlands. URL https://web.archive.org/web/20050121093827if_/http://www.wcc.nl/doc/WCCGROUP_2003.pdf.

WCC (2005) Target markets - Other markets. URL https://web.archive.org/web/20050407133407/http://www.wcc.nl:80/page.aspx?menu=targetmarkets_004&page=targetmarkets_other&lang=en.

WCC (2009a) HSPD-24 white paper now available from WCC Smart Search & Match. URL http://web.archive.org/web/20220806035706/https://www.securityinfowatch.com/home/news/10492664/hspd24-white-paper-now-available-from-wcc-smart-search-match.

WCC (2009b) Meeting the challenges of HSDP-24: A layered approach to accurate real time identification. White paper, WCC Smart Search & Match.

WCC (2019) Partners. URL https://web.archive.org/web/20190704094310/https://wcc-group.com/partners.

WCC (2020) UMF a solution for EU interoperability. URL https://web.archive.org/web/20200713152705/https://www.wcc-group.com/company/post/2020/06/10/umf-a-solution-for-eu-interoperability/.

Weltevrede EJT (2016) *Repurposing Digital Methods: The Research Affordances of Platforms and Engines*. PhD Thesis, Universiteit van Amsterdam. URL https://hdl.handle.net/11245/1.505660.

Wendt A (1992/ed) Anarchy is what states make of it: The social construction of power politics. *International Organization* 46(2): 391–425. DOI: 10.1017/S0020818300027764.

Wigand RT (2020) Whatever happened to disintermediation? *Electronic Markets* 30(1): 39–47. DO

I: 10.1007/s12525-019-00389-0.

Williams R and Pollock N (2012) Moving beyond the single site implementation study: How (and why) we should study the biography of packaged enterprise solutions. *Information Systems Research* 23(1): 1–22. DOI: 10.1287/isre.1110.0352.

Winkler WE (2014) Matching and record linkage. *WIREs Computational Statistics* 6(5): 313–325. DO I: 10.1002/wics.1317.

Winter T (2014) Russia warned U.S. about Tsarnaev, but spelling issue let him escape. *NBC News* URL https://web.archive.org/web/20210330213852if_/https://www.nbcnews.com/storyline/boston-bombing-anniversary/russia-warned-u-s-about-tsarnaev-spelling-issue-let-him-n60836.

Woolgar S (1990) Configuring the user: The case of usability trials. *The Sociological Review* 38(S1): 58–99. DOI: 10.1111/j.1467-954X.1990.tb03349.x.

Zampagni F (2016) Unpacking the Schengen visa regime: A study on bureaucrats and discretion in an Italian consulate. *Journal of Borderlands Studies* 31(2): 251–266. DOI: 10.1080/08865655.2016.1174605.

Zech J, Husk G, Moore T and Shapiro JS (2016) Measuring the degree of unmatched patient records in a health information exchange using exact matching. *Applied Clinical Informatics* 7(2): 330–340. DOI: 10.4338/ACI-2015-11-RA-0158.

Zijderveld M, Ridderhof W and Brattinga M (2013) Basis start architectuur architectuur van de vreemdelingenketen: kennis delen, informatie gebruiken, samen doen. Technical report, Ministerie van Binnenlandse Zaken, Den Haag. URL http://web.archive.org/web/20220901074647/https://www.digitaleoverheid.nl/wp-content/uploads/sites/8/2017/01/architectuur-van-de-vreemdelingenketen.pdf.

Zuboff S (2015) Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30(1): 75–89. DOI: 10.1057/jit.2015.5.

Zureik E and Hindle K (2004) Governance, security and technology: The case of biometrics. *Studies in Political Economy* 73(1): 113–137. DOI: 10.1080/19187033.2004.11675154.

Zureik E and Salter MB (eds.) (2005) *Global Surveillance and Policing: Borders, Security, Identity*. Devon: Willan Publishing. ISBN 978-1-84392-161-5.