VRIJE UNIVERSITEIT

MISSING LINKS

Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J J.G. Geurts,
volgens besluit van de decaan
van de Faculteit der Bètawetenschappen
in het openbaar te verdedigen
op woensdag 5 november 2025 om 13.45 uur
in de universiteit,
De Boelelaan 1105

door

Go Sugimoto

geboren te Kanagawa, Japan

promotor:	prof.dr. J.R. van Ossenbruggen
copromotor:	dr. V. de Boer
promotiecommissie:	prof. J.D. Richards prof.dr. J.J. Noordegraaf dr. L. Stork dr.ir. E.J.A. Folmer dr.mr. C. Gerritsen

SIKS Dissertation Series No. 2025-54
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: xxxxx

Declaration of Authorship

I, Go Sugimoto, hereby declare that the thesis titled *"Missing Links: Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities"* and its content are the result of my own work.
I confirm that:

- This work was primarily conducted during my pursuit of a research degree at these universities.
- If any portion of this thesis has been previously submitted for an academic degree or any other qualification at these universities or any other educational institution, I have explicitly disclosed this information.
- I have consistently acknowledged the sources of published works by other authors that I consulted.
- In cases where I have included excerpts from the work of others, I have consistently provided proper source attributions. Aside from these quotations, the entirety of this thesis represents my independent effort.
- I have acknowledged all significant sources of assistance.
- In cases where this thesis draws upon collaborative efforts with others, I have provided a clear distinction between the contributions made by collaborators and my own individual contributions.

21.05.2025

CONTENTS

ACKNOWLEDGMENTS

This thesis could not have been published in this form without the support from many people. I would like to express my profound gratitude for their help and encouragement.

Amsterdam

First, I would like to thank the two supervisors of this thesis for their professionalism and commitment. Victor de Boer provided me with the research opportunity in the first place. His generosity and support are beyond words. Jacco van Ossenbruggen is an exceptional researcher in the field, and his collaboration with Victor was crucial in ensuring the research stayed on the right track. I always took his sometimes direct "Dutch" remarks as encouragement. This thesis is a reflection of their valuable insights into scientific and critical thinking.

Second, my thanks go to the PhD committee, consisting of prof. J.D. Richards, prof.dr. J.J. Noordegraaf, dr. L. Stork, dr.ir. E.J.A. Folmer, and dr.mr. C. Gerritsen. Their feedback is as valuable as the thesis itself. I deeply appreciate that they took the time to read what is, admittedly, a rather lengthy thesis and traveled all the way to Amsterdam for my defense.

Egmond aan Zee

Third, the people at VU were an integral part of my PhD journey. The User Centric Data Science group provided the most productive environment for writing this thesis. Most importantly, the talented group of researchers were the true gems of the department. We supported and inspired each other to conduct cutting-edge research in the field. I would like to thank Shuai Wang, Margherita Martorana, Sarah Shoilee, Xueli Pan, Emma Beauxis-Aussalet, Maryam Alimardani, Elena Beretta, Tobias Kuhn, Roderick van der Weerdt, Mirthe Dankloff, Dayana Spagnuelo, Lea Krause, Xander Wilcke, Andre Valdestilhas, and Maddalena Ghiotto. Special thanks to Ronald Siebes for providing technical expertise on the server setup at VU, and to Peter Stol from the IT help desk. After participating in the High-Performance Computing course that Peter organized, I continued a friendship with him. His advice contributed to the success of the NLP work in Chapter 4.

I also met fascinating people in the Computer Science department. Mojca Lovrenčak, an extremely friendly secretary, helped me navigate all the tedious administrative issues with her charming Slovenian touch. One of the department's highlights was the overnight "outje." Spending time at a Dutch seaside resort was truly refreshing and re-energizing. During the event, our team's short sci-fi film "Genesis" won the Best Film award. I developed the concept, created the storyboard, directed most of the shooting, and even acted in the film. With the help of wonderful actors and film editors from the team, it became one of the most intense and creative moments of my life.

Odense, Stuttgart, Helsinki

Fourth, I would like to thank all members of the InTaVia project. The project was closely related to my research for Chapters 2, 4, and 5. ReKisstory was inspired by many ideas from the project and was an attempt to create synergies. Angel Daza, one of the coauthors

of the paper resulting in Chapter 4, played a key role. His recommendations were often spot on, not only with regard to the use of NLP tools but also for sightseeing in Tübingen, Germany. Several InTaVia colleagues helped test the NLP annotation, including Jouni Tuominen, Rafael Leal, and Matthias Schlögl. Although this task was eventually canceled, the experience and their contributions inspired my work. Additionally, Eva Mayr and Florian Windhager, outstanding researchers from Krems, managed the project exceptionally well and provided invaluable support in tackling challenging tasks.

This thesis was partially supported by the EU Horizon 2020 project InTaVia: In/Tangible European Heritage - Visual Analysis, Curation and Communication (http://intavia.eu) under grant agreement No 101004825.

Vienna

Another factor in my PhD enrollment was the result of consultations with my former colleague Davor, a freelance software engineer from Serbia. Although our communication was mostly remote, our lively conversations often made me laugh during the COVID-19 lockdowns and inspired me to dream about what we could achieve with IT. We also co-authored a paper in Vienna in the past. Another friend from Vienna is Julius, a Lithuanian designer who provided invaluable advice on the design of ReKisstory. To a novice software developer like me, with limited experience, he was like a magician with modern CSS and JavaScript, making my app visually appealing.

Anywhere On Earth

The participants of the online focus group workshops for ReKisstory should not be forgotten. Chapter 5 is the only research in this thesis conducted with direct contact with potential end-users. I must admit that some participants were my former colleagues (including those from InTaVia) and friends. I was pleasantly surprised by their interest in my work and their continued engagement in the domain. Many also sent their apologies for being unable to participate. For the sake of anonymity, I will not name them here.

The Hague, Hamburg, Miri, Kyiv, Brussels, Barcelona, Zurich, Helsinki, Munich, Manchester, Cardiff, Mexico City

I would also like to thank all my friends in The Hague and elsewhere. During my PhD journey, they helped me relieve stress and pressure on badminton courts, football pitches, in pubs, cafes, restaurants, and even online.

Tokyo

Last, but not least, my family supported me from Japan. They had absolutely no idea what I worked on for PhD, but their continuous love was always a source of my motivation. It is time to relax with them, playing a tasting game, practicing football skills, and dipping in an Onsen in front of the Pacific Ocean again.

This thesis is the result of international support and collaboration. I literally traveled across the globe and was warmly welcomed wherever I went. My PhD defense will be one of those moments when I take pride in my international career, having met hundreds of people from various countries and backgrounds.

At the same time, I must confess that this has been one of the most challenging periods of my life. The impact of the COVID-19 pandemic affected me in many ways—physically, mentally, economically, politically, technically, and socially. I was sometimes deflated and nearly gave up on my PhD, as well as my career.

Therefore, it feels even more unbelievable that I have earned a PhD more than a decade after receiving my Master's degree from the University of York in the United Kingdom. I feel a bit like Brian May of Queen, only missing the music career and the title of Knight! I have proven to myself that it's never too late and anything is possible. Now, let me pick up my guitar and see what happens next.

Thank you all for your precious support.

Go Sugimoto
In The Hague with Stratocaster, July 2025

In memory of Kaneharu Sugimoto (1930 - 2025)

1

INTRODUCTION

The Semantic Web aims to enable both humans and machines to process and interpret data more efficiently on the web by using structured data formats. Linked Data (LD) principles were designed as a building block for this approach. They provide a means to create interconnected data across the web through standardised frameworks such as URI, HTTP, and RDF. The thesis examines the application of LD in the fields of Cultural Heritage (CH) and Digital Humanities (DH). Despite significant progress, LD quality issues have been identified in the research community. In this thesis, we investigate these issues by employing a strategic approach combining two dimensions and three stakeholders for analysis. The two dimensions are the quality of data and the quality of tools. Three stakeholders are data producers, data consumers, and developers. Through the lens of stakeholders, we analyze current LD practices, identify challenges, and propose solutions to improve data content, connectivity and tool usability in CH and DH. Our research combines quantitative and qualitative methods to gain different perspectives of quality analyses. The research is structured around four primary research questions, each corresponding to a specific chapter. In this introduction, we provide a background for these themes.

1.1 PROLOGUE

2012. In London. "This is for everybody"[1], Tim Berners-Lee stated during the opening ceremony of the Olympic Games, standing amidst hundreds of dancers and music. This moment symbolised a celebration of human invention: the creation of the World Wide Web (WWW) in 1989. Nowadays it is hard to imagine life before the Web. It seemed that Berners-Lee's dream for a new information age was fully realised by that time.

This chapter is partly based on *Sugimoto, Go. (2019), Open Data Empowerment of Digital Humanities by Wikipedia/DBpedia Gamification and Crowd Curation -WiQiZi's Challenges with APIs and SPARQL. https://doi.org/10.5281/zenodo.3465654* [82] and *Sugimoto, Go (2018), Who is open data for and why could it be hard to use it in the digital humanities? Federated application programming interfaces for interdisciplinary research. In: International Journal of Metadata, Semantics and Ontologies (IJMSO) 12, p. 204-218. doi:10.1504/IJMSO.2017.10014806* [123].

[1] https://webfoundation.org/2012/07/sir-tim-berners-lee-closes-out-2012-olympic-opening-ceremony-this-is-for-everyone-one-web/ Accessed 2024-09-16

However, in the 1990s—and even today—Berners-Lee's more ambitious vision [47] was not fully understood by many. It is known as the Semantic Web-—also referred to as "Web 3.0" or the "Web of Data" [47]. At that time, most ordinary users of the Web were just beginning to grasp the basic concepts of the technology and its functionality. They simply enjoyed publishing websites to share information and communicate with millions of people around the world within seconds ("Web 1.0"). The WWW is truly a revolution and a turning point in human history, solving some critical problems of information management. But the success of Web 1.0 was certainly not the end of the evolution of information technology.

The main issue with Web 1.0 (and later "Web 2.0", which focused on user-generated content via social media and similar platforms) was that it was not designed for efficient machine consumption, where smarter automation could take place [90]. As the volume of information continued to grow beyond human capacity, automation became increasingly essential. The information explosion was imminent, but the world was not yet ready for it. Berners-Lee anticipated this challenge and envisioned the development of the (Semantic) Web. Although the concept of the Semantic Web originates in the early stages of the Web development, it has not achieved market adoption as quickly as other advancements in Web technology [92].

Berners-Lee's original vision of the Web did not only include the Web of Documents (i.e. web pages), but also the Web of Data (or the Semantic Web). According to Hogan [90], "The core objective of the Web of Data is to publish content on the Web in formats that machines can process more easily and accurately than the human-friendly HTML documents forming the current "Web of Documents"".[2] So as early as 1998, researchers sharing Berners-Lee's vision started to explore the potential of the Semantic Web. Their efforts have led to the development of standards, languages, protocols, tools, and optimizations as steps toward the new generation of the Web [90].

This thesis explores the potential and challenges of the Semantic Web in Cultural Heritage (CH) and Digital Humanities (DH) in the current research landscape. In particular, we focus on the quality issues of the technology. Where do we stand in the evolution of the (Semantic) Web? How has it affected the CH and DH fields, and what further developments lie ahead? Is there something crucial missing from this evolution?

1.2 MISSING LINKS IN CULTURAL HERITAGE AND DIGITAL HUMANITIES

One of the key indicators for the development of the Semantic Web is the standardization effort by the World Wide Web Consortium (W3C) which works with international stakeholders to develop web standards to meet requirements for such areas as accessibility, internationalization, privacy, and security [13].

In the context of the Semantic Web, W3C recommended the first version of the Resource Description Framework (RDF) in 1999 [16]. Subsequent milestones included the introduction of the Web Ontology Language (OWL) in 2004 [14] and SPARQL Protocol and RDF Query Language (SPARQL) in 2008 [21]. In 2006, Tim Berners-Lee published the Linked Data (LD) principles [48], establishing a cornerstone for the Semantic Web's architecture

[2]This will be explained in the following sections.

(see Section 1.3.1). Initially developed by academics and researchers in laboratories, LD has since spread to companies and various governmental agencies [90].

In essence, LD makes data on the web more machine-readable by following common technical standards and principles. As a result, machines can exchange data with one another and interpret and act on that data in more sophisticated and meaningful ways [90] (See technical background of LD in Section 1.3). In this thesis, we focus on LD as a practical implementation of the Semantic Web.

The CH sector has embraced LD technologies [72]. Edelstein et al. [72] introduce six stages of a Linked Open Data (LOD)[3] life cycle to group LD projects in CH, which we will revisit in the Section 6.2.2 in Chapter 6: 1) developing datasets, 2) linking data, 3) documenting processes for reuse, 4) developing user interfaces, 5) promoting a culture of reuse, and 6) expanding the definition of cultural heritage.

In the realm of data production, digital library projects have played a crucial role for developing LD datasets [121]. Leading national libraries and digital services have developed platforms allowing remote access to previously offline collections, including books, newspapers, magazines, and multimedia resources. This digital transformation has been boosted by long-term digitization efforts in CH. The digital representations of both analogue and born-digital objects seamlessly integrate with LD-encoded metadata.

Concurrently, the Open Data movement has gained momentum [90]. Many public organizations, including museums, libraries and archives, have recognised the potential of sharing their data more freely and effectively with the public [54]. When LD adheres to Open Data principles, it is termed Linked Open Data (LOD). Specifically, LD released under an open license qualifies as LOD. As a consequence, LD, particularly in its open form, has become a catalyst for sharing valuable digital data on the web. This ethos of free and open access to knowledge aligns closely with the principles of librarianship [128]. In McKenna et al. [105], the benefits of LD for CH are listed: a)interoperability and re-usability, b) discoverability and visibility, c) interlinking and integration, and d) reliability and authority control.

In terms of standards, one of the milestones in CH was the development of CIDOC Conceptual Reference Model (CIDOC-CRM) [1]. CIDOC-CRM provides a framework for describing a wide range of CH phenomena, including complex data collections. Becoming an ISO standard in 2006, CIDOC-CRM was designed to anticipate the potential of the Semantic Web in CH documentation management. While RDF is not mandatory for implementation [70], it has been adopted in numerous LD projects, particularly in DH [94, 117].

The CH sector often refers to memory organizations, commonly known as GLAM (Galleries, Libraries, Archives, and Museums). However, these institutions are, in reality, composed of individual professionals employed by GLAM organizations. These professionals play a crucial role in the use, adaptation, and implementation of LD.

In fact, individual researchers are increasingly recognizing new opportunities to utilize CH data in digital, machine-processable formats [92]. At the intersection of computer science, humanities, and social sciences, the vast availability of digital data in the CH domain has given rise to a new research paradigm known as Digital Humanities (DH). This paradigm is exemplified by a shift in research methodology, from "close reading" to "distant reading" [108]. Jänicke et al. [99] argue that the transition from analyzing a single analogue

[3]Linked Data with Open License. See below

text (close reading) to the ability to browse and process large collections of digital texts (distant reading) is one of the foundational pillars of DH. By analyzing statistical patterns across large datasets, DH enables researchers to ask new research questions about CH data.

In recent years, there has been significant growth in the use of LD within DH research. For instance, Antopoksky [43] classified various LOD projects in DH, concluding that a methodology for creating information resources based on LD technology has emerged globally. In addition, Zhao [144] identified 195 research articles published between 2016 and 2021 within DH and interdisciplinary fields that referenced Wikidata, one of the most widely adopted and systematically studied LD resources in DH.

While LD implementations have increased, several challenges have been reported in LD research within the fields of CH and DH. This thesis explores the challenge of quality of LD within the context of DH, which we believe is one of the most pressing issues in the current research landscape.

The LD community has intensively discussed about quality for the recent years [64, 65, 115, 116, 142]. In CH, Candela et al. [54] claim that it is a challenge for data researchers to identify candidate LD sources for reuse and enrichment. By analyzing LD in libraries, they provide an overview of the quality achieved by the LOD repositories developed by digital libraries. Although the interest in this area of research has increased, studies on LOD quality for a broader cultural heritage including museums and archives remains relatively limited.

As the LD quality is determined by the LD stakeholders such as end-users and developers, we investigate it from their perspectives. What makes this challenge more complicated is that these stakeholders have different levels of awareness and understanding of LD, varying skill levels, and specific requirements related to LD. These gaps often lead to miscommunication, which limits recognition, adoption, and progress in the CH and DH sectors. For instance, in the archival domain, Hawkins [88] highlights gaps recognised by DH and CH experts. The study examines the barriers currently hampering Digital Humanists from fully leveraging the potential of LD for archives. With such barriers in mind, we investigate what the stakeholders look for, what they have been working on, and what they can do with LD.

In this thesis, we shed light on these aspects, which we refer to as the "Missing Links". This metaphor represents the gaps in LD implementations. Specifically, we explore the quality of LD from the perspectives of a) two quality dimensions and b) three types of stakeholders. We will explain these in more detail in the following sections. By analyzing existing LD implementations, we propose recommendations and solutions for the issues identified, in order to improve LD applications in these domains.

In this introduction, we provide an overarching background of the thesis, clarifying fundamental concepts, including terminology and definitions. Additionally, we outline the motivation for our research. Furthermore, we lay out our research framework, including the research questions, the structure of our research in the following chapters, and the fundamental approaches and methodologies.

1.3 LINKED DATA TECHNOLOGY

As we use the term Linked Data (LD) in all chapters, we first explain the basic technical background of this concept.

1.3.1 LINKED DATA PRINCIPLES

LD is a set of principles to create interlinking data on the web as a best practice [48]. In short, the principles include:

- Use Uniform Resource Identifiers (URIs)[49] to name things
- Use Hypertext Transfer Protocol (HTTP)[46] to look them up on the web
- Use standards such as Resource Description Framework (RDF)[16] to describe them
- Include links to other URIs to discover more things

The first principle concerns identification. A URI consists of unique characters to identify a physical or abstract resource such as an analogue photograph and a digital photograph. A URI is similar to a Uniform Resource Locator (URL) that is colloquially referred to as the address of a web page. URLs are essential for the Web to work. While URLs are designed to uniquely locate resources on the web, URIs have a broader scope to "name things" for resources on the web and outside the web. Therefore, every URL is a URI, but not vice versa. If a URI does not provide a means to locate a resource on the web, it would be called a Uniform Resource Name (URN). URIs are essential for LD. They are useful to disambiguate similar resources and ensure the uniqueness of the resource. In Chapter 3, we define a URI syntax to refer to date information.

Now, if we need to refer to the same resource on the web, no matter if it is a concept, physical object, or digital image, there are two ways. Theoretically, but rather unrealistically, all web users use the same URI to refer to the same thing, say, Angkor Wat. If this happens, we can link all resources by creating links to this particular URI, so that we use this URI to retrieve all information about Angkor Wat on the web. However, this is a very unlikely scenario. Given that the Web is a distributed system, it is much more likely that multiple users will use different URIs for the same Angkor Wat. Thus, an alternative solution is to create links between those different URIs. For instance, Angkor Wat in one LD source (DBpedia) links to Angkor Wat in another LD source (Wikidata). This is often realised through an owl:sameAs link, which is one of the central points of discussion in Chapter 2.

In LD, the use of URIs is tightly coupled with the second principle, HTTP. It is a technical protocol for machines to communicate on the web. Although HTTP is not a prerequisite for URIs, if a URI uses HTTP, we can look up the resource on the web. For example, we can access information about Angkor Wat with this HTTP-based URI: http://dbpedia.org/resource/Angko_Wat. Similarly, the URI of the corresponding entity in Wikidata is https://www.wikidata.org/wiki/Q43473. HTTP allows us to publish date entities on the web in our case study (Chapter 3).

As URIs and HTTP are also building blocks of Web 1.0, they are not entirely unique to LD[4]. Instead, the third principle, RDF, is the technical core of LD, because it allows us to add semantics to data. The Semantic Web represents this notion of adding semantics to the

[4]This also implies that LD is built upon the existing infrastructure of the web without reinventing the wheel. The excellence of Berners-Lee's grand design can be seen here in terms of the long-term architecture of the web

1

data on the web. We discuss the quality of semantics in LD in Chapter 3, Chapter 4, and Chapter 5.

RDF is similar to HTML (HyperText Markup Language), which creates most of the websites today, in that they can be used to publish and share data about "resources"[5] on the Web. However, RDF is significantly different from HTML. While the hyperlinks of HTML are essentially merely connections between HTML web pages, RDF uses typed links to connect various resources. By adding ontologies (see Section 1.3.3), they can describe what kind of relations (semantics) the resources have. In other words, RDF data can be seen as a directed graph that can be used to make statements about resources/things. An RDF statement consists of three components: subject, predicate, and object. This is called a triple.

While the subject denotes a resource, the object denotes another resource or literal (textual information related to the subject). The predicate describes the relationship between the subject and the object. For instance, the statement "Angkor Wat is located in Cambodia" can be encoded in RDF. In DBpedia, it is described below, using the RDF Turtle syntax[6]. The first two lines define syntactical shortcuts for URIs that are often too verbose. For instance, the first line allows us to use dbr: in place of a base URI http://dbpedia.org/resource/. Thus, http://dbpedia.org/resource/Angkor_Wat can be shortened as dbr:Angkor_Wat.

```
@PREFIX dbr: <http://dbpedia.org/resource/> .
@PREFIX dbp: <http://dbpedia.org/property/> .

dbr:Angkor_Wat dbp:location dbr:Cambodia
```

In this case, the resources are the physical site of Angkor Wat and the country of Cambodia. The semantics of the connection is the location of the former in the latter. This is an RDF representation of the statement in the format called Turtle, but RDF can be encoded in different serializations of the same information such as XML and JSON-LD [18]. In Chapter 4, we generate RDF statements semi-automatically from a Wikipedia article, by means of Natural Language Processing (NLP). Interesting features of RDF are: a) a subject can be used in different statements, and b) an object can become a subject of another statement. For example, in addition to the above-mentioned statement in DBpedia, we can add more statements:

```
@prefix geo:    <http://www.w3.org/2003/01/geo/wgs84_pos#> .

dbr:Angkor_Wat geo:geometry "POINT(103.86666870117␣13.41250038147)"^^ogcgs:wktLiteral
```

and

```
@PREFIX dbo: <http://dbpedia.org/ontology/> .

dbr:Cambodia dbo:language       dbr:Khmer_language
```

The former provides information about the geographical coordinates of Angkor Wat next to the country of location. In this case, it is a literal (text strings with quotation marks) with a data type indicated by two carets (a coordinates format). The latter provides information about Cambodia. In this case, the language used in the country. By adding more statements related to the previous statement, we can create chains of statements to describe potentially unlimited amounts of data. The chains are created in the form of HTTP URIs. In fact, this is

[5] Resources can be anything people would like to describe, ranging from websites, digital photos, and software, to analogue photos, concepts, and facts

[6] https://www.w3.org/TR/turtle/

the fourth principle. Links in LD facilitates the interweaving of data, namely "Web of Data" [90].

By developing LD, experts claim that we can perform intelligent searches using the semantic relations between resources [112]. For this thesis, the LD principles are sufficient to understand the importance of semantic relations, and detailed knowledge about query languages and ontologies that support the intelligent search systems is not required. However, we briefly explain them in order to provide a bigger picture of the the Semantic Web. The query languages and ontologies are not explicitly included in the LD principles, but, in the context of the Semantic Web, they fall into the use of standards in principle three. Three standards below (SPARQL, RDFS, and OWL) are already widely acknowledged in this sense [90].

1.3.2 QUERY LANGUAGE

SPARQL stands for SPARQL Protocol and RDF Query Language [21]. It is a standard for query languages for RDF. In Chapter 5, we use various SPARQL queries to retrieve RDF data in a web application by specifying graph patterns. Although it is named as a query language, it can also be used to transform RDF data into other data. In addition, SPARQL offers "standard mechanisms for reasoning, where additional solutions can be inferred from the explicit semantics" [90]. As reasoning makes implicit conclusions from explicit information, we can open an avenue to new knowledge discovery or serendipity.

Typically, LD is deployed in a triple store where RDF triples are stored and a SPARQL endpoint is provided, which is an interface to query RDF triples. In case of the Angkor Wat example, SPARQL can retrieve all resources located in Cambodia from given triple stores. Angkor Wat is found as one of the results (?subject) among others (e.g. dbr:Kampot_Zoo) from the following query:

```
PREFIX dbp: <http://dbpedia.org/property/> .
PREFIX dbr: <http://dbpedia.org/resource/> .

SELECT ?subject
WHERE
{
  ?subject dbp:location dbr:Cambodia
}
```

A key feature of SPARQL is federated queries. Federated queries allow us to search multiple RDF triple stores through distributed SPARQL endpoints. Thus, RDF triples can be searched and retrieved from multiple triple stores on the web.

1.3.3 ONTOLOGY

Ontologies are formal descriptions of data structure to support knowledge representation. Two languages that can be used to express ontological information are Resource Description Framework Schema (RDFS) and Web Ontology Language (OWL). RDFS and OWL are the standards recommended by W3C for this purpose [17]. RDFS can describe the basic semantics of RDF triples. It provides vocabularies, primarily defining classes (definitions and types of individual data) and properties (relations between subject and object), and their relationships to describe the knowledge that individual RDF resources represent (e.g. relations between classes, relations between properties, data types). For example, CIDOC-CRM [1] uses RDFS to encode the knowledge representation of CH data. It describes the

concepts and relationships of entities including place, time, event, person, and object to document CH phenomena. [14] extends and reuses the RDFS vocabulary to specify other semantics such as property and class constraints. With OWL, a dataset's semantics can be more richly defined, enabling the automatic integration of data from diverse sources [90].

In RDFS and OWL, a member of a class is called an instance of the class; the class of a resource is called its type [90]. In the example of Angkor Wat in DBpedia above, Angkor Wat (dbr:Angkor_Wat) is called an instance of historic building, because it is one of the members of historic buildings. Thus, the class is historic building (dbo:HistoricBuilding), which is a type of the resource, Angkor Wat. Predicates such as dbp:location, geo:geometry, and dbo:language are the properties. OWL and RDFS can be used to define such classes and properties. For example, they can define which type of class can be used in the subject and object positions. In case of dbp:location, the subject is the historic building class (rdfs:domain) and the object is the country class (rdfs:range)[7]. OWL and RDFS can also define subsumptions, using classes and properties. For instance, if dbo:HistoricBuilding is a sub-class of dbo:Building, all instances of dbo:HistoricBuilding are also instances of dbo:Building. This inheritance of class relations is called transitivity [90]. A sample OWL syntax is provided as follows:

```
@PREFIX dbr: <http://dbpedia.org/resource/> .
@PREFIX dbp: <http://dbpedia.org/property/> .
@PREFIX dbo: <http://dbpedia.org/ontology/> .
@PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@PREFIX owl: <http://www.w3.org/2002/07/owl#> .
@PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

dbo:HistoricBuilding  rdf:type  owl:Class .
dbr:Angkor_Wat  rdf:type  dbo:HistoricBuilding .

dbp:location  rdf:type  owl:ObjectProperty ;
              rdfs:domain dbo:HistoricBuilding ;
              rdfs:range  dbo:Country .
```

As OWL specifies well-defined relationships between classes and properties, various communities created their own ontologies, tailored to specific domains. For example, the concept of time (relationships between time-related concepts such as year and week) can be expressed using the Time Ontology in OWL [57], which we reuse in Chapter 3 to design an ontology for numeric dates.

1.4 CHALLENGES OF LINKED DATA

Based on the LD principles, LD has been generated in a broad range of domains, which is visually illustrated by the LOD Cloud[8].

To investigate LD quality in this thesis, we take Juran's definition of quality as "fitness for use" [78]. This includes parameters such as availability, reliability, maintainability, and manufacturability. We also follow ISO 9126 "Software engineering — Product quality" where quality is defined based on three quality models: quality in use, intrinsic, and extrinsic. The last two deal with static properties on the structure and the product behavior in its environment respectively. This thesis focuses on quality in use, which concerns "perceived quality by the end user in their context" [78]. From the perspective of information systems,

[7]This is an example and may differ from the actual DBpedia ontology

[8]https://lod-cloud.net/#about, last accessed 2022-01-22.

we aim to measure systems quality and information quality through user surveys in terms of user satisfaction. Those categories are described in the Delone and McLean IS success model [78].

The quality of LD is critical for LD applications, therefore, computer scientists have made various efforts to analyze and improve the LD quality (for example: [39, 45, 56, 63–65, 68, 69, 85, 109, 113, 115, 116, 137, 142]). Although tools are important, typically research on LD quality concerns the data quality and not the tool quality. This is problematic because tools are needed in many aspects of data: it is often discovered, accessed, processed, visualised, shared, and preserved by tools. Therefore, this thesis divides the quality of LD into two dimensions:

- Linked Data Quality
- Linked Data Tooling Quality

In order to study two dimensions of LD quality, we also take the LD stakeholders into account. As Hawkins [88] points out, it is of the highest importance to evaluate the quality in their interests. In addition, Zaveri et al. [142] recognize a research gap on the focus of LD quality between the use and publication of LD. Therefore, it is crucial to examine the quality identified by different stakeholders such as LD users and LD publishers. To this end, we identified three types of stakeholders. We discuss this in Section 1.5. By intersecting two dimensions and three stakeholders, we can build a matrix of our research coverage for each chapter in this thesis (Table 1.1).

1.4.1 LINKED DATA QUALITY

The first dimension of LD quality is the quality of the data itself. An obvious challenge is to define what quality is. In this respect, we can start from more generic quality criteria, namely the quality of Open Data. In 2009, Berners-Lee advocated the Five-Star Open Data scheme in which five stars are the most advanced [48]. A paraphrased version by Hogan [90] is as follows:

* Publish data on the Web under an open licence

** Publish structured data

*** Use non-proprietary formats

**** Use IRIs to identify things

***** Link your data to other data

As publishing data on the web is often costly, this is a bottom-up approach to bootstrapping LD in an inclusive manner for data publishers. Although LD is not mentioned, the fourth and fifth stars implicitly suggest LD as a best practice. The scheme is an excellent concise guideline for data publishers intending to create valuable Open Data for data consumers. However, this does not deal with the detailed quality of LD.

Data quality has become a critical area of study in LD in recent years. Consequently, there are some attempts to specify the definition and quality criteria [64]. However, no universal consensus has been reached on what constitutes high-quality LD. This may be due

to the fact that data quality is often specific to the needs of individual users. Hogan [90] argues that the quality of LOD depends on applications and that a clear definition of purpose is missing in LOD, allowing for various interpretations of the user. Although there could be general quality criteria that most would agree upon, the final judgment on quality is difficult to standardize and may not benefit from strict standardization.

Nevertheless, quantitative and qualitative analyses are the two common approaches to evaluate data quality. The quantitative method deploys a high degree of statistical analyses on the entire or a large part of a dataset in question. The qualitative method focuses more on the details of the fine-grained individual data in the dataset. Both methods have advantages and disadvantages. The former can reveal the tendency of the dataset quality but lacks the details of quality for individual data, due to a lack of data observation on an individual level. The latter can obtain detailed information about the quality of sample data but lacks the statistical validity of a large portion of the entire dataset.

Many previous studies tend to analyze and evaluate the data quality quantitatively using computational calculations [45, 56, 65, 116]. However, this thesis deploys a middle-ground approach between quantitative and qualitative research in order to provide new insights into the data quality. As we define research strategies by defining the LD stakeholders (see Section 1.5), extra attention is paid to qualitative analysis which brings valuable information from the perspectives of data producers and consumers. In other words, our analysis is more closely connected to potential use cases in CH and DH. We discuss this often in analyses of data and tool quality in the next chapters.

Despite the challenges of definitions, we focus on the quality of data connectivity or linking for LD (Chapter 2). As seen in the fourth principle of LD, linking plays a vital role in LD. The more links are provided, the more data can be connected. We select eleven major LD sources widely known in CH and humanities and semi-automatically analyze the data quality of instances of certain classes. Our novelty lies in the evaluation of quality for instances in LD rather than the overall data quality among thousands of LD datasets.

As we identify issues of LD quality in Chapter 2, Chapter 3 explores the details of date entities. Not only investigating the current quality of the entities, but we also propose a solution to fill the gaps of the temporal entities in LD. We design Linked Date Entities (LODE) for numeric time entities, taking into account the existing resources such as the Time Ontology in OWL and Wikidata. An RDF model and lookup service are implemented. In total, 2.2 million entities are created, covering a duration of 6000 years. We test LODE with two use cases.

LD quality can also be evaluated by measuring the gaps between textual data and LD. For this purpose, we analyze quality gaps between Wikipedia and its related LD: Wikidata and DBpedia (Chapter 4). We use Natural Language Processing (NLP) techniques applied to a Wikipedia article to extract more fine-tuned information about a person (i.e. biography) than the information that the current Wikidata and DBpedia hold. Then, we generate a new set of LD which adds more semantic information to the current DBpedia and Wikidata. This potentially improves the quality of DBpedia and Wikidata.

1.4.2 LINKED DATA TOOL QUALITY

The second dimension of LD quality is the quality of LD tools. Tools and services are critically important in terms of data use and acceptance of a technology. In our thesis, the

quality of tools relevant to CH and DH will be explored.

Observing the history of the Semantic Web in CH and DH from a tooling perspective, it is in the middle of a transition. According to Hyvönen [93], the Semantic Web research in CH has concentrated on syntactic and semantic interoperability and data aggregation. He claims that a great deal of effort has been dedicated to metadata standards and data models for harmonizing data. Examples are CIDOC-CRM, Europeana Data Model (EDM)[9], IFLA Library Reference Model (LRM)[10], and The DPLA Metadata Application Profile (MAP)[11]. Consequently, a variety of web portals have been created such as Europeana[12], Digital Public Library of America (DPLA)[13], and the Sampo portals[14].

Hyvönen further states that the primary use case of the portals is to provide the user with enhanced information retrieval (searching and browsing) facilities for exploring the data. While he calls them the "First Generation Semantic CH Portals", he also observes an emerging trend of the "Second Generation" systems that provide the user with tools for solving DH research problems. The new systems enable the users to analyze a large volume of data statistically. He refers to the examples of BiographySampo, Six Degrees of Francis Bacon, and Epistolarium[93][15]. This transition is a shift from data publishing to data analysis.

Despite this trend, central discussions of recent LD research are often still geared toward data-related issues, especially data modelling, production, enrichment, integration, and sharing in the context of domain research questions and CH collections [121, 144]. In this regard, researchers tend to concentrate on the development of LD tools for data publishers. Consequently, discussions about tools tailored for data consumers in CH and humanities have not been thoroughly investigated.

In the thesis, two chapters are relevant to this topic. In Chapter 3, we first analyze the existing LD for date entities. Then, to enhance the availability and accessibility of such LD, we present an example solution with SKOSMOS software[16]. Chapter 5 is more concentrated on the topic of tooling for LD than Chapter 3. It investigates the landscape of the existing LD tools with a focus on CH and humanities. In particular, Wikidata-centric tools are scrutinised, as the use of Wikidata has been intensified over the recent years in the CH and humanities [121, 144]. To address the missing requirements of such tools, we devise a new web application to facilitate the use of Wikidata within the domain, specializing in the temporal data , which is fundamental for many sub-fields. The usability and usefulness of the application are measured by an end-user evaluation, consisting of focus group and questionnaires. As we identify that the studies of user evaluation for LD is under-explored, it would be a contribution to the LD, CH, and DH communities at large.

[9] https://pro.europeana.eu/page/edm-documentation (Accessed 2024-09-16)
[10] https://repository.ifla.org/handle/20.500.14598/40 (Accessed 2024-09-16)
[11] https://pro.dp.la/hubs/metadata-application-profile (Accessed 2024-09-16)
[12] https://www.europeana.eu (Accessed 2024-09-16)
[13] https://dp.la (Accessed 2024-09-16)
[14] https://www.europenowjournal.org/2019/09/09/linked-data-in-use-sampo-portals-on-the-semantic-web/ (Accessed 2024-09-16)
[15] https://ckcc.huygens.knaw.nl/epistolarium/ (Accessed 2024-09-16)
[16] `https://skosmos.org/` (Accessed 2025-09-04)

1

1.5 LINKED DATA STAKEHOLDERS

While the first line of investigation is two dimensions of the LD quality, the second line concerns the stakeholders of LD. We examine the critical aspects of LD in CH and humanities, regarding three most relevant stakeholders: data producers, data consumers, and developers. The concept of data producers (publishers) and consumers is borrowed from the W3C Recommendation "Data on the Web Best Practices" [36]. This document indicates that there is a fundamental need for a common understanding between data publishers and data consumers, because "data publishers' efforts may be incompatible with data consumers' desires".

As Folmer and Verhoosel rightly indicated by referring a previous study on quality [78], different stakeholders possess different perspectives on quality, reflecting their distinct interests and priorities. In this regard, it is highly valuable to introduce the stakeholder dimension for quality study in this thesis. We added one more stakeholder: developers who play a crucial role in the implementations of LD. To enhance impactful LD development in CH and DH, it is of utmost importance to understand these stakeholders. We define them as follows:

Data producers are "persons or groups responsible for generating and maintaining data" [36]. The key players in preserving CH are memory organizations [92]. Therefore, they are typically CH institutions, also known as GLAM institutions. They employ CH professionals such as curators, librarians, and archivists. Mckenna et al. [105] use the term Information Professionals (IPs) to refer to the keepers of CH data, and as experts in the field of metadata creation and knowledge discovery. The GLAM institutions generate LD often from their cultural collections and secondary information. The data coverage is extremely broad. For example, galleries hold information about collections of paintings as well as artists. Museums hold a wide variety of collections, ranging from archaeological and ethnological objects to biological and astronomical objects. Libraries have collections of prints including books, newspapers, and magazines. Archives hold manuscripts and modern maps. Other institutions host collections of music records, films, and performing arts. There are also organizations that deal with only digital data of GLAM institutions. They often compile and curate data for the end-users on the web.

Data consumers are "persons or groups accessing, using, and potentially performing post-processing steps on data" [36]. They are typically those who are interested in and visit the GLAM institutions recreationally and educationally as tourists, guides, students, and researchers. In our thesis, we often deal with the digital information from the GLAM collections, thus, the data consumers are the end-users of digital applications provided by GLAM institutions, including their websites, apps, and online databases. It is often the case that the data consumers are the least skilled among the three stakeholders in terms of LD. Their knowledge about the GLAM data may also be limited, unless they are highly interested in it. As an individual person, both humanities and DH researchers fall into this group. However, skills related to LD vary significantly (See also Chapter 5). As the number of end-users could grow considerably in the public, this group potentially would have the highest number of members among the three stakeholders.

Developers are the LD technicians. They are typically computer scientists or IT experts in universities, research institutions, and companies who have technical knowledge and skills not only to design and generate LD, but also to develop tools and applications for it. They may or may not collaborate with data producers and data consumers. They may be involved in the development of technical standardization of LD. Normally, they are not domain experts; therefore, they have limited knowledge about the data, but have opportunities to contribute to the implementations of LD in different domains. Given the level of expertise needed, the number of members may be the lowest among the stakeholders. A small number of DH and CH experts specializing in LD are also a part of this group. However, their technical skills in LD would generally be lower than those of computer scientists.

In some cases, the distinction between these three stakeholders is blurred. Data consumers may also act as data producers [36] and/or developers simultaneously. This overlap is increasingly common in the interdisciplinary field of DH. DH researchers often produce data and develop tools while also serving as end-users of other LD services. The LD survey by McKenna et al. [105] targets IPs as well as researchers and academics with experience in the GLAM and/or LD sectors. This highlights the complex landscape of LD stakeholders.

Despite the challenges in providing clear definitions, our research investigates the gaps between stakeholders to clarify the current situation of LD practices. As mentioned above, the matrix presented in Table 1.1 consists of two dimensions and three stakeholders, which illustrates the approximate coverage of each chapter of this thesis. For instance, Chapter 2 deals with LD data quality issues that concerns the (potential) engagement of data producers and data consumers. Chapter 3 instead focuses on both data and tool qualities and the stakeholder engagement for data producers and data consumers, while developers are only involved in the data quality aspect. We will discuss this in Section 1.6 in more detail. Consequently, Table 1.1 highlights that the thesis addresses all areas of the LD quality matrix.

Table 1.1: Chapter coverage for stakeholders and quality dimensions

	Data Quality	Tool Quality
Data Producer	Ch 2, 3, 4, 5	Ch 3
Data Consumer	Ch 2, 3, 4, 5	Ch 3, 5
Developer	Ch 3, 4	Ch 5

It should be noted that the challenges we have described are not totally new. Hogan [90] lists four challenges when it comes to LOD: a) the LD community focuses on publishing data, postponing the discussions on the use of data, b) purpose of LOD is not clearly defined to determine the quality, c) the cost of publishing LD is high, and d) techniques to build LD applications for novice users have not been properly explored. We deal with all the challenges to a certain degree throughout the thesis, except the cost of publishing.

1.6 RESEARCH QUESTIONS AND CHAPTER OUTLINES

Taking the challenges of LD into account, this thesis focuses on and address the following main question:

1

How can the quality of data and tools for Linked Data in Cultural Heritage and Digital Humanities be enhanced? To be more specific, we divide this question into four Research Questions (RQs) that roughly address the main research question of each chapter in the thesis:

- RQ1: How is the quality of Linked Data instances in Cultural Heritage, particularly in terms of their connectivity?
- RQ2: How can Linked Data connectivity for date entities be improved?
- RQ3: What are the quality gaps in biographical information between Wikipedia and Linked Data, and how can Information Extraction on Wikipedia be used to address them?
- RQ4: What are the effective designs and functionalities for Linked Data tools to support research using temporal information in Cultural Heritage and Digital Humanities?

Throughout this thesis, we employ two research strategies to address the RQs, namely analysis and engineering. In Chapter 2 (also partially in Chapter 3, Chapter 4, and Chapter 5), we perform structured analyses to identify the current state of the LD quality. In Chapter 3, Chapter 4, and Chapter 5, based on the analyses, we design and engineer data and tools to provide solutions to improve the LD quality. These strategies are combined with a design science approach. In design science, our RQs can be rephrased and translated into two problem dimensions, which are so-called "knowledge problems" and "real-world problems" [140]. In our case, the knowledge problem is that we do not have a clear picture of the current situations of LD in terms of data quality and tool quality. The real-world problem is that we have not yet fully unlocked the potential of LD in CH partly due to the quality issues. By addressing the knowledge and real-world problems, our main question can be answered. In addition, our research questions roughly correspond to pairs of question/problem and answer/solution in each chapter of this thesis. Therefore, the thesis fills those knowledge and reality gaps by analyzing the problems and devising solutions. This structure makes this thesis relatively comprehensive for the specific domains of problem/question. In Chapter 6, we reflect on the RQs and lessons learned in the conclusions.

Regarding the stakeholders, Table 1.1 illustrates the coverage in each chapter. The data producers and consumers are the main actors for Chapter 2. We evaluate the data quality in the context of research capability of the produced LD sources from the data consumer's perspective. As both parties share interest in data quality, our analysis and conclusion incentivize their commitment in the future. In contrast, Chapter 3 deals with the intersections of all stakeholders and two dimensions, except developers and tooling. The developers and tooling are less relevant for this chapter, as we use an existing tool. The other stakeholders have common interest in the newly generated data and, to a certain extent, the tool to access and use the data. Chapter 4 addresses LD production, thus, it is related to the data producers and consumers. Although tooling is relevant, in this case, we emphasize more on how to establish a workflow to use existing tools/services than how to develop those tools. Besides, the tools/services in this chapter are not tools for ordinary users, but code libraries for developers. Chapter 5 involves a tool development. Therefore, developers are the main

stakeholders. At the same time, this is the only chapter in which our research requires direct contact with the stakeholders. Potential data consumers are invited to focus group workshops to provide feedback on the developed tool. As we scrutinize the existing LD tools, we introduce the perspective of the developers. Data producers and consumers are relevant in terms of quality of the data used for the tool.

*

Now, the packing of our background is complete. Let us start our journey with Berners-Lee's dream for the Semantic Web.

2

Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data Integration

In cultural heritage, many projects employ Named Entity Linking (NEL) through global Linked Open Data (LOD) references in order to identify and disambiguate entities in their local datasets. It allows users to obtain extra information and contextualise the data with it. Thus, the aggregation and integration of heterogeneous LOD are expected. However, such development is still limited partly due to data quality issues. In addition, analysis on the LOD quality has not sufficiently been conducted for cultural heritage. Moreover, most research on data quality concentrates on ontology and corpus level observations. This chapter examines the quality of the eleven major LOD sources used for NEL in cultural heritage with an emphasis on instance-level connectivity and graph traversals. Standardised linking properties are inspected for 100 instances/entities in order to create traversal route maps. Other properties are also assessed for quantity and quality. The outcomes suggest that the LOD is not fully interconnected and centrally condensed; the quantity and quality are unbalanced. Therefore, they identify key challenges forautomatically identifying, accessing, and integrating known and unknown datasets. This implies the need for LOD improvement, as well as the NEL strategies to maximise the data integration.

This chapter is partly based on *Go Sugimoto. Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data Integration, 1 Jan. 2023, the Semantic Web, 1, 55 - 100.* [125].

2.1 INTRODUCTION

This chapter mostly focuses on the data quality, addressing RQ1: "How is the quality of Linked Data instances in Cultural Heritage, particularly in terms of their connectivity?" (Section 1.5). The main stakeholders are data producers and data consumers who would ideally assess the data quality in close collaboration. As stated in Section 1.2, Linked Open Data (LOD) is the term used to describe Linked Data (LD) that adheres to Open Data principles. In particular, LD released under an open license qualifies as LOD. In this chapter, we use the term LOD.

In recent years, Linked Open Data (LOD) has been widely acknowledged and data rich institutions have generated a large volume of LOD. As of May 2020, the LOD Cloud website reports 1,301 datasets with 16,283 links[1]. The real power of LOD originates from a very simple philosophy of the Web inventor. Berners-Lee [48] states “include links to other URIs. so that they can discover more things”, hence the name “Linked” (Open) Data. LOD transforms distributed data in Resource Description Framework (RDF) into a connected global knowledge graph and allows us to find and formulate new information and knowledge [50]. This vision seems to be particularly suited for research activities. However, it seems that this scenario is not happening as quickly as we expected. It is still unclear whether we have discovered something significant in this manner. One of the reasons for this problem is the gap between the LOD producers and consumers, which is heavily attributed to data quality. Zaveri et al. [142] state that there is less focus on how to use good quality data than to how to publish it.

In this chapter, we explore the problems of LOD quality from the user’s point of view. In particular, we analyse the linking quality of LOD from a research perspective in the field of cultural heritage and Digital Humanities (DH). Our study on this fundamental aspect of LOD should be able to provide a better understanding of a bottleneck of LOD practices. Although we concentrate on these domains, we believe that our analysis is equally valuable in other domains, because the analysed data is highly generic.

In cultural heritage and DH, many projects create and use a wide range of LOD for research purposes. In the course of populating and improving LOD, they often execute curatorial tasks such as Named Entity Recognition (NER), entity extraction, entity/coreference resolution, and Named Entity Linking (NEL) [54, 62, 72, 132, 143]. These are the tasks to identify, disambiguate, and extract entities/concepts from data, and to reconcile and make references to entities in another data. Thus, we can find more information on the web. In this article, we use NEL as a catch-all term for all these tasks.

For example, Europeana executes NEL in a large number of cultural heritage datasets and creates links to widely known LOD sources including GeoNames, DBpedia, and Wikidata that we discuss [120, 122]. Jaffri et al. [96] echo this view, stating that many datasets are linked with DBpedia entities through the owl:sameAs property. In practice, this means that information about the same entity (e.g., place, person, event etc.) is stored in different LOD datasets on different servers. As Tomasuzuk and Hayland-Wood [129] indicate, RDF enables us to join data stored at disparate sources and provide the user with an integrated perspective of this data. This is called data integration. If one dataset only supplies partial information

[1] https://lod-cloud.net/#about, last accessed 2022-01-22.

about an entity, NEL allows us to retrieve more information from all linked datasets, by "merging" data through links. For example, a local record may describe Pablo Picasso in Spanish, but lacks information in other languages. If we perform NEL for the database of this record and create links to other LOD, we can fetch missing information, including a label in Chinese and related artists in Arabic. Consequently, a more comprehensive record about Picasso can be created. In this regard, NEL serves as a building block of LOD, fostering connection, compilation, aggregation, and contextualisation of (distributed) information.

What is not investigated in cultural heritage and DH is, what impact NEL and subsequent data integration have for future research? Currently, there is a tendency for entity linking to become a purpose by itself, without examining the consequences of the linking. Due to the relative infancy of LOD in the field, perhaps most effort has been put into the aspect of data discoverability on the web, which NEL also facilitates. This function of LOD may not require extensive use cases after NEL is performed. In any case, data producers are often not fully aware of the next steps for research using LOD, as well as the needs of the data users. Although not limited to these domains, Data on the Web Best Practices[2] observes: "the openness and flexibility of the web create new challenges for data publishers and data consumers, such as how to represent, describe and make data available in a way that it will be easy to find and to understand".

Currently, the benefit of data integration using NEL is often restricted to the data sources within a single institution or domain. For instance, an advanced semantic search is developed for the historical newspapers in the Netherlands [134]. In fact, the investigation of the aggregation and integration of heterogeneous LOD from different data providers is rather rare [72], or done with relatively small multiple sources. A few exceptional cases are found in museums and institutions in France [37] and Spain [104]. Still, the formation of new knowledge based on complex queries across distributed LOD resources is not easily implemented. As such, the full potential of LOD has been neither fully explored nor verified. The practice of LOD-based research using distributed data still faces many challenges.

In terms of data linking quality, computer science communities have intensively worked on this issue in the past years. Critical quality issues of linking have been frequently raised and discussed in the studies of LOD [39, 45, 56, 63–65, 68, 69, 85, 109, 113, 115, 116, 137, 142]. We discuss this in Section 2.2 in more detail. However, one specific aspect helps here to explain our motivation. Most previous research regards owl:sameAs as a central property for LOD linkages, because it is a W3C recommended standard and serves as a bridging link between identical entities. We also think that it plays an important role to automate data processing using federated SPARQL queries in dispersed datasets, because we know the property beforehand without knowing heterogeneous and complicated ontologies of individual datasets. At least there is no doubt that LOD information can be automatically traversed and aggregated by simply following the links through this property. Therefore, we are interested to understand the future prospect of LOD automation by examining commonly used properties.

Taking this background into account, this chapter aims to evaluate the quality of widely known (referential) LOD as the target resources of NEL. In particular, the linking quality and connectivity is analysed in detail in order to provide an overview of the current "state of NEL ecosystem". To this end, we examine LOD entities/instances through lookups. With a

[2] https://www.w3.org/TR/dwbp/ last accessed 2021-01-26.

special emphasis on multi-level traversability in the LOD cloud, we can estimate the impact of NEL for end-users. In other words, our research questions are as follows.

- RQ1.1: When a local dataset links to a global LOD, what level of information can we find?
- RQ1.2: How can we follow links "to discover more things"?
- RQ1.3: How are the entities in (the core part of) the LOD cloud connected to each other and can be navigated?
- RQ1.4: What kind of information can be obtained by automatic graph traversals through standardised properties like owl:sameAs?
- RQ1.5: What are the linking and content patterns for different types of entities?

In terms of RQ1.2, it concerns not only an engineering problem but also issues of data quality, as the mechanism for following links influences both the usability and usefulness of the data for end users. This mechanism involves how users navigate graphs within LOD sources (RQ1.3). The effectiveness of this mechanism determines the amount and quality of new information that can be obtained (RQ1.4).

As LOD potentially enables us to undertake machine-assisted research with the help of more automated data integration and processing, this project serves as a reality check for the current practices of LOD in the field.

The structure of this chapter is as follows. Section 2.2 explores the related research. Section 2.3 describes objectives, scopes, and methodology. Section 2.4 presents the analysis of 100 entities in five categories relevant to cultural heritage data integration and contextualisation. The final section summarises the discussions and outlines ideas for future work.

2.2 RELATED WORK

Over the last years quantitative research has been carried out intensively for the LOD quality. The landscape of previous studies is examined in an in-depth survey by Zaberi et al. [142]. They analyse 30 academic articles on data quality frameworks and report 18 quality dimensions and 69 metrics, as well as 20 tools. Many studies investigate the linking quality, but some aim to assess broader aspects of LOD quality. For instance, Färber et al. compare DBpedia, Freebase, OpenCyc, Wikidata, and YAGO with 34 quality criteria [79]. They span from accuracy, trustworthiness, and consistency to interoperability, accessibility, and licences. Schmachtenberg et al. [116] update the 2011 report on LOD, using the Linked Data crawler, analysing the change of LOD (8 million resources) over the years. Debattista et al. [65] provide insights into the quality of 130 datasets (3.7 billion quads), using 27 metrics. However, the linking on which we would like to focus is a small part of the metrics. Mountantonakis and Tzitzikas [109] have developed a method for LOD connectivity analysis, reporting the results of connectivity measurements for over 2 billion triples and 400 LOD Cloud datasets. A rather different project has been conducted by Guéret et al. [84]. They concentrate on the creation of a framework for the assessment of LOD mappings using

network metrics. They specifically look into the quality of automatically created links in the LOD enrichment scenario.

In parallel, a number of valuable contributions have been made to scrutinise owl:sameAs and "problem of co-reference" [97]. Firstly, there are critical discussions about the proliferation of owl:sameAs semantics [85]. Secondly, several large scale statistical analyses uncover the status of owl:sameAs networks to detect errors for 558 million links [113], verify the proliferation [68, 69](4352 and 8.7 million links respectively), and propose solutions. Most projects concentrate on macro-level studies and statistical observations of the comprehensive cross-domain LOD cloud, applying metrics to measure the data quality through dumps and SPARQL endpoints. Their methodologies help us to gain a holistic view of the development of the LOD cloud in terms of linking quality.

There are also a few examples of "semi-micro" level research, using domain specific datasets. Ahlers [39] analyses the linkages of GeoNames (11.5 million names). He reveals some cross-dataset and cross-lingual issues and distribution biases. Debattista et al. [63] inspect the Ordnance Survey Ireland (50 million spatial objects) in order to identify errors in the data mapping for the LOD publishing and check the conformance to best practices. Although the datasets pass the majority of 19 quality metrics in the Luzzu framework [64], the low number of external links (only DBpedia) is clearly our concern.

The studies for the cultural heritage domain are relatively new. Candela et al. state that there has been so far no quantitative evaluation of the LOD published by digital libraries [55]. They systematically analyse the quality of bibliographic records from four libraries with 35 criteria covering 11 dimensions to provide a benchmark for the library community. The research on the LOD quality for a broader cultural heritage including museums and archives is scarce.

Apart from Mountantonakis and Tzitzikas, macro-level research projects oftentimes treat data sources (or corpora) as a whole, when investigating owl:sameAs link connectivity. In other words, the data connectivity is examined regardless of the user mobility at an instance level. For example, their research does not reveal if the connection for a specific instance such as Mozart is available between data source A and B, even if they detect many links between the instances in the two sources. This is because the domain coverage may be different: A originates from a Polish library and B from a Greek museum. Mozart could be found in both, but could be in neither. To this end, it is necessary to observe trees (Mozart as an instance) not forests (the data source A and B as a collection of instances).

In addition, most macro-level analyses are not designed for multiple graph traversals. One of the exceptions is Idrissou et al. [95] who indeed claim that gold standards for entity resolution do not go beyond two datasets. Interestingly, they develop hybrid-metrics that combine structure and link confidence score to estimate the quality of links between entities for six datasets from the social science domain. Although we agree that accurate automated evaluation of links is much needed, our study aims to gain deeper understanding of smaller sampling entities. Going back to our analogy, we currently cannot know how much and what kind of data we can find by following a link from Mozart in data source A to an entity in source B, which provides links to an entity in source C. Therefore, a close observation of instances is needed. The instance level maneuverability indicates whether and how users can navigate themselves in the knowledge graphs and can obtain related information from various data sources, and potentially integrate them.

2.3 OBJECTIVES AND METHODOLOGY

We explain the process of defining objectives and methodology in four sub-sections. The first section describes the scope of the linking quality evaluation. The second section discusses the nature of research in cultural heritage and DH in relation to conceptual models and ontologies, in order to specify the object of analysis. The third section details the data sampling. The fourth section deals with the technical methods of a wide range of analyses.

2.3.1 SCOPE OF ANALYSIS AND GRAPH TRAVERSALS

Our research will not repeat the comprehensive statistical analyses on the LOD quality according to the existing or newly created comprehensive metrics. In contrast to previous research, we deploy a micro-level analysis. Our research deals with a small ecosystem of LOD in the cultural heritage NEL, based on an empirical qualitative and quantitative method. In particular, it focuses on user maneuverability for arbitrary LOD entities. We analyse multi-level graph traversability using standardised properties, especially bearing the automatic data traversals and integration in mind.

The primary goal is to create "traversal maps" of major LOD data sources at an instance level. "Traversal maps" are maps illustrating all possible routes of graph traversals in the LOD cloud (RQ1.3). We specialise in the route of standardised properties including owl:sameAs (RQ1.4). Naturally, the collections of instances covering the same topic (i.e. categories in Section 2.3.2) are of vital importance for the analysis (RQ1.5). Subsequently, it is expected to provide a better understanding of which referential resources are accessible in what way between multiple sources (RQ1.1, 1.2). This scope enables us to deliver an observation more from the data user's perspective than the producer's. The traversal maps should be helpful for the end-users to orient themselves in the LOD cloud and formulate strategies for data navigation and integration to capitalise NEL.

The use case for the LOD traversals in this article is the following: we/user manually look up a LOD entity/resource identified. Then, they follow available links in the entity to reach identical and/or the most related LOD resources. For example, one may traverse an RDF graph from a resource in DBpedia to a resource in Wikidata via owl:sameAs:

dbr:1969 owl:sameAs wd:Q2485 .

Hyperlinks are documented and counted to generate traversal maps. To support the link quality analysis, information about other content is also documented and counted (RQ1.5). It includes the amount of rdfs:label, rdf:type, skos:prefLabel and skos:altLabel as well as rdf:resource, and rdf:about (see Section 2.3.4). The traversal continues as long as it is within the specified datasets boundaries (see Section 2.3.3). The reason to evaluate lookups instead of data dumps and SPARQL queries is that they play a vital role to publicly and openly raise awareness of the data existence that NEL essentially needs. To our knowledge, none of the previous studies works on lookups.

Regarding the link types, the W3C recommended properties, owl:sameAs, rdfs:seeAlso, and skos:exactMatch are used[3]. It is a common practice that information providers set owl:sameAs links to URI aliases [45, 50]. In addition, schema:sameAs is included, due to its popularity. One of the advantages of those standards is that the properties are widely known (see Section 2.3), implying no prior knowledge is required to access and process data.

[3]One property per ontology is selected.

As Hartig [86, 87] observes, it is highly important that the end users can obtain data from initially unknown data sources. In other words, they should be able to discover new LOD sources at runtime by following RDF links [50].

Since rdfs:seeAlso may be asymmetric, our analysis is not limited to LOD and symmetric graphs. This means that the sources and destinations of incoming and outgoing links are not 100% synchronised as identical LOD entities. For example, "Italy" in Getty TGN contains rdfs:seeAlso for an HTML representation (`http://www.getty.edu/vow/TGNFullDisplay?find=&place=&nation=&subjectid=1000080`). This is allowed in the specification[4]. Another reason to avoid strict co-references is that it is hard to find and evaluate the same identity only by URIs. For instance, a VIAF record provides a link to Getty ULAN in the following syntax: `http://vocab.getty.edu/ulan/500240971-agent`. This resolves to `http://vocab.getty.edu/ulan/500240971`. In general, redirects introduce technical complexity for the analysis. As a consequence, the links to the same domain name in the URIs (e.g. getty.edu is same as vocab.getty.edu) are regarded as the same destination, regardless the identity and format of the entity. In this way, our analysis attempts to bypass complicated discussions over the accurate semantics of properties such as owl:sameAs [85].

When assessing the quality of LOD, proprietary properties cannot be ignored. They often contain interesting and specialised information. However, we put less emphasis on them. Compared to standardised properties, these properties may not be frequently used as a means to connect the data sources within the core part of the LOD cloud. Another reason is extensively explained in Section 2.3.4 in the context of difficulties in the data quality comparison, and our compromised approach is described.

Documentation on an instance is recorded in separate tabs in a spreadsheet for each source. VBA scripts are created to aggregate and/or facet datasets. Subsequently, various types of tables and charts are generated. In order to increase the research transparency and reproducibility, our datasets and documentation are fully archived in the Zenodo Open Access repository (https://doi.org/10.5281/zenodo.5913136).

2.3.2 CORE QUESTIONS AND CONTEXTUALISATION IN CULTURAL HERITAGE ONTOLOGY

In order to narrow the scope of the LOD evaluation, this article focuses on addressing typical and generic core questions for cultural heritage and DH alike. For instance, one of the largest cultural heritage data platforms is Europeana. It has created the Europeana Data Model (EDM)[5] in order to capture heterogeneous cultural heritage information. Its Primer[6] notes that "EDM will let users browse Europeana in revealing new ways. It answers the 'Who?', 'What?', 'When?', 'Where?' questions, and makes connections between the networks of stories that will animate Europeana's content". EDM features five classes (agent, event, place, time-span, concept) for this purpose, which are called contextual entities, because they enrich and "contextualise" cultural heritage objects. Although these 4'W' questions are common sense for scientific research in general, they manifest the essence of cultural

[4] https://www.w3.org/TR/rdf-schema/, last accessed 2021-01-26.

[5] https://pro.europeana.eu/resources/standardization-tools/edmdocumentation, last accessed 2021-01-26

[6] `https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf`, last accessed 2021-01-26

heritage research: without them, researchers are hardly able to solve any other research questions in their disciplines. Thus, they provide the contextualisation or foundation of research.

The importance of the four core questions is also reflected in other cultural heritage ontologies. CIDOC-CRM "provides the "semantic glue" needed to mediate between different sources of cultural heritage information, such as that published by museums, libraries and archives"[7]. It centres "Event" as a core entity, connecting "Agent", "Time-Span", "Objects", and "Place". In the library sector, DCMI Metadata Terms[8] also defines almost identical entities: "Agent", "PeriodOfTime", "PhysicalResource", and "Location" among others. In addition, FRBR[9] is a conceptual reference model for libraries which introduces hierarchical concepts of cultural works (i.e. work, manifestation, expression, and item). The Group 1 entities (the products of intellectual and artistic endeavor) are relevant to the What question, whereas the Group 2 entities (person and corporate body) are related to Who. Group 3 (the subjects of intellectual or artistic endeavor) is associated with other W-questions.

Therefore, the evaluation of LOD in this article concentrates on these four questions and use them as categories of our investigation. We employ the following terminology to be more specific: agents (for Who), events (for What), objects and concepts (for What), dates (for When), and places (for Where). Due to the genericness of the categories, investigating the five categories not only helps us to answer our research questions, but also makes our analysis valuable for research outside the cultural heritage field.

2.3.3 DATA SOURCES

Our study introduces two basic strategies for the selection of datasets/data sources. It examines LOD in 1) RDF/XML with 2) unrestricted look-up access (i.e. no API keys). Although there are other RDF serialisation formats, RDF/XML is the only commonly available one for all the data sources described below[10]. On top of the technical setup, we consider popularity (through literature [55, 72, 121, 143]), data volume, coverage, and actual linkages for the selection. The aforementioned LOD cloud is also taken into account, as one of the comprehensive visualisations of LOD networks. Consequently, the following nine data sources which include significant content for the cultural heritage and DH are chosen for examination: 1) Getty vocabularies (ULAN (Union List of Artist Names), AAT (Art & Architecture Thesaurus), and TGN (Thesaurus of Geographic Names)), 2) GeoNames, 3) VIAF , 4) WorldCat FAST, 5) DBpedia, 6) Wikidata, 7) the Library of Congress, 8) BabelNet, and 9) YAGO.

There are two exceptions for the selection criteria. Wikipedia delivers its articles in HTML, but it may be studied as an indicator, because it has a unique position as a global reference on the web inside and outside the LOD context [38, 39, 107, 132]. Indeed, the data in DBpedia and YAGO are derived from Wikipedia[11]. Wikidata has a close relationship with the Wikipedia project. The other case is Europeana. It provides an alpha version API with

[7] http://www.cidoc-crm.org/, last accessed 2021-01-26

[8] https://www.dublincore.org/specifications/dublin-core/dcmiterms/, last accessed 2021-01-26

[9] https://www.ifla.org/publications/functional-requirements-forbibliographic-records, last accessed 2021-01-26

[10] This is mainly due to GeoNames that only provides RDF/XML, KLM, and HTML representation for lookups. This is already a discovery of LOD quality in terms of standardisation.

[11] YAGO2 is used for our study

a public API key[12]. However, it is one of the most valuable LOD resources in the cultural heritage sector, and therefore, it is included. Only those who read the documentation can find the API key and the URI syntax to access the lookup service.

As this study deploys a qualitative analysis, a manageable level of data sampling is considered. It selects twenty representative instances/entities from five categories defined in the Section 2.3.2 (Table 2.1), resulting in 100 entities in total[13]. In order to objectively and systematically select the most relevant entities, we consulted the "Wikipedia most referenced articles"[14] (2011) for the top 20 places and dates, whereas a scientific article about the interaction of top people in Wikipedia is used for the 20 agents [73]. In addition, the top 20 events are retrieved by a SPARQL query from the EventKG endpoint[15] as follows (Listing 2.1):

Listing 2.1: The SPARQL query to retrieve top 20 events from EventKG

```
PREFIX eventKG-s: <http://eventKG.l3s.uni-hannover.de/
schema/>
PREFIX eventKG-g: <http://eventKG.l3s.uni-hannover.de/
graph/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sem: <http://semanticweb.cs.vu.nl/2009/11/sem/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbr: <http://dbpedia.org/resource/>

SELECT ?dbp ?links {
    ?event rdf:type sem:Event .
    GRAPH eventKG-g:dbpedia_en { ?event owl:sameAs ?dbp . } .
    {
        SELECT ?event (SUM(?link_count) AS ?links)
        WHERE {
            ?relation rdf:type eventKG-s:Relation .
            ?relation rdf:object ?event .
            GRAPH eventKG-g:wikipedia_en { ?relation eventKG-s:links ?link_count
              . } .
        }
        GROUP BY ?event
    }
}
ORDER BY DESC(?links)
LIMIT 30
```

It is not trivial to nominate 20 objects and concepts, because cultural heritage and DH cover an extremely broad field. In fact, there are countless numbers of material entities such as museum objects and buildings. Moreover, millions of archaeological objects are even unnamed. Indeed, many object entities are not globally and uniquely identifiable, because they have not (yet) been created in the global references. As such, it is much more challenging to implement entity linking for those entities. Nevertheless, we manually selected 20 entities from the featured articles of Wikipedia[16]. They aim to represent a wide range of chronological, geographical, and thematic diversity[17].

[12] https://pro.europeana.eu/page/entity, last accessed 2021-01-26.

[13] For practical reasons, it concentrates on the English version as the primary resource of an entity when multiple language versions exist (e.g. DBpedia). Nonetheless, other language versions are documented as a reference.

[14] https://en.wikipedia.org/wiki/Wikipedia:Most-referenced_articles, last accessed 2019-09-25

[15] http://eventkginterface.l3s.uni-hannover.de/sparql (last accessed 2019-09-25)

[16] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles,last accessed 2020-03-10

[17] This research investigates tens of thousands of global entities that are reasonably well known and one could look

Table 2.1: 100 entities in five categories selected for analysis

ID	Agents[1]	Events[2]	Dates	Places[3]	Objects and Concepts
1	Carl Linnaeus	World War II	1987	United States	Book of Kells
2	Jesus	World War I	1986	United Kingdom	Vasa (ship)
3	Aristotle	American Civil War	1985	France	The Garden of Earthly Delights (paining)
4	Napoleon	FA Cup	1983	England	Rosetta Stone
5	Adolf Hitler	Vietnam War	1980	Germany	Palazzo Pitti (building)
6	Julius Caesar	Academy Awards	1984	Canada	Boeing 747
7	Plato	Cold War	1982	Australia	Sgt. Pepper's Lonely Hearts Club Band (album)
8	William Shakespeare	Korean War	1968	Japan	Tosca (opera)
9	Albert Einstein	American Revolutionary War	1979	Italy	Blade Runner (film)
10	Elizabeth II	UEFA Champions League	1969	Poland	Uncle Tom's Cabin (novel)
11	Michael Jackson	UEFA Europa League	1978	India	Ming Dynasty
12	Madonna (entertainer)	Olympic Games	1967	Spain	Ukiyo-e (art)
13	Ludwig van Beethoven	Stanley Cup	1981	London	Angkor Wat (building)
14	Wolfgang Amadeus Mozart	Super Bowl	1977	Russia	Toraja (ethnic group)
15	Pope Benedict XVI	Iraq War	1976	New York City	Byzantine Empire
16	Alexander the Great	War of 1812	1975	Brazil	Mars (planet)
17	Charles Darwin	Gulf War	1964	California	Tamil language
18	Barack Obama	Spanish Civil War	1966	New York	Influenza (disease)
19	Mary (mother of Jesus)	World Series	1965	The Netherlands	The King and I (musical)
20	Queen Victoria	EFL Cup	1960	Sweden	Like a Rolling Stone (song)

[1] The priority is given in the following order: page rank, 2Drank (24 languages), and page rank (female).
[2] International events are prioritised, thus a couple of specific events such as US censuses are removed.
[3] Top 20 places are extracted from the general list.

The actual number of entities analysed is 836 (859 occurrences), since some sources do not have the entities the others have. Full details of the entity coverage per data source are provided in Appendix A. Statistically speaking, in case of missing entities, they are included in the calculation and the data values are counted as null[18]. In addition, there are double identity/occurrence (or a kind of "duplicate"[19]) in some sources. The double identities are consolidated as one identity[20]. When an entity lookup is not accessible for technical reasons, the data is included in the statisticsas a zero value[21].

In practice, it is not feasible to fully automate the analysis process. In order to properly

up and refer to as sources for NEL, rather than millions or billions of instances of cultural heritage objects that could be hard to refer to globally. On one hand, encyclopedia-based and authority-file based LOD sources such as Wikidata and VIAF deal with the former and generate LOD by a top-down approach. On the other hand, Europeana takes a bottom-up data aggregation approach to build LOD for over 50 million digital objects from the records held by thousands of cultural heritage organisations. Most of them are unique and not well known. Next to their instance-level LOD, Europeana offers a highly limited amount of entity lookups relevant to their LOD that our study evaluates.

[18] For example, WorldCat does not seem to have entities 1976, 1979, and Europa League.

[19] This article only tries to identify the data about the same entity without judging if the data contents are duplicated or not. It seems that the double identity is a leftover of merging entities during data aggregation. Such examples include Aristotle in VIAF (https://viaf.org/viaf/268271999/ and https://viaf.org/viaf/7524651/) and California in YAGO (http://lod.openlinksw.com/describe/?url=http%3A%2F%2Fyago-knowledge.org%2Fresource%2FCalifornia)(last accessed 2021-01-26).

[20] During the entity identification process, we already recognise interesting patterns in the coverage of entities across the data sources. A typical case is the mosaic of availability for the objects and concepts. In the Getty Vocabulary, Ukiyo-e would be included as an artistic style, not an individual artwork, whereas Book of Kells, Garden of Earthly Delights, Sgt. Papers, Blade Runner, Uncle Tom's Cabin, and the King and I are not, because they are unique. Symbolically the latter group is all included in WorldCat, the Library of Congress, and VIAF as well as BabelNet, DBpedia, and YAGO. It seems to make sense to consider this pattern as the coverage difference between record-orientated library authority files and concept-orientated museum vocabularies.

[21] For example, unfortunately Italy in BebelNet has constantly returned HTTP 500 error during our analysis (http://babelnet.org/rdf/page/s00047705n).

document the data quality, it is required to search, identify, and verify the same entity across 11 data sources. The quality of each entity needs to be manually double checked. The main problem of our analysis is semantic disambiguation. It is even not always possible to accurately find an entity. For instance, the challenges of disambiguation and entity matching across multiple LOD sources are presented by Farag [76]. In our case, three reasons are worth mentioning: a) the lack of cross linking between data sources makes it hard to find all available entities, b) the entities are confusingly organised and hidden from the mainstream contents, especially in aggregated LOD, and c) the search functionalities on the website of the data sources may have limited capacity and have not been optimised. In these cases, lookups are executed on a best-effort basis[22]. Another justification of our manual evaluation is the lack of gold standard. In fact, the research on the LOD quality in digital libraries requires manual reviews for several metrics [55].

2.3.4 ANALYSIS METHODOLOGIES

In this study, we conduct both qualitative and quantitative analysis. As for the qualitative approach, we present some examples that are found during the manual inspection of LOD instances. As for the quantitative approach, we generate chord diagrams in R[23] to examine the basic flow of incoming and outgoing links within the 11 data sources. We deploy Data to Viz, based on the circlize package[24]. For the creation of traversal maps, we import matrix data from spreadsheets to R and generate network diagrams with igraph[25] packages. In addition, we calculate the amount and percentage of links and provide different views on the quality. Moreover, a basic network analysis is also conducted with R to objectively evaluate the characteristics of the small LOD network. It turns out that this approach is useful, because Guéret et al. [84] subsequently proposed a linking quality method with some of the network metrics we use in the R analysis.

Furthermore, we also analyse other data content (such as literals) in addition to the links. This is important, as we cannot obtain a full picture of link quality without studying the content of the link destination. In an RDF graph, there can be three types of nodes: IRIs[26], literals, and blank nodes[27]. As the blank nodes are not heavily used in our target datasets and add extra complexity, we limit ourselves to literals. For this purpose, first we simply extend our calculation to check the use of four W3C standardised properties, mainly for literals. The amount and percentage of rdfs:label, rdf:type, skos:prefLabel, and skos:altLabel are calculated. In addition, the total amount of content associated with rdf:resource and rdf:about is assessed. These two properties are at the centre of RDF/XML and are used to describe and connect resources. Although there are other important properties than the six properties described above, they are the most fundamental and frequently used properties to

[22] In addition, it is noted that this study does not guarantee technical feasibility of traversing via lookup services in reality. The project only documents and analyses the availability of links, not the validity of links. For example, it is the responsibility of LOD providers to adequately implement and maintain content negotiation and HTTP redirect.

[23] https://www.r-project.org/, last accessed 2021-01-26

[24] https://www.data-to-viz.com/graph/chord.html, last accessed 2021-01-26

[25] https://igraph.org/r/ last accessed 2021-01-26

[26] Internationalised Resource Identifier is the generalisation of URI that supports Unicode characters. For our convenience, URI is used as a synonym of IRI in our research.

[27] https://www.w3.org/TR/rdf11-concepts/, last accessed 2022-01-20

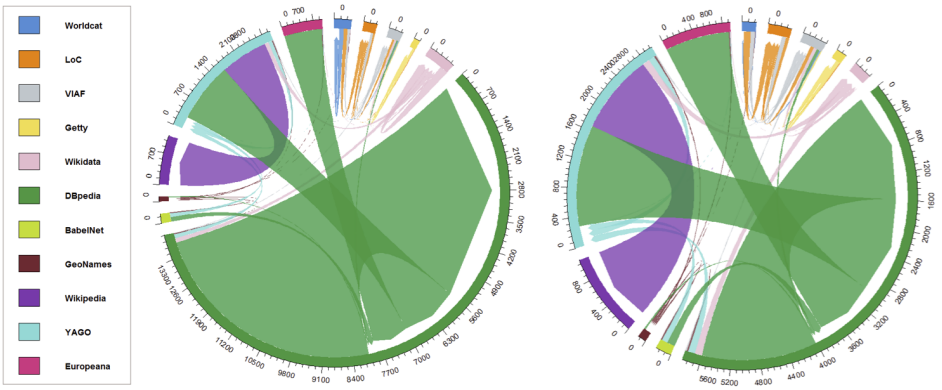

Figure 2.1: Chord diagram illustrating the amount of link flows between 11 data sources (left) and after removing inverse links (right)

describe entities. These statistics allow us to obtain basic holistic views on the data content. However, they are not sufficient to draw conclusions.

The challenge is how to objectively compare and evaluate the content quality of different LOD sources. The major problems are: a) there is no standard theory about what is regarded as high quality, and b) it is hard to evaluate the quality of semantics. In terms of a), for example, the number of links (edges) or labels/literals (strings) alone would not be able to indicate the data quality. In terms of b), the same hyperlinks and labels can be found in different context. For example, the link "http://www.example.com" can be found in skos:exactMatch or dcterms:isPartOf, while the string "Book of Kells" can be in skos:prefLabel or rdfs:label. Both of these cases carry the same information, but there is no easy way to assess the quality of semantics of the properties. This is especially the case when proprietary properties are used. It is practically impossible to judge the quality, due to the nature of freedom in LOD. Moreover, we cannot give any preference to a hyperlink or literal as the object of a property.

To minimise the impact of a biased evaluation, Python scripts[28] are developed to supplement our analyses. They compare the overlap of data content in each LOD source without any interpretations/assumptions. Technically this means that the scripts analyse the objects of the main entity with string matching, and calculate the amount of unique content. The objects include both edges and literals, where URIs are considered as string values to be compared. In other words, the semantics of the properties are not evaluated. Although this method may not be the most accurate way to measure the content quality, it allows us to perform systematic and automatic measurements. It provides us with a sense of the amount of information and the coverage or diversity of data contents.

It is anticipated that a broad mix of above-mentioned methods can provide new insights into the linking quality at different levels.

[28] Available at https://github.com/GO5IT/LOD_analysis and https://doi.org/10.5281/zenodo.5913595 including the data generated

Table 2.2: The total and average number of outgoing links (to the 11 data sources) held by the data sources

ID	A	B	C	D	E	F	G	H	I	J	K	
Source	YAGO	WorldCat	Wikidata	VIAF	Lib of Congress	Getty	GeoNames	European	DBpedia	BabelNet	Wikipedia	Total
Total	2713	259	192	171	102	69	23	903	5832	210	0	10474
Average	27.4	2.7	1.9	3.1	1.1	1.6	1	36.1	58.3	2.1	0	12.5

2.4 LINKED OPEN DATA ANALYSIS

2.4.1 OVERALL TRAVERSAL MAP

The first analysis starts with chord diagrams. Figure 2.1 primarily focuses on the number of links and their origins and destinations within the 11 data sources. The source data which produce Figure 2.1 is found in Appendix B.

The total number of links amounts to 10474. The dominance of DBpedia is obvious, occupying over 66.2% of the entire linkages (Figure 2.1 left). It is also noticeable that self-links significantly contribute to the volume of the links. YAGO supplies a substantial amount of links to DBpedia and Wikipedia. This results in the influential position of Wikipedia (5.2%), although it is not LOD. Surprisingly, Europeana comes fourth, despite the significantly limited amount of available entities (Appendix A). WorldCat, the Library of Congress, and VIAF somewhat share similar numbers of links. The outgoing and incoming links are unbalanced for Europeana.

From these numbers we can derive the following: the average number of links in all sources is 952.2, whereas the medians are 2.1 and 149 for both outgoing and incoming links. In fact, the amount of outgoing hyperlinks found in each source is moderate, given the entire size of those datasets (i.e. millions of triples); on average it is mostly under four links per entity (Table 2.2). These small figures are alarming, because this survey focuses on well-known sources often used for NEL for the cultural heritage datasets. It is clear that there is a great deal of room for improvement. Nevertheless, DBpedia, Europeana, and YAGO stand out, showing more promising quality for LOD with high number of links per entity.

When inverse traversals are removed from the statistics, the situation looks largely different (Figure 2.1 right). The sum of the links decreases to 6166. DBpedia loses an ample number of links (47.3%), whereas YAGO gains most (24.2%). Such a dramatic shift is an evidence of abundance of inverse properties described in DBpedia. If we scrutinise the data closely, we notice that this is mostly due to the inverse use of rdfs:seeAlso in DBpedia. For instance, the entity of Sweden contains:

dbr:Lund rdfs:seeAlso dbr:Sweden .

Figure 2.2 is the simplified overall "traversal map" for all data sources. It is a network diagram, illustrating all possible paths between the 11 data sources. However, since we observe a very high volume of links in DBpedia, YAGO, and Europeana, volumes and self-links/loops (i.e. links pointing to the same data source/domain) are not included in this figure. Thus, the diagram concentrates on the routes of traversals (i.e., the users' mobility and traversability).

It is clear that the traversing routes are not equally available across the data sources, and thus, it may be hard to navigate the LOD network. It is found that YAGO delivers four connections as well as one to Wikipedia. The next contenders are Europeana and DBpedia with four outgoing connections. In contrast, Wikidata has no outgoing connections[30]. Whilst

[30]192 self-links are omitted.

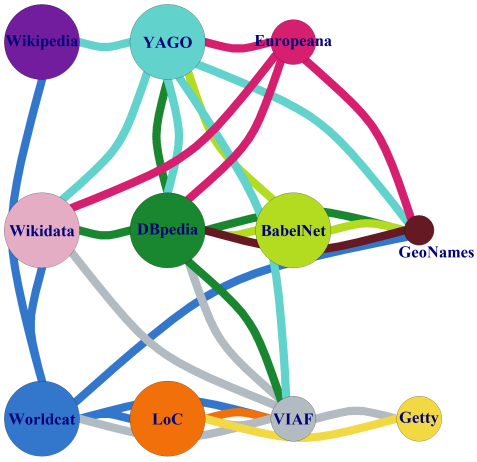

Figure 2.2: The overall "traversal map" shows available links/paths through four standardised properties between the 100 entities in 11 data sources (after self-links to the same domain is removed)[29]

GeoNames only links to DBpedia, the Library of Congress and Getty have one channel. With regard to incoming connections, GeoNames is an attractive destination to which five sources refer. Wikidata and DBpedia are also a centre of gravity, inviting five connections. On the other hand, Europeana and BabelNet receive no links. Whereas the lack of incoming links to BabelNet may be surprising, in Europeana's case it is not, because it is not equipped with a truly public lookup. This would mean that the generation of LOD dump and/or SPARQL endpoint may not be sufficient. It is best to publicly declare entities that are resolvable via lookups without access restrictions. WorldCat and Getty are both only reached by VIAF.

It is particularly remarkable that reciprocal links are quite rare. There are several nodes/vertices which can be reached via only particular edge(s)/path(s). This implies that network is not desirably populated by the standard properties, and that the users would not be able to efficiently obtain information through these properties. They need to follow the best paths to retrieve the identical or closely matching information. It is possible for data publishers to use other RDF properties, but it would be an irregular practice.

Idrissou et al. [95] stress that a full mesh (fully connected network) has the highest quality in their link quality metrics. When they compare different structures (e.g. ring, line, star, mesh, tree), the more a network resembles a fully connected graph, the higher the quality of the links in the network for all metrics (bridge, diameter, closure). One might argue that a full mesh is not necessarily a prerequisite of high data quality. This may be true for much LOD, however, let us remember that we focus on the most well-known data sources that many other LOD tend to link to. Therefore, it helps the connectivity of LOD on the web as a whole. Guéret et al. [84] use clustering coefficient and owl:sameAs chains as their criteria for high quality.

Figure 2.3 depicts traversal maps faceted by four link types. From now on, inverse properties are included but loops are excluded for the traversal map visualisation. Thus, the distortion of the "route diagram" that we avoided in Figure 2.2 is minimal. However, the rest of the statistics (matrix data and in the texts) include both inverse properties and loops, so

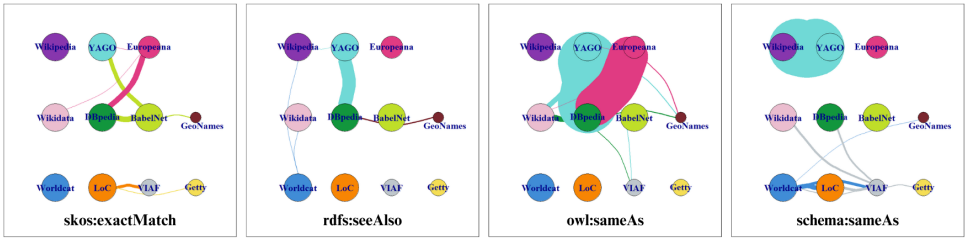

Figure 2.3: The overall traversal map by each standardised property (after removing self-links to the same domain)

that they reflect the actual situation.

Although we decided to avoid discussions on interpretations of link semantics, there is at least a clear difference between owl:sameAs (as well as skos:exactMatch and schema:sameAs) and rdfs:seeAlso. It can be clearly seen that Europeana, the Library of Congress, and BabelNet are the only data publishers using skos:exactMatch. rdf:seeAlso is used mostly by YAGO, while GeoNames and WorldCat are also visible. However, the proportions of owl:sameAs and schema:sameAs are higher. In particular, Europeana and YAGO provide a large amount of connections to either DBpedia or Wikipedia. We also realise that WorldCat and VIAF opt more for schema:sameAs. In general, Figure 2.3 suggests that the data creators made different ontological decisions on the choice of standardised properties. We will explore this further in the following sections.

2.4.2 AGENTS TRAVERSAL MAP

Figure 2.4 depicts the traversal map for agents. Appendix B includes the source matrix data and the traversal maps for all four properties. In general, agents have much less influence from loops than from other categories, because 72.4% of links are still present after removing recursive links, compared to the overall 42.0%. The most eye-catching result is Europeana. Especially, it uses owl:sameAs to link to DBpedia. In cultural heritage, VIAF plays a valuable role for agents as an aggregation of authority files of national libraries. For instance, it is the only source which offers four outgoing paths. This category has only three sets of nodes that have bilateral links.Therefore, segmentation is visible in the network and truly standardised LOD connectivity is limited.

In Table 3 in Appendix B, the role of DBpedia is expectedly prominent for incoming links, attracting 1555 links (80%). Unlike the outgoing links, Wikidata captures 121 referrals, making it the second highest source. Manual examination found that VIAF had only 72 incoming links, however, it contains more links which connect its entity to data sources outside the 11 sources, than any of the other sources. For instance, only four links with schema:sameAs are recorded for Beethoven. However, the destinations of a further eight links include the national libraries of France, Germany, Japan, Spain, and Sweden.

The amount of outgoing links held by 11 data sources in each entity is visualised in Table 2.3. When comparing the total amount in this table and in Table 3 in Appendix B, we notice that 1945 incoming links are received within the 11 data sources, out of 2412 outgoing links (80%)[31]. Whereas Europeana has 798 outbound links (33%), DBpedia and

[31] In the coming sections, we will compare outgoing links (the tables in the text) with incoming links (the overall

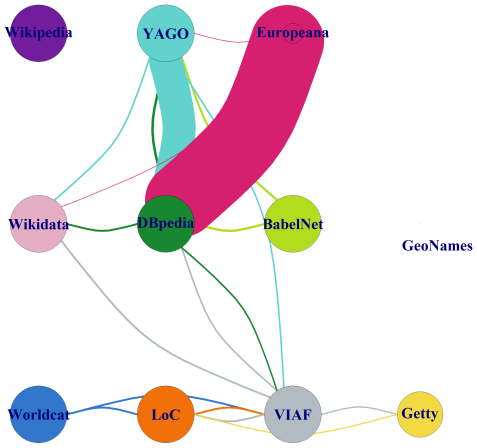

Figure 2.4: The overall traversal map for agent entities

Table 2.3: The amount of outgoing links that the 11 data sources hold in each agent entity (* means duplicate consolidation)

		Linneaes	Jesus	Aristotle*	Napoleon	Hitler	Caesar*	Plato	Shakespeare	Einstein	Elizabeth II	M Jackson	Madonna	Beethoven	Mozart	Benedict XVI	Alex Great*	Darwin	Obama	Mary	Q Victoria	SUM
A	YAGO	23	38	24	28	28	24	24	24	26	27	37	34	25	31	29	23	28	31	0	26	530
B	WorldCat	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	60
C	Wikidata	5	0	5	3	6	4	2	2	4	0	1	0	3	1	1	1	4	4	2	7	55
D	VIAF	12	3	11	12	12	20	10	12	12	10	13	11	12	12	11	13	12	11	6	12	227
E	LoC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
F	Getty	3	0	1	3	3	3	3	3	3	3	1	0	3	3	0	1	3	1	0	3	40
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	0	122	0	119	0	117	119	0	0	0	88	114	119	0	0	0	0	0	0	798
I	DBpedia	24	59	25	31	29	26	26	25	27	28	39	36	26	32	30	24	29	34	19	27	596
J	BabelNet	4	0	4	7	6	7	4	6	4	4	4	4	4	4	3	5	4	3	5	4	86
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	75	104	196	88	207	88	190	195	80	76	99	177	191	206	78	71	84	88	36	83	2412

YAGO follow at 596 and 530 respectively. There is a considerable gap between the highest number of outgoing links across 11 sources (Hitler, 207) and the lowest (Mary, 36). The highest cluster are from Europeana, however, the outgoing links in Europeana are unevenly distributed. Only Aristotle, Hitler, Plato, Shakespeare, Madonna, Beethoven and Mozart are present. This would offer evidence that art and cultural figures are more important for the cultural heritage objects that Europeana deals with than politicians and scientists. DBpedia and YAGO show a similar pattern, mainly due to the tight connections between them. In there, we observe relative popularity for Jesus, Michael Jackson, and Madonna.

WorldCat holds exactly three links per entity[32]. One is caused by the description of a new WorldCat identifier via the inverse property of rdf:seeAlso. The other two are schema:sameAs which links to the Library of Congress and VIAF. Similarly, the Library of Congress has exactly one link per entity (skos:exactMatch to VIAF)[33]. These two cases suggest evenly distributed and highly normalised RDF content, probably due to systematically generated links between the library sources.

Whilst most data sources cover all 20 agents, Jesus Madonna, Benedict XVI, and Mary are totally missing in Getty vocabularies. Similarly, the number of VIAF links is sharply reduced for Jesus and Mary. This is understandable since Getty ULAN and VIAF are typically orientated toward artists and authors in the context of libraries and museums, and religious figures are harder to be recognised as agent entities. Indeed, Jesus has the lowest number of links for five data sources (Mary for four data sources). As such, it is remarkable that Jesus is relatively high in DBpedia (59 links). It is also interesting that non-artists figures such as Einstein, Elizabeth II, and Obama are found in ULAN.

2.4.3 EVENTS TRAVERSAL MAP

Figure 2.5 clearly illustrates the lack of links. Bilateral links are extremely rare: only between YAGO and DBpedia. As a result, it is not possible, for example, to move from the Library of Congress to Wikidata. This implies that the entry point to a network determines the movement within it. DBpedia contains far more links than other sources. Although Europeana has only one entity in this category (i.e. World War I), it manages to draw a thick line (skos:exactMatch) in the figure (111 links).

In general, events were not found in VIAF during the manual data exploitation, however, it turns out that WorldCat and the Library of Congress refer to it seven times each. For example, the former links to the World Series in French (skos:prefLabel is Séries mondiales (Base-ball) and skos:altLabel is World Series (Base-ball)). Another 13 cases are all sporting events and awkwardly labelled as corporate entity in VIAF. Although those entities may be exceptional cases, they also reveal interesting cataloguing practices (or perhaps errors) by libraries in data modelling or mapping. Whatever the reasons are, we may face challenges in the future to tackle errors and inconsistency for semantic reasoning.

In terms of each entity (Table 2.4), the most appealing entity is World War II, followed by World War I and the Iraq War. Europeana's contribution to World Wart I is considerable. Although the EFL Cup is the lowest, the gaps between entities are relatively subtle except

tables in Appendix B)

[32]There is the forth link (rdfs:seeAlso). It is provided by not rdf:resource, but the anyURI typed literal, therefore, it is excluded from the analysis.

[33]skos:closeMatch is excluded from the analysis.

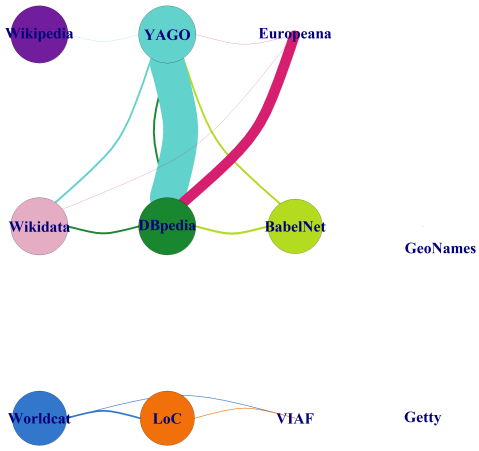

Figure 2.5: The overall traversal map for event entitiesentities

the top three (i.e., median 49.5, average 57.5). The principal reason for the prominence of DBpedia for the World War II is rdfs:seeAlso inverse links which include the DBpedia entities of agents (e.g. Winston Churchill), places (e.g. Leipzig), ships (e.g. USS Hornet), and the lists and articles derived from Wikipedia (e.g. tanks in the German Army, history of propaganda). In this case it is advantageous for the users to discover and access detailed information about the war. However, as RDF representation is not guaranteed for rdfs:seeAlso, this situation would hamper predicting the source of link destination and decreasing the possibility of efficient and/or automatic data processing.

2.4.4 DATES TRAVERSAL MAP

It is striking that the volume of links is very low (Figure 2.6). Out of 881 outgoing links, 863 links are consumed within the 11 data sources, implying a high level of closure in the network. In addition, only three sources are referenced: DBpedia, Wikidata, and the Library of Congress. Although YAGO provides many links to DBpedia and Wikidata via owl:sameAs, it does not receive any incoming links. Since bilateral links do not exist, the movement in the network is highly restricted. There are only three possible paths. Consequently, the fluctuation of linking patterns is also minor (Table 2.5).

The economy of the creation of date entities may show serious issues. 1978, 1979, and 1976 do not seem to exist in YAGO, the Library of Congress, and WorldCat, while other consecutive years in the 1970's are available (see Appendix A). Such inconsistency would become problematic, when queries are constructed to look for answers to research questions on years and periods. In semantic queries, erroneous links and data omissions require careful presentation to LOD users in the future, in order to avoid misinterpretation and misjudgment.

One reason for this phenomenon is the lack of recognition and/or needs for numeric date instant entities, in comparison with other date representations, including textual dates (e.g. "End of the 17th century"), numeric durations (e.g. "1880-1898"), and periods and eras (e.g. "Bronze Age" and "Roman Republic"). For example, a quick search indicates the entity for "Neolithic" exists in all our data sources except GeoNames, VIAF, and Europeana.

Table 2.4: The amount of outgoing links that the 11 data sources hold in each event entity (* means duplicate consolidation)

		WW II	WW I	World Series	War of 1812	Vietnam War	Super Bowl	Stanley Cup	Spanish Civil War	Olympic Games*	Korean War	Iraq War	Gulf War	FACup	Europa League	EFLCup	Cold War	Champions League	American Rev War	American Civil War	Academy Awards	SUM
A	YAGO	21	24	21	19	33	23	18	22	25	22	34	23	26	21	16	21	23	20	20	22	454
B	WorldCat	2	2	3	2	2	3	3	2	3	2	2	2	3	0	3	1	3	2	2	2	44
C	Wikidata	3	4	0	0	3	2	1	1	2	2	3	1	2	2	0	3	1	1	3	2	36
D	VIAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	LoC	2	2	1	2	2	1	1	2	1	2	2	2	1	0	1	2	1	2	2	2	31
F	Getty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	113
I	DBpedia	122	61	23	28	39	24	19	26	23	25	41	29	27	22	17	29	24	23	28	24	654
J	BabelNet	4	7	3	4	4	3	3	0	9	4	3	4	3	3	3	4	3	7	7	4	82
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	154	213	51	55	83	56	45	53	63	57	85	61	62	48	40	60	55	55	62	56	1414

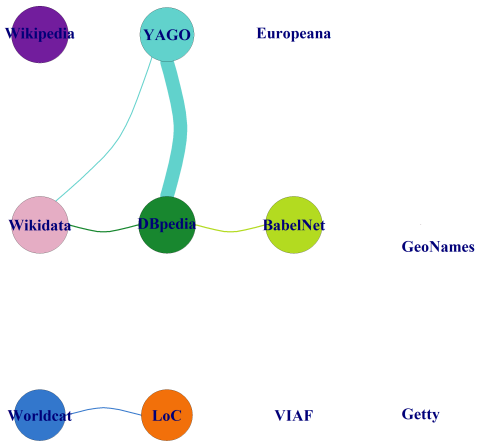

Figure 2.6: The overall traversal map for date entities

Table 2.5: The amount of outgoing links that the 11 data sources hold in each date entity

		1987	1986	1985	1983	1980	1984	1982	1968	1979	1969	1978	1967	1981	1977	1976	1975	1964	1966	1965	1960	SUM
A	YAGO	13	13	13	13	13	13	13	13	13	13	0	13	13	13	13	13	13	13	13	13	247
B	WorldCat	2	2	2	2	2	2	2	2	0	2	2	2	2	2	0	2	2	2	2	2	36
C	Wikidata	4	3	3	4	3	4	4	1	2	2	2	3	3	3	3	2	4	4	3	3	60
D	VIAF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	LoC	2	2	2	2	2	2	2	2	0	2	2	2	2	2	0	2	2	2	2	2	36
F	Getty	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	GeoNames	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	Europeana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	DBpedia	26	24	25	25	26	27	26	28	23	21	23	20	25	26	26	25	24	23	20	19	482
J	BabelNet	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	48	45	46	47	47	49	48	47	39	41	30	41	46	47	43	45	46	45	41	40	881

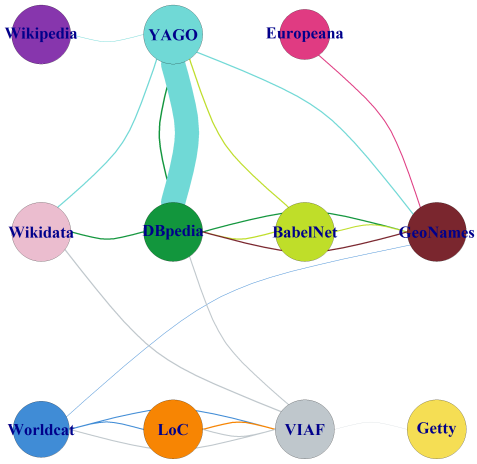

Figure 2.7: The overall traversal map for place entities

In cultural heritage, numeric dates are often stored in a database as string/literal data type, when encoded in XML or RDF. They can be typed as date in the XML Schema (e.g. xsd:date). Thus, they are not designed for NEL, although it would have many advantages, especially for data linking and integration. What is clear is that users have currently a very limited possibility to execute NEL for numeric dates. To fill this gap, we have recently started a project to create LOD for the numeric date entities [124].

2.4.5 PLACES TRAVERSAL MAP

Traversability for places is better than in other categories. YAGO dominates the scene for outgoing links (Figure 2.7). Interestingly VIAF comes third despite its focus on agent entities. The Library of Congress, Getty TGN, and GeoNames contain an almost consistent number of links, each typically pointing to DBpedia. Users need to be careful regarding Europeana, because it does not provide the entities for the USA at all (USA, California, New York, and New York City). This type of inconsistency may be problematic for NEL implementers. They should scrutinise the occurrences of their place entities in their local datasets before selecting the right NEL targets. Strangely, no outgoing links are found for Australia, Canada, France, Germany, Italy, Japan, and Russia in Wikidata.

The presence of GeoNames, in particular, facilitates more fluid movements in the network. Although Ahlers [39] claims that it is the largest contributor to geospatial LOD and is intensely cross-linked with DBpedia, it is a disadvantage that it only connects to DBpedia. This makes the overall mobility less ideal.

Apart from a link to VIAF, Getty TGN only contains 20 self-links mostly in the form of rdfs:seeAlso for a HTML representation. RDF/XML for New York City (tgn:7007567) holds:

tgn:7007567 rdfs:seeAlso `<http://www.getty.edu/vow/TGNFullDisplay?find=&place=&nation=&subjectid=7007567>` .

Therefore, it is a dead end in terms of network traversals, of which the users need to be

Table 2.6: The amount of outgoing links that the 11 data sources hold in each place entity

		USA	UK	France	England	Germany	Canada	Australia	Japan	Italy	Poland	India	Spain	London	Russia	New York City	Brazil	California	New York	Netherlands	Sweden	SUM
A	YAGO	43	37	40	34	35	31	34	33	35	34	35	44	32	35	31	39	35	31	45	46	729
B	WorldCat	4	4	4	3	3	4	4	4	3	3	4	3	4	3	0	3	3	3	3	3	65
C	Wikidata	4	1	0	5	0	0	0	0	0	4	2	4	2	0	2	3	2	5	5	2	41
D	VIAF	11	7	7	5	8	9	11	10	10	7	11	10	11	8	11	6	10	10	9	4	175
E	LoC	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
F	Getty	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
G	GeoNames	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	21
H	Europeana	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	17
I	DBpedia	52	49	469	131	481	253	266	187	314	417	238	164	63	275	78	133	208	231	45	121	4175
J	BabelNet	1	7	6	6	6	6	7	6	0	6	6	6	9	6	6	6	9	7	6	6	118
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	119	109	531	189	537	307	326	244	366	475	300	235	125	331	131	194	270	290	117	186	5382

aware during their traversing. Europeana is disappointing including only 17 outgoing links only to GeoNames.

If loops are included, DBpedia holds 86% of all outgoing links. This is caused by a vast number of inverse links. For example, in case of Australia, 255 out of 266 outgoing links in DBpedia are those inverse rdfs:seeAlso links to DBpedia itself. It is possible to find both important and less important links:

dbr: Health_care_in_Australia rdfs:seeAlso dbr:Australia .

On one hand, the DBpedia loops may be confusing, especially due to the use of ambiguous rdfs:seeAlso links and the flexibility of information provided. On the other hand, they allow users to find unexpected related information that other LOD sources do not provide, leading to the serendipity that LOD is good at.

In Table 2.6, the lowest entities are surprisingly: the Netherlands, United Kingdom, and United States. This is chiefly attributed to fewer numbers of DBpedia links. However, the reason for this is unclear. On the contrary, the top entities receive a large quantity of links, which include Germany, France, and Poland.

The outgoing links are the lowest for United Kingdom, followed by the Netherlands, and United States. In contrast, Poland, Germany and France are the top three. The cause is obvious: the numbers are affected by the uneven pattern of links in DBpedia. The amount of links in other sources are instead more or less evenly spread across different entities. It would be intriguing to investigate the reasons by inspecting the corresponding entities in Wikipedia articles and the linking mechanism behind the DBpedia transformation. It would reveal pros and cons of a crowdsourcing approach to LOD, as opposed to authority approach such as the Library of Congress, VIAF, and Getty from libraries and museums.

2.4.6 OBJECTS AND CONCEPTS TRAVERSAL MAP

Objects and concepts are the subject matter in which cultural heritage researchers would be most interested. To a large degree, they are the target entities of contextualisation which is substantiated through data integration and inferences, thus, the contextualised entities are out of our scope. Rather we analyse them as the entities supporting contextualisation (Figure 2.8). 1844 outgoing links are recorded of which 91% are bounded for the 11 data sources. Network closure also persists in this category. 81.3% of all incoming links concentrate on Wikipedia (1085), with DBpedia (100) and Wikidata (43) lagging far behind. The same can be said

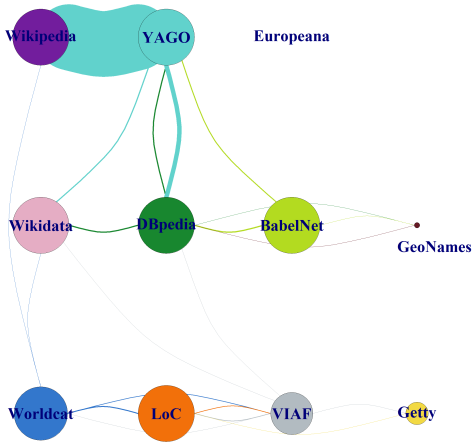

Figure 2.8: The overall traversal map for objects and concepts entities

for outgoing links: YAGO(1212) and the rest. This happens, because YAGO provides a considerable number of links to Wikipedia.

Although Europeana produces LOD out of digital cultural heritage objects, its entity API is merely an experimental reference point, thus, no contribution is observed in our traversal scenarios. Interestingly, VIAF plays an authoritative role for this category. It serves a small number of links to five sources. Although the number of outgoing links from BabelNet is not high, it performs better in this category. During the process of identifying and collecting the entities, some data quality issues are recognised. The significant concepts of cultural objects in FRBR, namely Work, Manifestation, Expression, and Item, are not easily conceptualised and encoded in the LOD observed. For example, taking a book as an example, we consider a single physical copy of a book as Item. Then, all published copies of the book which share the same ISBN are defined as Expression. Manifestation is considered as a book in a specific language by a specific author, whereas Work is a higher level of abstraction to cover the idea or the fundamental creation of the book by an author. Therefore, for instance, VIAF holds records on The King and I as Expression (motion picture) and Work (the original artwork). However, partly due to the technical mechanism of VIAF, Work may not be easily created. Similarly, Wikipedia has a disambiguation page for the King and I to distinguish the original musical from films and music products associated to the musical. This implies some difficulties in terms of co-reference resolution during NEL, as well as graph traversing.

As this category is deliberately broad and vague in principle, it is not possible to see clear-cut results. For example, GeoNames has entities for Palazzo Pitti and Angkor Wat, which could be classified as places and object simultaneously. Nevertheless, it reminds us that the data modelling for cultural heritage entities is intentionally complex. There could be entities that have multi-types. Depending on the perspective, the data modelers and users would need to find a common view on both practical use and theoretical truth and/or fuzziness of datasets. For instance, Palazzo Pitti could be a geographical place, as well as a building structure, concept, or organisation. However, complicated roles may introduce unnecessary complexity for real usage, confusing end users.

Table 2.7: The amount of outgoing links that the 11 data sources hold in each object and concept entity

		Book of Kells	Vasa	Garden of E Delights	Rosetta Stone	Palazzo Pitti	Boeing 747	Sgt. Pepper's	Tosca	Blade Runner	Uncle Tom's Cabin	Ming Dynasty	Ukiyo-e	Angkor Wat	Toraja	Byzantine Empire	Mars	Tamil language	Influenza	King and I	Like a Rolling Stone	SUM
A	YAGO	52	50	44	88	55	77	18	53	65	8	113	59	112	31	153	23	161	0	10	40	1212
B	WorldCat	3	3	3	3	3	2	3	3	3	3	0	4	2	2	3	2	2	2	3	3	52
C	Wikidata	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	VIAF	5	2	4	3	8	0	4	2	5	5	1	0	4	0	3	2	0	0	6	2	56
E	LoC	1	1	1	1	1	2	1	2	1	1	2	2	2	2	1	2	6	3	2	1	35
F	Getty	0	0	0	0	2	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	9
G	GeoNames	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	3
H	Europeana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	DBpedia	17	16	16	19	18	20	21	21	21	19	23	19	20	12	39	25	25	24	11	20	406
J	BabelNet	3	3	3	4	4	3	3	3	3	3	5	3	4	3	5	5	5	4	2	3	71
K	Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	SUM	81	75	71	118	92	104	50	84	98	39	145	88	147	50	205	60	200	34	34	69	1844

Another interesting finding is that Mars appears in TGN of the Getty vocabulary. It is normally considered that the vocabulary contains place names on earth, as one expects from GeoNames. There could be some surprise for LOD users in terms of how data is conceptualised and modelled, and from where data is obtained, especially when automatic data collection and integration are implemented in the future.

Regarding the individual entities (Table 2.7), Byzantine Empire and Tamil language in YAGO display a distinct pattern. The cause of this pattern seems to be clear; it includes links to language orientated resources such as language codes, maybe suggesting an important role of language resources in the LOD scenario. For other entities in YAGO it is hard to find exact causes and correlations between the entities with more links (Rosetta Stone, Ming Dynasty, Angkor Wat) and the ones with fewer links (Uncle Tom's Cabin, Influenza, King and I). The results from Getty imply the exclusion of specific objects.

2.4.7 NETWORK ANALYSIS

We deploy a network analysis using R to supplement the so far relatively subjective impressions and interpretations of the traversal maps (Table 2.8). Although the work of Idrissou et al. [95] is highly relevant here, unfortunately we are unable to use their metrics, because they are based on undirected weighted graph with link strength (confidence scores). As seen in the traversal maps, reciprocity is generally low. The unavailability of bilateral links are obvious for dates and events. Mean distance is short, mostly under 2.0. Diameter is the length of the longest geodesic. We have rather short diameters, implying connections are limited within a small circle. Edge density is the ratio of the number of edges and the number of possible edges. Here we observe low density.

In addition, centrality is calculated, using three methods: Closeness (in and out), Eigen Vector, and Betweenness (Figure 2.9). The Closeness statistically suggests the LOD hubs of outgoing and incoming links. The overall Closeness is similar across 11 sources. However, the contrast between Wikidata and Wikipedia as an incoming source and BabelNet and Europeana as an outgoing source can be observed. It is rather unexpected that there are no big differences between the sources for the centrality by Eigen Vector. Thus, the dominance

Table 2.8: Network analysis measurements by category

Measurement	Overall	Agents	Events	Dates	Places	Objects &Concepts
Reciprocity	0.345	0.316	0.154	0.000	0.381	0.381
Transitivity	0.505	0.600	0.692	0.600	0.420	0.447
Mean Distance	1.919	1.791	1.235	1.167	1.826	1.878
Diameter	4	4	2	2	4	4
Edge Density	0.264	0.173	0.118	0.045	0.191	0.191

of DBpedia (and to a less extent YAGO) is not clearly visible in the chart. VIAF and DBpedia seem to sit in-between position, mediating the linking flows. Moreover, a radar chart (Figure 2.10) shows the indicator by R for the roles of vertex. The vertex is called a "hub" if it functions as a node to hold many outgoing edges, while it is called "authority" if it serves as a node to attract many incoming edges. Whereas YAGO, WorldCat, and Europeana are hubs, Wikidata and GeoNames are authorities. DBpedia has both characteristics, and is, therefore, a strong influencer for the analysed LOD sources.

Generally speaking, the overall situation shows a mosaic of segmentation even in a small LOD cloud. It is far from a full mesh network, if not data silos, which LOD is supposed to resolve. Our result simultaneously indicates a couple of tightly connected LOD clusters at best. Thus, it is currently hard to implement automatic traversals among the datasets without studying non-standardised properties (i.e. ontologies) and traversal maps.

2.4.8 CONNECTIVITY AND LINK TYPES IN DETAIL

In order to better understand the overall connectivity of LOD datasets, we additionally generated more segmentation and detailed statistics.

Figure 2.11 illustrates how close the 11 data sources are connected to each other through four standardised properly links. It displays the ratio of the hyperlinks bounding for the domains of the 11 datasets. Thus, it should represent the openness or closure of this small network. A high level of exclusivity for our data sources is observed. On average, 87.8% of links are within the 11 dataset boundary. Except Wikipedia, VIAF remains the lowest source in terms of links to the other datasets, but still holds over 37.3%. The statistics clearly indicate the closed and close connections of the 11 data sources in terms of standardised traversability.

When combing with analysis in the previous sections, this closure and the homogeneity and centrality of the 11 datasets are a worrying sign in the sense that the users of 11 datasets are not able to identify and explore new and unknown datasets beyond those giants of LOD, hampering serendipity for users' research. This phenomenon would also decrease the diversity of the LOD cloud. Our analysis indicates that the identical entities in local cultural heritage datasets cannot be effectively connected to each other through NEL via the 11 global LOD sources. Data integration and/or contextualisation would only be possible if the users know the connectivity of datasets in advance and conduct a federated SPARQL query at known endpoints.

In fact, Ding et al. [69] note that the typical size of sameAs networks either remains a small constant or increases slowly, and that single central resources are connected to a number of peripheral resources. This condensed view of LOD is adequately depicted in

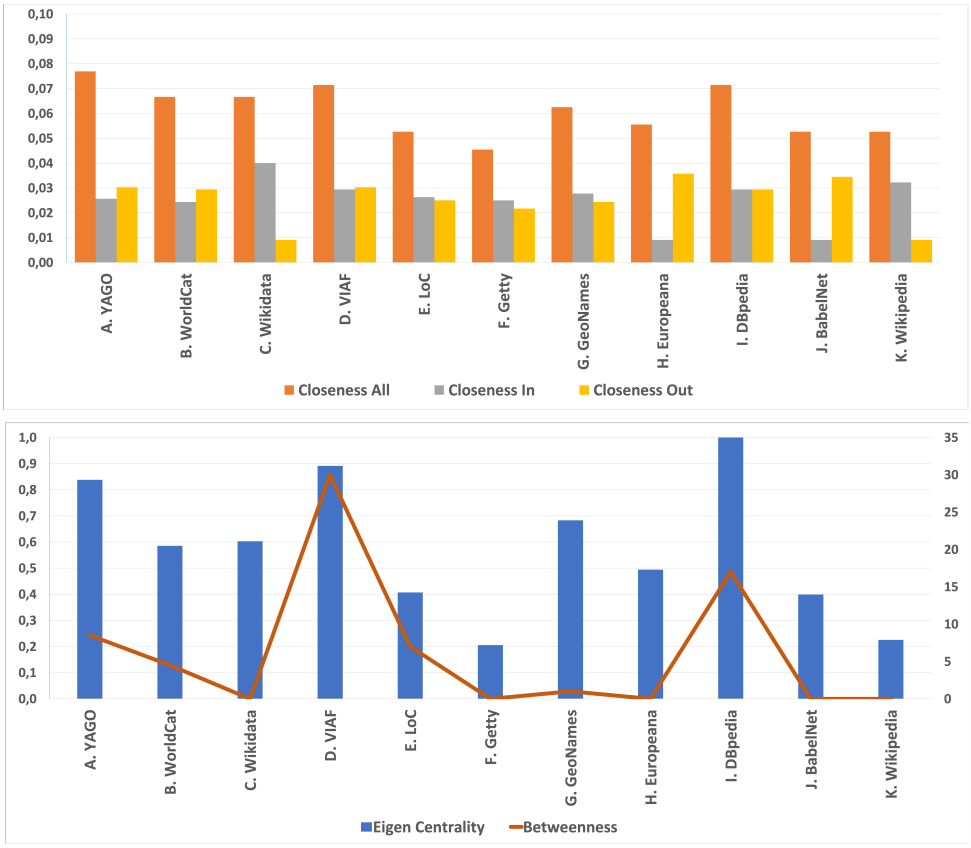

Figure 2.9: Closeness (above) and centrality (below, left Y axe) and betweenness (below, right Y axe) for 11 data sources

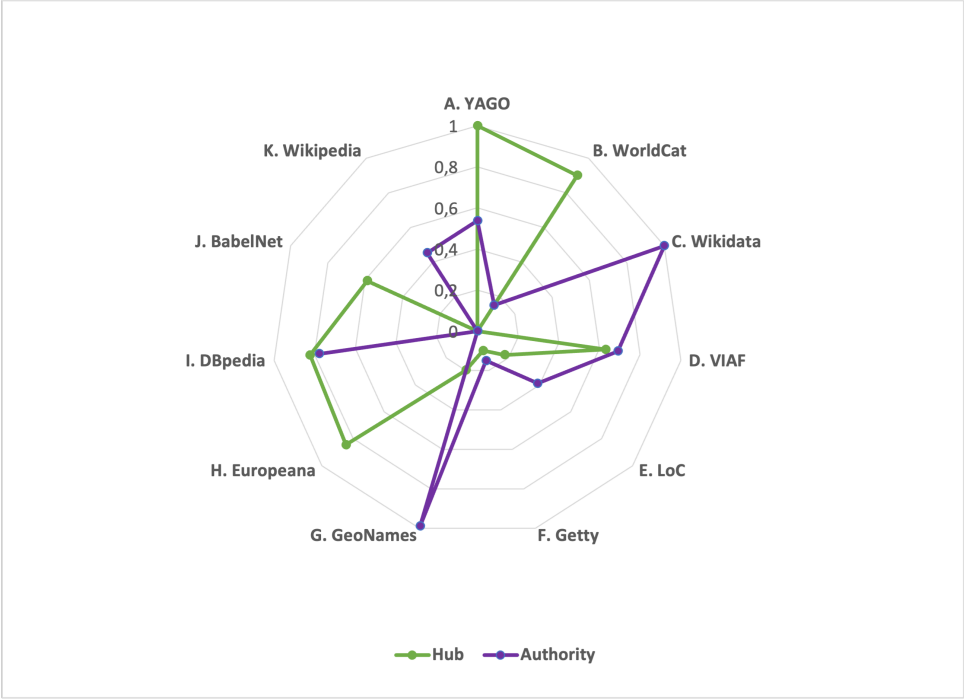

Figure 2.10: Indicator by R if a data source is authority or hub

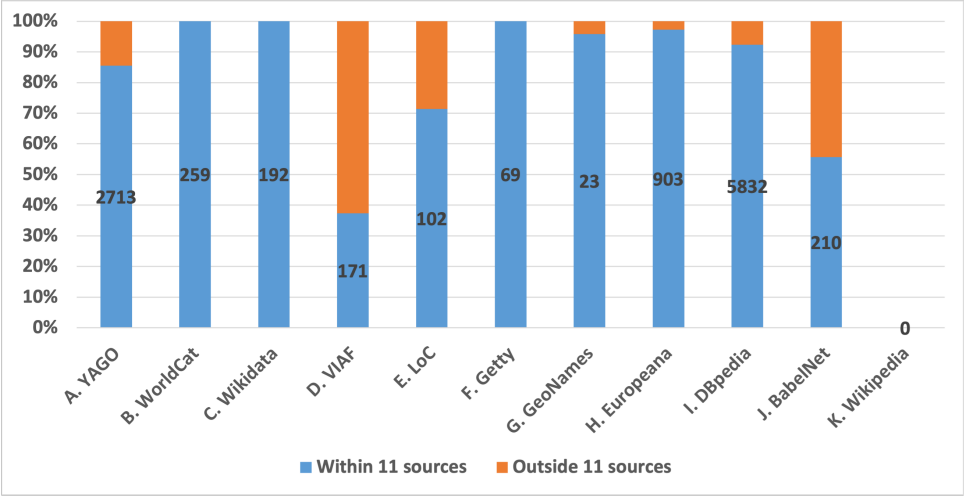

Figure 2.11: The ratio of the four standardised property links going within and outside 11 data sources

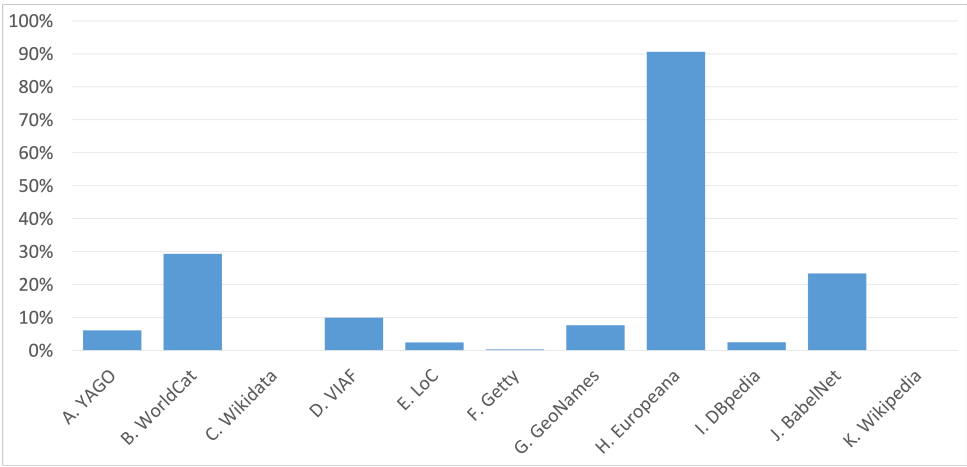

Figure 2.12: The percentage of four srandardised properties used for the purpose of rdf:resource linking in 11 data sources

their cluster analysis and visualisation, where a few LOD data sources investigated in this research are clearly seen as in-degree or out-degree hub nodes such as DBpedia, GeoNames, Wikipedia, and WorldCat. Correndo et al. [56] also report a power-based LOD network. Moreover, recent research discovers two high-centrality nodes (DBpedia and Freebase) and domain specific naming authorities/hubs such as GeoNames among others [45]. The added value of our study is to reveal the extent of this phenomenon for four different properties at an instance level.

Now, let us take a close look at link types. Figure 2.12 presents the percentage of the four standard properties used within rdf:resource. In RDF/XML, rdf:resource is the property to indicate the URI of the object node in a graph[34]. In this sense, it should normally contain all the outgoing links. By dividing the ratio of the four properties, we can highlight the balance between them and other properties including proprietary ones.

The overall percentage is, unsurprisingly, low because the four properties are normally a small part of RDF content. Nevertheless, the range varies from 30% to close to 0%. An exception is Europeana. 90.6% of links use them, demonstrating a high conformity to the standardised RDF properties and highly limited use of proprietary properties. The result suggests relatively high importance of the four properties in the WorldCat and BabelNet datasets. In contrast, Getty vocabularies and Wikidata use other properties almost exclusively. Indeed, a query on WDProp[35] lists 8732 unique properties in Wikidata as of 26 January 2021. A manual examination of Wikidata entities further justifies the outcome: the properties are organised by its proprietary wdt: with P prefix, while wdtn: is the entities with Q prefix[36].

For example, the entity of France contains 9500 rdf:resource, while wdt: is used 294 times with rdf:resource. 7292 rdf:type are included in combination with rdf:resource. The W3C properties of our concern are not available at all. owl:sameAs only appears occasionally

[34]https://www.w3.org/TR/rdf-syntax-grammar/, last accessed 2021-01-26
[35]https://rawgit.com/johnsamuelwrites/wdprop/master/index.html, last accessed 2021-01-26
[36]https://www.wikidata.org/wiki/Wikidata:SPARQL_tutorial, last accessed 2021-01-26

to provide inverse relations for obsolete (mostly duplicate) properties that offer redirects. Erxleben et al. [74] explain that Wikidata is keen to faithfully represent the original data using the language of RDF and linked data properly. In particular, they claim that owl:sameAs would often not be justified to relate external URIs to Wikidata. This leads to their hesitation to use this property as well as to include links to many external data.

On the one hand, proprietary properties in Wikidata enable the users to refine the semantics of outbound links. It is useful in some cases where one needs to identify a particular link among tens of owl:sameAs links. On the other hand, they make it more difficult to automate graph traversals, when used with other LOD. In addition, there is a question of manageability and usability. As the outgoing link properties can be suggested by the users, the number of the properties could grow sharply. Then, the complication of selecting them will be amplified.

Another issue is that the Wikidata entities do not use human "guessable" URIs, even if they are not absolutely opaque URIs such as hash. For instance, the syntax of the entity URI for Cold War is https://www.wikidata.org/entity/Q8683. They are agnostic about their semantics and are language independent, which prevents human users from guessing the meaning of properties and/or hacking the URIs[37] without examining the ontology behind. We should recognise that self-describing URIs are rated high for the quality metrics of Candela et al. [55].

When we manually examined France in Getty, we found that there were 1783 rdf:resource. 1349 SKOS properties are used among which 10 skos:prefLabel, 18 skos:altLabel, and 1246 skos:narrower are present. Whereas 251 Dublin Core Metadata Terms (dct:)[38] and 202 Getty Ontology (gvp:)[39] are in use, 60 PROV (prov:)[40] and 56 SKOS-XL (skosxl:)[41] are also found. Although not all properties use rdf:resource, the figures provide us a clue about the relation between linking and property usage.

Figure 2.13 illustrates the ratio of each property among the four properties. Despite the wide spread of research concerning owl:sameAs, its use for outgoing links is less than the majority for all outgoing links (42.2%). While 38.4% use rdfs:seeAlso, schema:sameAs and skos:exactMatch are in the minority. As GeoNames provides the link to DBpedia with rdfs:seeAlso, the equivalent identity cannot be inferred. skos:exactMatch is present in BabelNet, Europeana, Getty vocabularies, and the Library of Congress. VIAF exclusively uses schema:sameAs, whilst more than half of WorldCat entities are described with it. YAGO also uses it for more than one third of its entities. However, its use is debatable, since the schema.org ontology is not a W3C recommendation[42]. Moreover, Beek et al. [45] point out that it is semantically different from owl:sameAs.

From Figure 2.12 and 13, it becomes clear that some data providers set different strategies to design their ontologies in spite of the W3C recommendations. The results indicate that it is not feasible to traverse LOD and collect information, if the users specify only one type of

[37] In this context, hacking means the manipulation of URIs to access another data, for example, by changing prefix or suffix. See also http://www.jenitennison.com/2009/07/25/opaque-uris-unreadable-uris.html, last accessed 2021-01-26

[38] https://www.dublincore.org/specifications/dublin-core/dcmiterms/, last accessed 2021-01-26

[39] http://vocab.getty.edu/ontology, last accessed 2021-01-26

[40] https://www.w3.org/TR/prov-o/, last accessed 2021-01-26

[41] https://www.w3.org/TR/skos-reference/skos-xl.html, last accessed 2021-01-26

[42] https://schema.org/docs/howwework.html, last accessed 2021-01-26

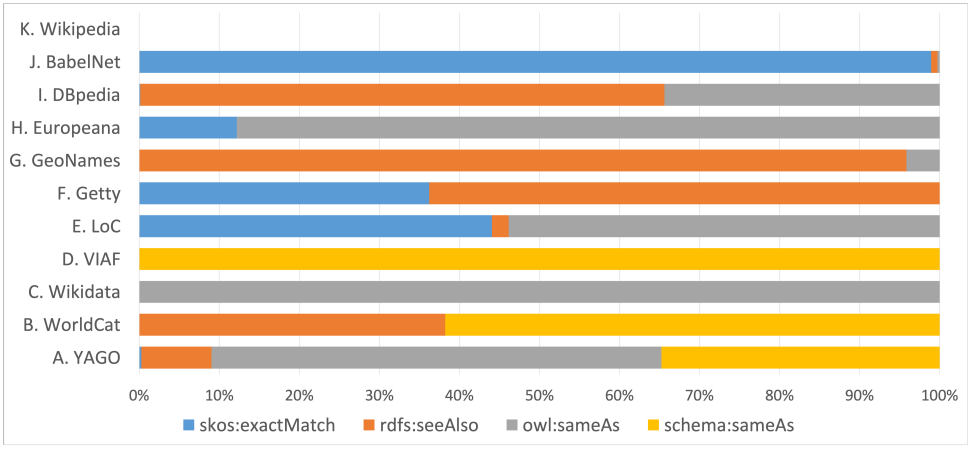

Figure 2.13: The ratio of each property among the four standardised properties used in 11 data sources

property. As seen throughout Section 2.4, the need of traversing strategies is also verified from this perspective.

2.4.9 LITERALS

This section examines the quality of other data content to supplement the analysis of link quality. The content-related four W3C standard properties are analysed, namely, rdfs:label, rdf:type, skos:prefLabel, and skos:altLabel. Figure 2.14 shows the ratio of each property among the four properties used in the 11 data sources.

Here one can also observe the characteristics of data sources. The contrast between rdfs:label and SKOS vocabularies is one focal point. Interestingly BabelNet prefers to use the former this time, in place of the latter. It is noted that GeoNames only uses rdf:type, primarily because it employs proprietary properties for the name of places (gn:):

`<https://sws.geonames.org/6251999/>` gn:name "Canada"; gn:alternateName "Kanuadu"@olo .

The library sector (VIAF, the Library of Congress, and WorldCat) uses skos:altLabel extensively Generally speaking, it is evident that the use of properties is diverse and not standardised. Therefore, automatic retrieval of basic information such as entity labels would require good understanding of each data source before data processing begins.

We further investigate the core constructs of RDF/XML. The use of rdf:resource and rdf:about is analysed. The average amount of rdf:resource, rdf:about, and literals is shown in Table 2.9. In general, contrast is clearly visible between the data providers with a high volume of content (Wikidata, YAGO, DBpedia) and the rest. Somehow Getty has competitive numbers. We are also curious about the low average of 1.1 for rdf:about in YAGO. When we had a close look at the dataset, we discovered that it used a single instance of rdf:about for the entity itself, for example, as follows:

<rdf:Description rdf:about="http://dbpedia.org/resource/World_War_II"> </rdf:Description>

Similarly, each entity in GeoNames contains it exactly twice (2.0 for rdf:about):

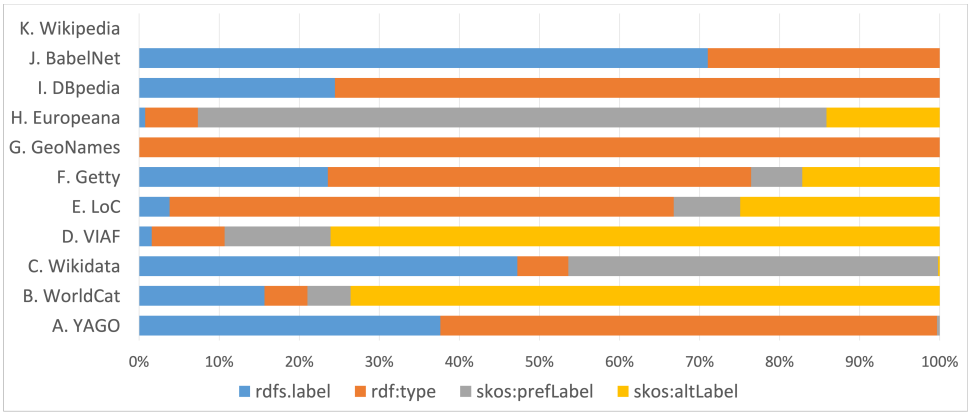

Figure 2.14: The ratio of each content-related property among the four content-related properties used in 11 data sources

Table 2.9: The average number (per entity) of rdf:resource, rdf:about, and literals for each data source

ID	A	B	C	D	E	F	G	H	I	J	K	
Source	YAGO	WorldCat	Wikidata	VIAF	Lib of Congress	Getty	GeoNames	Europeana	DBpedia	BabelNet	Wikipedia	Total
rdf:resource	530.6	9.3	4696.1	84.1	62.2	595.4	14.3	41.0	2546.5	16.3	0.0	8595.9
rdf:about	1.1	8.1	2164.8	27.5	93.5	73.6	2.0	1.3	2285.8	5.9	0.0	4663.7
Literals	105.2	43.7	50723.9	448.1	230.5	207.2	176.1	138.1	82.6	2.9	0.0	52158.3

<gn:Feature rdf:about="http://sws.geonames.org/2077456/"></gn:Feature><foaf:Document rdf:about="http://sws.geonames.org/2077456/about.rdf"></foaf:Document>

The second rdf:about preserves the technical metadata about the entity such as a Creative Commons license and creation date.

Moreover, we investigate the amount of literals. However, they have to be treated carefully, as they may include less relevant information about the entity. Despite the caveats, the figures do provide a rough idea of how much content is described in each LOD instance. Manual inspection indicates that the number of literals in some LOD is extremely high. This is not only due to an enormous amount of technical metadata, but also to repetitions (e.g. literals expressed in several schemas) and language variations in them. For example, there are in total over 4.5 million literals and, on average, more than 50 thousand for the 100 entities in Wikidata.

2.4.10 CONTENT COVERAGE

This section presents our attempt to further enhance the results of Section 2.4.9. Our Python scripts compare the content differences of the 100 instances across the 11 data sources (see Figure 1 in Appendix C). The amount of unique content of a single entity and the ratio are automatically calculated, and the aggregated view for the 11 data sources is shown in Table 2.10. In theory, they should represent the coverage and diversity of content (for a data source). The table is grouped by categories (i.e. all entities within are aggregated), because the instances tend to show similar patterns within the same category. “Full coverage” indicates the total amount of the unique content that 11 data sources hold as a whole (thus 100% coverage). It means that overlapping content is calculated once. The percentage of a

Table 2.10: The number of unique data content per data source in each category (values in parentheses indicate coverage in percentage)

ID	Data Source	Overall	%	Agents	%	Events	%	Dates	%	Places	%	Objects& Concepts	%
A	YAGO	251293	56.2	19201	34.2	24227	55.0	418	0.9	202215	72.5	5232	24.2
B	WorldCat	3667	0.8	886	1.6	346	0.8	276	0.6	1876	0.7	287	1.3
C	Wikidata	69183	15.5	18944	33.8	6708	15.2	4074	8.8	34068	12.2	5389	24.9
D	VIAF	11207	2.5	5695	10.2	0	0.0	0	0.0	4702	1.7	810	3.7
E	LoC	11980	2.7	2587	4.6	2253	5.1	774	1.7	4997	1.8	1369	6.3
F	Getty	23894	5.3	1605	2.9	0	0.0	0	0.0	21783	7.8	506	2.3
G	GeoNames	3284	0.7	0	0.0	0	0.0	0	0.0	3200	1.1	84	0.4
H	Europeana	3746	0.8	1375	2.5	256	0.6	0	0.0	2115	0.8	0	0.0
I	DBpedia	128307	28.7	20212	36.0	20003	45.4	40951	88.1	35469	12.7	11672	53.9
J	BabelNet	1866	0.4	359	0.6	345	0.8	358	0.8	497	0.2	307	1.4
K	Wikipedia	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
-	**Full Coverage**	**447065**	**100.0**	**56068**	**100.0**	**44044**	**100.0**	**46489**	**100.0**	**278807**	**100.0**	**21657**	**100.0**

data source indicates the ratio of the unique content against the full coverage.

In the overall column of Table 2.10, YAGO holds the largest amount of unique content (56.2%), which also implies that it is the data source with the most diverse content. It is nearly double the size of DBpedia. It may be also surprising that Wikidata contains just over a half of the DBpedia data. When we look at this from a cross-domain LOD perspective, the Library of Congress and WorldCat are considered as small-scale datasets, while the number of BabelNet content is even smaller. Obviously, data sources containing fewer entities provide less content.

Regarding the agents category, DBpedia exceeds YAGO and Wikidata. As expected VIAF is also prominent. However, the number is rather disappointing, compared to these three sources.

With regard to events, the reasons why the Library of Congress has relatively high number of contents is mostly due to bflc:subjectOf link. DBpedia provides a large number of seemingly Wikipedia derived content, ranging from links (related persons, places, events, and digital resources) to literal descriptions in different languages.

In the dates category, DBpedia has substantial advantage (88.1%). Other sources are unlikely to offer highly informative content. We also conducted manual inspection on our data sources. We discovered that the high volume of DBpedia in general was most likely due to a large number of links (derived from Wikipedia article dbo:wikiPageExternalLink (i.e., external links, further reading in Wikipedia) and dbo:wikiPageWikiLink (i.e., many useful links in Wikipedia). Wikidata is the second highest source (as it contains labels in many languages), but it is hard to understand the target resource with opaque entity names (wd:Qxxxx). The Library of Congress has useful links to their library resources related to the date (bflc:subjectOf). The Library of Congress and WorldCat use SKOS to connect to broader concepts of decade. It is noticeable that the library-based LOD sources (WorldCat, the Library of Congress, VIAF) have many overlapping content. BabelNet also uses skos:broader, but it seems the links are generated programmatically and it uses proprietary IDs (like Wikidata). Thus, it is hard for machine (and humans) to understand the meaning of the links. In addition, for some reason, the RDF representation of an entity has a significantly lower number of links compared to the HTML representation, therefore, some useful information may be lost.

YAGO shows strength in the places category, given that the ratios are more evenly

Table 2.11: The amount of overlaping content per category (from 1 source to 5 sources)

Category	1 Source	%	2 Sources	%	3 Sources	%	4 Sources	%	5 Sources	%
Agents	42871	76.3	12373	22.0	627	1.1	231	0.4	58	0.1
Events	34308	77.9	9437	21.4	262	0.6	29	0.1	8	0.0
Dates	46163	99.3	290	0.6	36	0.1	0	0.0	0	0.0
Places	255184	91.5	20051	7.2	1500	0.5	823	0.3	582	0.2
Objects & Concepts	16880	84.2	2694	13.4	386	1.9	65	0.3	19	0.1
Overall	**393028**	**88.9**	**43934**	**9.9**	**2764**	**0.6**	**1131**	**0.3**	**664**	**0.2**

Table 2.12: The amount of overlaping content per category (from 6 sources to 10 sources)

Category	6 Sources	%	7 Sources	%	8 Sources	%	9 Sources	%	10 Sources	%	SUM	%
Agents	20	0.0	5	0.0	14	0.0	0	0.0	-	-	56199	100.0
Events	0	0.0	0	0.0	-	-	-	-	-	-	44044	100.0
Dates	0	0.0	-	-	-	-	-	-	-	-	46489	100.0
Places	173	0.1	234	0.1	100	0.0	160	0.1	0	0.0	278807	100.0
Objects & Concepts	10	0.0	2	0.0	2	0.0	0	0.0	-	-	20058	100.0
Overall	**201**	**0.0**	**241**	**0.1**	**116**	**0.0**	**160**	**0.0**	**0**	**0.0**	**442239**	**100.0**

distributed across all sources due to the availability of the entities in this popular category. Interestingly, Getty Vocabularies (TGN) performs relatively well, whereas GeoNames is not as good as we expected. New and diverse information may not be found in the latter.

As for objects and concepts category, the strength of DBpedia persists. It seems that it extracted a great deal of data from Wikipedia. Understandably, Wikipedia articles would be more exciting for human users than a collection of factual data in LOD.

In general, this analysis suggests: a) the concentration of (diverse) content in DBpedia, YAGO, and Wikidata, and b) data richness in specific proprietary properties. A critical question is how the 11 LOD producers facilitate users to find them among hundreds of properties, in order to access rich information, especially if they are unfamiliar with their on tologies. The hurdle could be higher for the data integration by federated queries in multiple LOD sources.

Table 2.11 and Table 2.12 illustrate the amount of data overlaps per category. While the one-source column indicates the number of non-overlapping content for the source (i.e., unique content), other columns indicate the number of overlapping content (i.e, two to ten sources hold identical string). Interestingly, the content covering all data sources does not exist at all. This implies that even the most standard English label cannot be found in every source. Over 75% of content is unique. However, overlaps in two sources are relatively high for agents, events, and objects and concepts. The numbers drop sharply for the overlap in more than two sources. However, very high coverage is also seen for agents, places, and objects and concepts. One reason for these phenomena would be the contrasting volume of data sources. As we have seen earlier, the disproportionately high volume of DBpedia, YAGO, and Wikidata makes the rest of the sources look insignificant. Therefore, although there are some highly overlapping content, the percentages remain very low.

Our assumption is two-fold: 1) the higher the coverage, the more accessible the data, yet the more redundancy in the LOD cloud, and 2) the lower the coverage, the more serendipity with unique content, yet redundant traversals. From this perspective, it is too early for us to judge how much users benefit from a large amount of unique content, and/or how much they suffer from redundant information in multiple sources, because we do not have gold

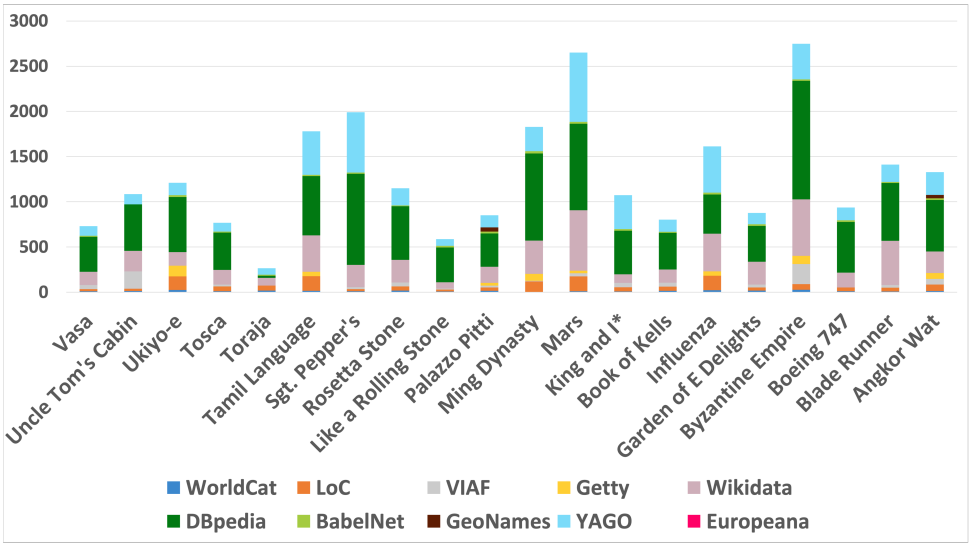

Figure 2.15: The amount of content for objects and concepts entities per data source

standard for data quality.

We additionally created intriguing views of the amount of unique content per entity for each category. Figure 2.15 provides a view for the objects and concepts category. In this case, content diversity is clearly visible, ranging from the rich volume of Byzantine Empire and Mars, to poor volume of Traja and Like a Rolling Stone. The details of other categories and short comments are found in Appendix C.

2.5 CONCLUSIONS

2.5.1 CHALLENGES FOR CULTURAL HERITAGE DATASETS

This research strives to uncover gaps between the data producers and consumers. Indeed, our evaluation of 11 LOD providers reveals a clear sign of data quality issues from a user perspective, which have neither been examined in this detail nor on an instance level by other studies. While it verifies some results of the previous research, it also pinpoints additional issues, in particular, issues specific to the cultural heritage domain, as well as the different types of link properties and literals.

Our analysis confirms the observations of Ahlers and Debattista et al. [39, 63] that a limited number of links are found for major LOD datasets, with the exception of the relatively ample amount for DBpedia (RQ1.1). A large proportion of LOD sources may not be fully connected and unevenly interlinked for the representative entities (RQ1.2, 1.3). This result also reflects previous LOD studies on the overall quality and owl:sameAs networks [56, 69]. In particular, power-law-based networks and closures have been found for the LOD cloud. Moreover, centrality can be observed for not only linkages, but also for data content.

"High-volume and high-quality" datasets are biased toward a couple of data sources, especially generic knowledge bases (RQ1.3). Consequently, it is uncertain if users and

researchers would be able to find new information, let alone to answer more specialised questions that they are interested in. As Zaveri et al. pointed out [142], assuring data quality is particularly a challenge in LOD as the underlying data stems from multiple autonomous and evolving data sources.

Some valuable information about the same entity is not easily reachable due to the lack of links, and/or redundantly long traversing (RQ1.2). For example, it is not possible for a user looking at Beethoven in Getty ULAN to obtain relevant artists and songs in BabelNet. This is discussed in the case of Pablo Picasso in Section 2.1. Generally speaking, due to the heterogeneity of LOD quality and linking patterns, it seems that the automation of graph traversals and the subsequent data integration currently require more human effort than necessary (RQ1.4).

Those are serious shortcomings for our research scenarios. In other words, the quality of a hundred representative entities from major LOD providers has not yet met the basic needs of researchers.

From a user's perspective, our analyses also provide an insight into LOD that previous research has not been able to deliver. For example, it became clear that some objects and concepts may introduce complication, because links between LOD resources may be missing and/or confusingly created (RQ1.3, 1.5). There seem to be a different number of corresponding records, depending on the type of concepts in FRBR (work, manifestation, expression, and item). Unlike skilled librarians, average users on the web would not be able to distinguish four types of FRBR resources and solve co-references on their own. However, this is not a technical problem of LOD, but an issue about the different perceptions and/or understanding of users about the conceptualisation of entities. This "semantic gap" between the data consumers and data producers has the potential to cause problems for research in the future.

As we have seen, an obstacle for interoperability and data processing automation is proprietary properties. LOD is not as powerful as it can be, as long as human users analyse related data every time when traversing data, because they are not initially aware of data sources and their ontologies in their query time [131]. This is particularly true for a large amount of data for which manual analysis is unrealistic. According to Bizer et al. [50], it is a good practice to reuse terms from well-known RDF vocabularies wherever possible, and only if they do not provide the required terms should data publishers define new, data source-specific terminology. In the interoperability metric of Candela et al. [55], the use of external vocabularies is also favoured for the LOD quality assessment. At the same time, we found that rich information tended to be "hidden" in proprietary properties among many other properties (RQ1.1, 1.2). Without close manual examination of ontology and data itself, it would not be easy to automate data processing (RQ1.4).

2.5.2 LIMITATIONS OF OUR ANALYSIS

Admittedly, this article has some limitations. It focuses on the analysis of LOD entities which provide a context for cultural heritage research. For example, as mentioned earlier, Europeana has enriched its digital object datasets with named entities. One may find cultural heritage objects with owl:sameAs links to GeoNames or DBpedia. However, the entity collection of Europeana analysed in this research has been created separately from the object datasets. Europeana offers a) LOD instances (i.e. digital cultural heritage objects

via OAI-PMH and SPARQL endpoint), b) their related entities (i.e. contextual entity via REST APIs that we analysed), and c) the ontology (i.e. Europeana Data Model). Therefore, co-reference resolution should occur in situations such as SPARQL queries, so that the related instances could actually "meet" via an identical entity in the same repository. Thus, it is usually not possible to see such data integration in the lookup scenario we used in our research (RQ1.2).

In addition, in case of external entity linking, federated queries are required to investigate the data integration across different LOD sources, which is slightly out of the scope of our research[43]. For the same reason, we could not apply such sophisticated network metrics as developed by Idrissou et al. [95], because they cannot be easily evaluated in the lookup scenario. Moreover, due to different characteristics of graphs (i.e. weights and directions), it is necessary to heavily customise the metrics. We take those issues for the upcoming research.

Furthermore, largely due to the manual-based methodology, the sample size remains the bare minimum. However, LOD is oftentimes populated programmatically, although crowdsourced LOD such as Wikidata would have more manual curation by human users. In fact, we show that much LOD content is relatively standardised or normalised; the number of links at a data source is relatively similar and consistent across entities in the same category (RQ1.5). It is therefore doubtful that if a large-scale sampling would make our results considerably different.

Nevertheless, our research should aim for the fusion of manual and automatic evaluation in the future. As Idrissou et al. [95] stress, we agree that the links must often be human validated, since entity resolution algorithms are far from being perfect. We also consent to computer support that can accurately estimate the quality of LOD, because the manual analysis is both a costly and an error-prone process.

It is also worth mentioning that there are some technical challenges concerning the automatic analysis of LOD. We encountered many small problems to collect and analyse the data. For example, data is sometimes not consistent (RQ1.1, 1.2, 1.4, 1.5). YAGO has an issue with special characters in the data. We observed this for Uncle Tom's Cabin and Sgt Peppers Lonely Heart Club Band. In case of the former, YAGO's URI is different from that of the DBpedia URI, while all other URIs are identical for the two sources. Thus, error handling was required for those exceptional entities in Python scripts. In addition, the stability of URIs is extremely important, but not always guaranteed. If we look at a broader range of LOD resources, we know that, for example, there was certain impact, when the GND, the German integrated authority records, changed their entity URIs from HTTP to HTTPS in 2019[44].

2.5.3 RECOMMENDATIONS FOR DATA CONSUMERS AND PRODUCERS

Despite those caveats for limitations, the investigation in this research clearly indicates that NEL in local databases may not be as sufficient as one may think (RQ1.1). Our study observes an iceberg of a large variation in data quality on the web [142]. Thus, it would

[43] There is also a serious technical problem with scalability for federated SPARQL queries on the web, which makes it hard to conduct analysis of our kind.

[44] https://wiki.dnb.de/display/DINIAGKIM/HTTP+vs.+HTTPS+in+resource+identification, last accessed 2022-01-18

be wrong to expect that NEL automatically generates synergies for LOD data integration. Indeed, successful projects applying such data integration are highly limited so far in our field. Careful strategies are required to identify efficient traversals and obtain data such as multilingual labels and links to global and/or local databases, and integrate heterogeneous datasets in a useful fashion (RQ1.2, 1.3). One recommendation for the NEL strategy would be to refer to hubs such as YAGO, DBpedia, and WorldCat as much as possible, from where the W3C standardised links to other major LOD resources are available. At the same time, one should be aware that YAGO and WorldCat would be the best choice to find information in Wikipedia. While WorldCat is not connected to DBpedia, it has links to the Library of Congress, which DBpedia does not. Contrary to many practices of NEL in cultural heritage, links to Wikidata would be recommended if the users have a good understanding of its proprietary properties to access other data sources. In addition, our traversal maps can be used as an orientation guide for the NEL implementers.

It is ironic that although Wikidata generally receives high numbers of incoming links from other sources and holds a substantial amount of information, it does not offer the standardised way of providing outgoing links at all. This could be a controversial issue for the efficiency and/or "democratisation" of LOD. A limited amount of new data could be obtained from WorldCat, BabelNet, and GeoNames. It is therefore not promising to carry out serious research with such data as it seems that some datasets tend to serve merely as global identifiers, rather than new sources of information (RQ1.1).

Simultaneously, the use of opaque URIs and a large number of proprietary properties in Wikidata should be more intensively discussed by the LOD publishers and consumers, especially by the NEL implementers, because Wikidata is becoming a NEL standard in cultural heritage [121].

In any case, providing multiple links during NEL will increase interoperability, because it may avoid redundant traversals and give us more flexibility (RQ1.2). At the same time, we can also advise the maintainers of 11 LOD sources to fully link to each other, as well as to provide more links to other local datasets as much as possible. The reciprocal links will allow users to integrate truly interdisciplinary and heterogeneous datasets. In a way, our study identifies the myth of NEL and verifies the obstacles of LOD (RQ1.1). NEL is a step necessary to the use of multiple datasets in LOD [95]. However, linking is the means, not the goal.

2.5.4 DISCUSSIONS ON LOCAL DATASETS

The connection between local datasets and globally known reference resources that we deal with has been largely uninvestigated (RQ1.2). This entails that the local-to-local (L2L) connections via global sources are not well known, although LOD and NEL are designed to perform this task. One exception is demonstrated by Waagmeester et al. [138], describing four cases with federated SPARQL queries to connect Wikidata with local datasets. Yet, our research clarifies that the 11 global LOD sources do not easily enable us to integrate local datasets due to the lack of links to them (RQ1.1). In addition, if two local datasets point to different global sources, they need to traverse more than one graph in order to link each other. This means that the destination of NEL determines the usability of L2L data integration. In any case, a feasibility study on the L2L data integration would be one of the next tasks for our research. We could extend it further by exploring what innovative research

we could actually do after NEL and federated queries. Pilot use cases are needed to simulate and evaluate data aggregation, contextualisation and integration as the outcomes of NEL in the cultural heritage field, followed by semantic reasoning and creation of new knowledge. Otherwise there is a risk that LOD would end up with an idealistic vision without concrete impact on our society.

Related to this, there are also problems with local datasets. It is known that some LOD in cultural heritage is not adequately and sufficiently published. For instance, Francorum Online[45] has technical problems. Pleiades[46] provides RDF/XML, but does not offer links to major LOD that are available in JSON. Other LOD projects (LOCAH[47] and PCDHN[48]) have other problems such as sustainable funding. From a quantity perspective, it is hoped that more local LOD will be published and connected to improve the overall "researchability" for the domain.

2.5.5 FURTHER RESEARCH AND DEVELOPMENT IN THE SEMANTIC WEB

To enhance the analysis carried out in this article, it would be interesting to investigate the LOD traversability in comparison with all the LOD properties actually used. For instance, Linked Open Vocabularies[49] is a good starting point to analyse the acceptance of a broad range of properties for LOD and the implications of standardisation and proliferation of vocabularies. In addition, the automated graph traversals and data integration can be examined, using SPARQL queries. Although our research concentrates on lookup because of the NEL setting, analysis on federated queries can uncover the real research scenarios of the end users.

As Berners-Lee states [48] that "statements which relate things in the two documents must be repeated in each" and further, "a set of completely browsable data with links in both directions has to be completely consistent, and that takes coordination, especially if different authors or different programs are involved." As such, reciprocal links and lookups need to be added with care. For the next step, it seems necessary for the web community to help major LOD dataset maintainers to identify incoming LOD as much as possible, and enrich the datasets to create reciprocal links. Even if a full mesh network is not an aim for many LOD data sources, it would be critical for the LOD creators to be aware of and interconnect with other LOD data sources in order to provide a way to find as much new information as possible (RQ1.1, 1.2, 1.3).

Python analysis let us remember that data overlaps across data sources are duplicate information (RQ1.1, 1.5). On the positive side, fewer traversals are needed to find the same information. On the negative side, data is redundant. As the size of the LOD cloud grows, it may confuse users in the vast amount of information like a needle in a haystack. Use cases by researchers would help to evaluate the pros and cons of the LOD's distributed data approach. In this regard, we also need to find a way to adequately manage and use

[45] http://francia.ahlfeldt.se/index.php, last accessed 2021-01-26
[46] https://pleiades.stoa.org/, last accessed 2021-01-26
[47] http://data.archiveshub.ac.uk/, last accessed 2021-01-26
[48] https://dataverse.library.ualberta.ca/dataset.xhtml?persistentId=doi:10.7939/DVN/URXSGC, last accessed 2021-01-26
[49] https://lov.linkeddata.es/dataset/lov/, last accessed 2021-01-26

aggregation services of LOD.

One example which enables the users to compare LOD sources is SILK [137]. Although it is limited to two data sources, it provides support to create and maintain interlinks. Their update notification service is also particularly valuable. It is also possible and realistic that third-party services would be developed for the integration of LOD data sources [83, 96]. However, there are limited numbers of web applications capable of crawling the web and detecting incoming links of LOD. Some projects offer data dumps containing such information. Yet, they often do not provide an interactive interface. Furthermore, research on LOD search engines is advancing somewhat slowly. Although there are some projects including Swoogle, Sindice, and LODatio [83], many are experimental, out-of-date, or un-user friendly. It is hoped that next generation of search engines for LOD will be developed.

This chapter highlights the reality of a reasonable set of LOD datasets in cultural heritage, but the discussion is applicable for other domains. By removing the obstacles found in this article, LOD traversing and date integration become more feasible for end-users with help of automatised tools.

3

Building Linked Open Date Entities for Historical Research

Time is a focal point for historical research. Although existing Linked Open Data (LOD) resources hold time entities, they are often limited to modern period and year-month precision at most. Therefore, researchers are currently unable to execute co-reference resolution through entity linking to integrate different datasets which contain information on the day level or remote past. This chapter aims to build an RDF model and lookup service for historical time at the lowest granularity level of a single day at a specific point in time, for the duration of 6000 years. The project, Linked Open Date Entities (LODE), generates stable URIs for over 2.2 million entities, which include essential information and links to other LOD resources. The value of date entities is discussed in a couple of use cases with existing datasets. LODE facilitates improved access and connectivity to unlock the potential for the data integration in interdisciplinary research.

This chapter is partly based on *Go Sugimoto. Building linked open date entities for historical research. In Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, editors, Metadata and Semantic Research, Communications in Computer and Information Science, pages 323–335. Springer International Publishing, 2021. doi:10.1007/978-3-030-71903-6_30.* [124].

3.1 INTRODUCTION

This chapter mostly focuses on the data quality, addressing RQ2: "How can Linked Data connectivity for date entities be improved?" (Section 1.5). However, tool quality is also relevant in the context of examining and/or using improved data. The main stakeholders are data producers and data consumers. There is also an element of developers, when it comes to the methodologies and workflows for implementation.

Time is one of the most fundamental concepts of our life. The data we deal with often contain time concepts such as day and year in the past, present, and future. There is no doubt that historical research cannot be done without a notation of time. On the other hand, the advent of Linked Open Data (LOD) has changed the views on the possibility of data-driven historical research. Indeed, many projects have started producing a large number of LOD datasets. In this strand, entity linking has been considered a critical ingredient of LOD implementation. Digital humanities and cultural heritage communities work on co-reference resolution by means of Named Entity Linking (NEL) to LOD resources with an expectation to make connections between their datasets and other resources [122, 133, 134, 143]. It is often the case that they refer to globally known URIs of LOD such as Wikidata and DBpedia for the purpose of interoperability. Historical research datasets include such fundamental concepts as “World War I” (event), “Mozart” (person), “the Dead Sea Scrolls” (object), “the Colosseum” (building), and “Kyoto” (place). However, rather surprisingly, time concepts/entities are not fully discussed in this context. In the following sections, we investigate the time entities in LOD in detail.

The primary goal of this research is to find ways to improve Linked Data connectivity for date entities. More specifically, we aim to foster LOD-based historical research by modelling and publishing time concepts/entities, called “Linked Open Date Entities (LODE)”, which satisfies the preliminary requirements of the target users. In particular, we 1) design and generate RDF entities to include useful information, 2) provide a lookup and API service to allow access to the entities through URI, 3) illustrate a typical implementation workflow for entity linking (“nodification”), and 4) present use cases with existing historical resources.

3.2 RELATED WORK AND UNSOLVED ISSUES

Firstly, we examine published temporal entities in LOD. In terms of descriptive entities, DBpedia holds entities, including the 1980s, the Neolithic, the Roman Republic, and the Sui dynasty. PeriodO[1] provides lookups and data dumps to facilitate the alignment of historical periods from different sources (the British Museum, ARIADNE, etc.). Semantics.gr has developed LOD vocabularies for both time and historical periods for SearchCulture.gr [80]. However, the lowest granularity of time in Semantics.gr is in the early, mid, and late period of a century.

As descriptive time entities are already available, we concentrate on numeric time entities that could connect to the descriptive ones. In this regard, DBpedia contains RDF nodes of numeric time such as 1969. They hold literals in various languages and links to other LOD resources, and can be looked up. However, year entities[2] seem to be limited in the span

[1] https://perio.do (accessed July 20, 2020)

[2] See also https://en.wikipedia.org/wiki/List_of_years. (accessed July 20, 2020)

between ca. 756 BC and ca. AD 2071, while years beyond this range tend to be redirected to the broader concepts of decade. Moreover, there seem to be no or only few entities for a month and day of a particular year. SPARQL queries on Wikidata suggest that the year entities are more or less continuously present between 2200 BC and AD 2200.[3] Year-month entities seem to be merely available for a few hundred years in the modern period[4], and day-level entities are scarce.[5]

Situations are normally worse in other LOD datasets.[6] Therefore, it is currently not possible to connect datasets to the time entities comprehensively for a day and month, or a year in the remote past. This is not satisfactory for historical research. For instance, we could easily imagine how important time information could be in a situation in which the day-to-day reconstruction of history in 1918 during World War I is called for. The same goes for prehistory or medieval history, although lesser time precision would be required.

Secondly, we look for ontologies in order to represent temporal information in RDF. De Boer et al. [61] study TimeML to annotate historical periods, but its XML focus is out of our scope. The Time Ontology in OWL[7] builds on classical works [41, 42, 89] and addresses limitations in the original OWL-Time, which defined concepts like "instant" (a point in time) and "interval" (a period of time) but was restricted to the Gregorian calendar [58]. The updated ontology now supports modeling different temporal reference systems—such as the Jewish calendar or radiocarbon dating—allowing them to reference the same absolute point in time [145]. The specifications also state some advantages of their approach over a typed literal, echoing the vision of our proposal (Section 3.3.1). In the Wikidata ontology, two streams of temporal concepts are present. One is concepts for the unit of time, or time interval, including millennium, century, decade, year, month, and day. The other is considered to be an instance of the former. For example, the second millennium is an instance of millennium, while August 1969 is an instance of month. In the field of historical research, CIDOC-CRM[8] has similarity to Time Ontology in OWL, defining temporal classes and properties influenced by Allen et al. [42]. Zou et al. [145] apply Time Ontology in OWL for the ancient Chinese time, which demonstrates the importance of developing ontologies for non-Gregorian calendars.

Thirdly, a few examples are found along the line of data enrichment and entity linking. During the data aggregation process of Europeana, data enrichment is performed [122]. Some Europeana datasets include enriched date information expressed via edm:TimeSpan in relation to a digital object.[9] It contains URIs from semium.org, labels, and translations.[10]

[3] Currently the lower and upper limit would be 9564 BC and AD 3000

[4] A SPARQL query only returns 218 hits between AD 1 and AD 1600, while 5041 entities are found between AD 1600 and 2020

[5] A SPARQL query returns no hits before October 15 1582 (on the day when the Gregorian calendar was first adopted), and only returns 159691 hits between AD 1 to AD 2020

[6] For example, rare cases include https://babelnet.org/synset?word=bn:14549660n&details=1&lang=EN. (accessed July 20, 2020)

[7] https://www.w3.org/TR/owl-time/ (accessed July 20, 2020)

[8] https://www.cidoc-crm.org/ (accessed July 20, 2020)

[9] The DPLA Metadata Application Profile (MAP) also uses edm:TimeSpan (https://pro.dp.la/hubs/metadata-application-profile) (accessed July 20, 2020)

[10] See an example record at https://www.europeana.eu/portal/en/record/9200434/oai_baa_onb_at_8984183.html. For example, https://semium.org/time/1900 represents AD 1900. (accessed July 20, 2020). Note that semium.org is not longer in service.

Those URIs connect different resources in the Europeana data space. A time concept links to broader or narrower concepts of time through dcterms:isPartOf. Another case is Japan Search.[11] In its data model, schema:temporal and jps:temporal function as properties for time resources.[12] The SPARQL-based lookup service displays time entities such as and https://jpsearch.go.jp/entity/time/1100-1199, which often contain literal values in Japanese, English, and gYear, as well as owl:sameAs links to Wikidata and Japanese DBpedia. The web interface enables users to traverse the graphs between time entities and cultural artifacts in the collection.

Although we have seen examples of LOD for the numeric time entities, an obstacle for NEL for historical research is their limited coverage (including limited granularity) for globally known LOD such as DBpedia and Wikidata. Traversals between LOD entities via numeric temporal entities outside the coverage are not feasible.

3.3 IMPLEMENTING THE LINKED OPEN DATE ENTITIES

3.3.1 NODIFICATION

We shall now discuss why RDF nodes are beneficial. Time concepts in historical research datasets are normally stored as literal values, when encoded in XML or RDF. In fact, those literals are often descriptive dates, such as “early 11th century”, “24 Aug 1965?”, “1876”, and “1185 or 1192”, to allow multilingualism, diversity, flexibility, and uncertainty [80]. De Boer et al. [61] report that less than half of dates in the ARIA database of Rijksmuseum are 3 or 4 digit year. Sometimes literal values are more structured and normalised like 1789/7/14. However, they could be only a fraction. The syntax of “standardised” dates also varies in different countries (10/26/85 or 26/10/85). The tradition of analogue data curation on historical materials may also contribute to this phenomenon to a certain extent.

Whatever the reasons are, literals in RDF have three major disadvantages over nodes: a) new information cannot be attached, b) they are neither globally unique nor referable, and c) they cannot be linked. Since LOD is particularly suited to overcome those shortcomings, literals alone may hinder historical research in the LOD practices. This is the forefront motivation of the transformation of literals with or without data type into nodes/entities/resources. We may call it “nodification”. Figure 3.1 visualises a real example of nodification. ANNO[13] and the Stefan Zweig dataset[14] can be interlinked and the graph network is extended to other global LOD resources. By creating a numeric date node ("1862-05-15"), previously unconnected datasets (ANNO and Stefan Zweig) can be accessed and explored more seamlessly without calculating their common date information, using literals.

Some may still argue that nodification is redundant and/or problematic, because typed literals are designed for time expressions, and XMLSchema-based (e.g. xsd:date) calculations by queries cannot be done with nodes. But, this is not entirely true. First of all, the nodification of this project does not suggest a replacement of literals. When LOD datasets include typed literals, they can be untouched and/or fully preserved in rdfs:label of the new nodes. The temporal calculations are still fully supported, and are encouraged for

[11] https://jpsearch.go.jp/ (accessed July 20, 2020)
[12] https://www.kanzaki.com/works/ld/jpsearch/primer/ (accessed July 20, 2020)
[13] https://anno.onb.ac.at/ (accessed July 20, 2020)
[14] https://www.stefanzweig.digital (accessed July 20, 2020)

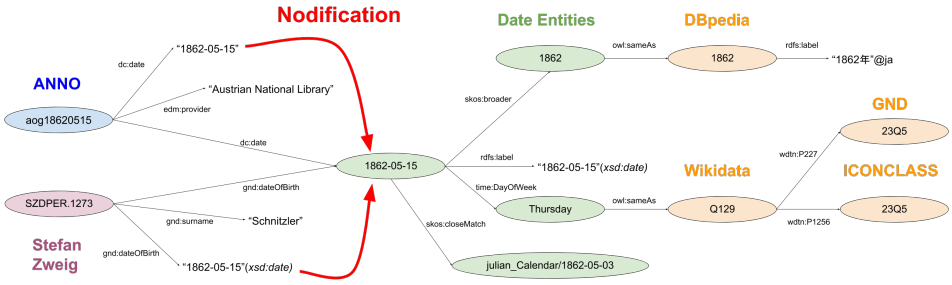

Figure 3.1: Interlinking of nodes by nodification of two datasets

mathematical operations. It is possible to use SPARQL to obtain not only dates in typed literals, but also dates without data types. It is also noted that the year entities in DBpedia do not seem to support data types for literals, thus arithmetic calculations may not be possible, while Wikidata does for the year, month and day entities.[15] Secondly, as the literals are intact, this proposal is a data enrichment and hence not a duplication. The enrichment provides additional possibilities to attach new information, which cannot be achieved by typed literals. Thirdly, a lookup service of LODE serves as a global and permanent reference point for the links across datasets. It encourages data owners to include the entity URIs in their datasets, so that they are able to connect to other datasets that refer to the same URIs. In addition, users often need data browsing before and/or without data querying, in order to understand the scope of data (e.g., data availability, coverage, connectivity, structure) by traversing graphs. Whilst the nodification offers an optimal use case for easy data exploration, literals have limited possibility. Lastly, without the nodification, LOD users have to connect datasets via date literals on the fly whenever they need to. Although it is possible to generate RDF nodes out of literals only when needed, URIs may not be assigned permanently for the nodes in this scenario. Therefore, long-term references are not assured. In addition, it is critical to openly declare and publish URIs a priori through the lookup. Otherwise, it is unlikely that NEL is conducted widely. In a way, the nodification also has a similar scope to materialisation[16] with regard to pre-computing of data for performance. In Table 3.1, several advantages of our approach (preprocessed nodification) are outlined over 1) the use of only literals, and 2) on-demand nodification.

3.3.2 URI SYNTAX PRINCIPLES

In order to execute the nodification, URIs are required. This section briefly highlights the design principles for the URI syntax of LODE. The date URIs consist of a base URI and a suffix. The base URI is set as https://vocabs.acdh.oeaw.ac.at/date/ as a part of an institutional vocabulary service, although it is misleading to be called vocabulary. As for the suffix, we follow the most widely accepted standard, ISO8601, which is the convention of many programming languages (SQL, PHP, Python) and web schemas (XMLSchema 1.1).

[15] However, there would be a problem because it sets January 1 as the value of xsd:dateTime for a time interval entity (e.g. 1987 is represented as 1987-01-01T00:00:00Z)

[16] Materialiastion is the term used in the Semantic Web to generate graphs based on inferences. Implicit knowledge is materialised in order to make it explicit for the purpose of query performance

Table 3.1: Pros and cons of the three approaches

	Preprocessed nodification (our approach)	Only literals	On-demand nodification
Connects to nodes in RDF	✓	✗	⚠ [a]
Includes literals	✓	✓	✓
Possibility to add other information	✓	✗	✗
Time calculation by XSD typed literals	✓	✓	✓
Stable URIs	✓	✗	✗
Lookup service	✓	✗	✗
Graph data browsing	Optimal	Not optimal	Not optimal
Access and query performance	Depends (probably better)	Depends	Depends
RDF data size (after enrichment)	Bigger	Smaller	Smaller
Data processing tasks for enrichment	Much more[b]	No, or much less	No, or much less

[a] Only on demand for selected datasets
[b] This may not be a disadvantage, if one would like to execute other types of data enrichment and normalisation to improve data quality in parallel

The most common formats should look like YYYY (2020), YYYY-MM (2020-01), and YYYY-MM-DD (2020-01-01). An important factor of adopting the subset of ISO8601 is that it can provide non-opaque numeric-based URIs. It enables human users to conjecture or infer any dates, including dates in a remote past and future, even if look ups are not available. In contrast, it is very hard for them to infer dates from opaque URIs such as the Wikidata URIs.[17] ISO8601-based URIs are also language-independent, as opposed to the DBpedia URIs. Those consideration helps researchers who deal with time spanning tens of thousands of years.

The use of ISO8601 also implies that the Gregorian calendar and proleptic Gregorian calendar are applied. The latter is the extension of the Gregorian calendar backward to the dates before AD 1582. Although ISO8601 allows it, the standard also suggests that there should be an explicit agreement between the data sender and the receiver about its use. Therefore, we provide a documentation to explain the modelling policy.[18] In addition, the ISO8601 syntax is applied for BC/BCE and AD/CE, although there is complicated discussions and controversy.[19] Year Zero does not exist, thus "0000" means 1 BC[20] and "-0001" is 2 BC. More than 4 digits ("-13000") allow time concepts in prehistory. As the syntax is the subset of ISO8601, exactly three digits (YYY) and two digits (YY) can be also used, representing a decade and century respectively.[21]

In order to accommodate other calendars (e.g. Julian, Islamic) and dating systems (carbon-14 dating), one can add a schema name between the base URI and the date. For example, we could define URIs for the Japanese calendar as follows:

3.3.3 MODELLING LODE IN RDF

The first implementation of our RDF model should at least include entities at the lowest granularity level of a single day for the duration of 6000 years (from 3000 BC to AD 3000). From the perspectives of historians and archaeologists, day-level references would be required for this temporal range. The number implies that there will be over 2.2 million URIs, counting the units of the whole hierarchy from days to millennia. In any case, this experiment does not prevent us from extending the time span in the future.

Regarding the RDF representation of time entities, we adopt properties from Time Ontology in OWL, RDFS, and SKOS. However, there is a clear difference between LODE and Time Ontology in OWL. The former aims to create historical dates as stable nodes, rather than literals that the latter mostly employs. The latter also does not have properties expressing broader semantic concepts than years; decades, centuries, and millennia are not modelled by default. Therefore, we simply borrow some properties from the ontology for specific purposes, including time:DayOfWeek, time:monthOfYear, time:hasTRS, time:intervalMeets, and time:intervalMetBy. In LODE, the URLs of DBpedia, YAGO, Wikidata, semium.org,

[17] For instance, https://www.wikidata.org/entity/Q2432 represents AD 1984

[18] Full details are available at https://vocabs.acdh.oeaw.ac.at/date/ together with the syntax principles

[19] See https://phabricator.wikimedia.org/T94064. There are confusing specifications in XML Schema Part 2: Datatypes Second Edition (https://www.w3.org/TR/xmlschema-2/#isoformats), XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes (https://www.w3.org/TR/xmlschema11-2/#dateTime), and the HTML living standard (https://html.spec.whatwg.org/#global-dates-and-times). (accessed July 20, 2020)

[20] As seen in Wikidata (https://www.wikidata.org/entity/Q25299). (accessed July 20, 2020)

[21] For example, "196" means the 1960s, and "19" is the 19th century. They should not be confused with "0196" (AD 196) and "0019" (AD 19). Years less than five digits must be expressed in exactly four digits

and Japan Search are included in our entities as the equivalent or related entities, where possible, especially for the entities of years and upward in hierarchy. Listing 3.1 illustrates a typical date entity for the day-level.

In order to generate 2.2 million entities, we have created dozens of Perl scripts, producing entities in RDF/XML for days, months, years, decades, centuries, and units of time, because of the complexity of generating the DBpedia and YAGO URIs as well as literal variations for different units of time. As there are only 6 millennia, they are manually created as the top-level entities. The Perl library of DateTime[22] is primarily used to calculate, for example, the day of a week, the day of a year, and the corresponding day of the Gregorian calendar in the Julian calendar. Some small functions are also developed to generate variations of descriptive dates in English and German and to calibrate entities for BC and AD as well as leap years.

The overall structure of various entities in LODE is visualised in Figure 3.2. There were two choices to create links between the date entities. One is the SKOS vocabulary, and the other is an ontology using RDFS/OWL. According to the SKOS Reference specifications[23], a thesaurus or classification scheme is different from a formal knowledge representation. Thus, facts and axioms could be modelled more suitably in an ontology, as formal logic is required. The date entities seem to be facts and axioms, as we are dealing with commonly and internationally accepted ISO8601. However, from a historical and philosophical point of view, one could also argue that they are also heavily biased toward the idea of the Christian culture. Therefore, the decision to adopt SKOS or OWL was not as simple as it seemed. We primarily use SKOS for two reasons: a) the implementation of a lookup service is provided by SKOSMOS, which requires SKOS, b) it is preferred to avoid debates on the ontological conceptualisation of time for the time being. It is assumed that even Wikidata ontology (Section 3.2) could be a subject of discussion. Moreover, there would be potential problems with using semantic reasoners, for example, due to the inconsistency of our use of decades and centuries.[24] In this sense, SKOS is more desirable thanks to its simple structure and loose semantics. Nevertheless, we adopted the same structure of the Wikidata ontology for interoperability reasons, by simply replacing all its proprietary properties with skos:broader and skos:narrower and a couple of Time Ontology in OWL. This builds a hierarchy of time concepts such as day, month, and year. Similarly, we separate the units of time from instances, with rdf:type connecting them. The sequence of the same time unit is encoded by time:intervalMetBy and time:intervalMeets.

[22] https://metacpan.org/pod/DateTime (accessed July 20, 2020)

[23] https://www.w3.org/TR/skos-reference/ (accessed July 20, 2020)

[24] For example, a popular usage is that a decade begins with a year ending in 0 and ends with a year ending in 9, while a decade starts with a year ending in 1 and concluding with a year ending in 0 in a rarer version. Wikidata adopts the former, resulting the 0s and 0s BC consisting of only 9 years. A similar conflict of constructs exists for the use of the century

Listing 3.1: The 1901-02-01 entity in Turtle

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix acdh-date: <https://vocabs.acdh.oeaw.ac.at/date/> .
@prefix acdh-ut: <https://vocabs.acdh.oeaw.ac.at/unit_of_time/> .

acdh-date:1900-02-01
    a acdh-ut:February_1, acdh-ut:day, skos:Concept ;
    skos:definition "1900-02-01 in ISO8601 (the Gregorian and proleptic Gregorian
      calendar). 1st February 1900."@en ;
    skos:prefLabel "1900-02-01"@en ;
    time:intervalMetBy acdh-date:1900-01-31 ;
    skos:broader acdh-date:1900-02 ;
    skos:altLabel "1 February 1900"@en, "1st February 1900"@en, "01-02-1900"@en,
      "02/01/1900"@en, "01/02/1900"@en ;
    time:DayOfWeek <http://www.w3.org/ns/time/gregorian/Thursday>, acdh-ut:Thursday ;
    skos:note "With regard to Date Entity modelling, documentation should be
      consulted at https://vocabs.acdh-dev.oeaw.ac.at/date/. It includes information
      about URI syntax, ISO8601 conventions, and data enrichment among others." ;
    time:hasTRS <http://www.opengis.net/def/uom/ISO-8601/0/Gregorian> ;
    time:intervalMeets acdh-date:1900-02-02 ;
    time:monthOfYear <http://www.w3.org/ns/time/gregorian/February>, acdh-ut:February
      ;
    skos:inScheme acdh-date:conceptScheme ;
    rdfs:label "1900-02-01"^^xsd:date ;
    skos:closeMatch acdh-date:julian_calendar_1900-01-20 .
```

A lookup service is implemented with SKOSMOS[25] (Figure 3.3). Once SKOS compliant RDF files are imported to a Jena Fuseki server, one can browse through a hierarchical view of the vocabulary and download an entity in RDF/XML, Turtle, and JSON-LD.

[25] https://skosmos.org/ (accessed July 20, 2020)

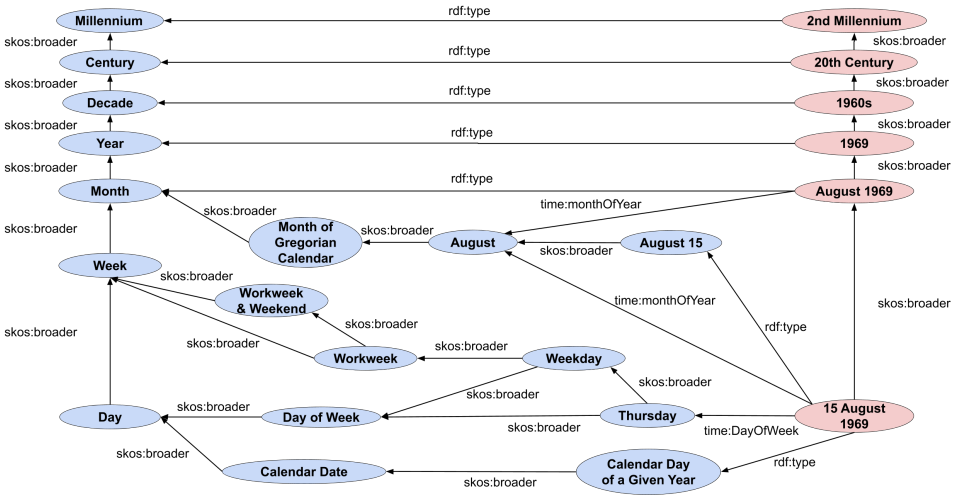

Figure 3.2: The structure of LODE based on the Wikidata ontology, in which the unit of time concepts (in blue) and its instances (in red) are linked each other (Color figure online)

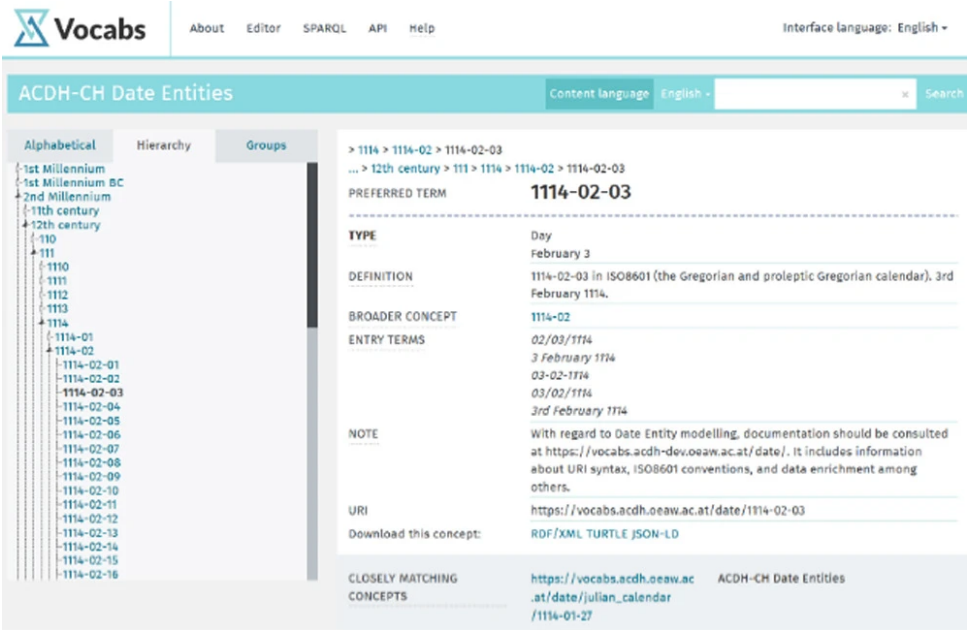

Figure 3.3: Date entity lookup in SKOSMOS

3.4 USE CASES

One benefit of LODE is the capability of handling multilingualism and different calendars. In a use case of Itabi: Medieval Stone Monuments of Eastern Japan Database[26], one may like to align the Japanese calendar with the Western one, when expressing the temporal information in the dataset as LOD. The trouble is that most records hold the accurate date (i.e. day and month) in the Japanese calendar, and only the equivalent year in the Western calendar. Thus, while preserving the original data in literals, it would be constructive to use nodification and materialisation techniques to connect relevant date entities to the artifact (Figure 3.4). LODE helps substantially in this scenario, because it allows us to discover the corresponding day both in the proleptic Gregorian calendar and the Julian calendar by inferences/materialisation, as well as the day of the week. The implementation is not possible yet, however, LODE plans to include mapping between the Japanese and Western calendar in the future. By extending this method, we could expect that LOD users can query LODE to fetch a literal in a specific language and use it for querying a full-text database that is not necessarily RDF-compliant, and does not support the Western alphabet and/or calendars. Such a use case is not possible with literals alone.

A more typical pattern of nodification is data enrichment. The Omnipot project in our institute aims to create an extremely large knowledge graph by ingesting local and global LOD into one triple store. The project evaluates the connectivity of heterogeneous graphs through LODE and the usability of data discovery and exploration. During the nodification of 1.8 million literals in ANNO, not only dates but also data providers and media types are nodified. In this regard, the nodification is not a labour-intensive obstacle, but a part of data improvement and NEL. A similar nodification is conducted for the Schnitzler-LOD datasets[28] and PMB[29] by Regular Expression. This practice verifies our approach with human-inferable non-opaque URIs. Unlike the Wikidata URIs, the LODE and DBpedia URIs were embedded with little effort. The simplicity of implementation incentivises data owners to nodify their data in the future.

Research Space[30] displays incoming and outgoing node links in the Omnipot project automatically. Figure 3.5 showcases connections between them, via the 1987 entity. Users could interactively compare art objects in Europeana with art works in Wikipedia from the same year via Wikidata.[31] This view is currently not possible with literals alone. By default many visualisation software offers a graph view, enabling users to focus on nodes as means to traverse graphs. Therefore, they do not have to worry about query formulations. As it is not trivial to construct the same view by a query using literals, user-friendliness should be considered as a selling point of nodification.

[26] https://www.rekihaku.ac.jp/up-cgi/login.pl?p=param/ita2/db_param (accessed July 20, 2020)
[27] https://www.rekihaku.ac.jp/doc/itabi.html (accessed July 20, 2020)
[28] https://schnitzler-lod.acdh-dev.oeaw.ac.at/about.html (accessed July 20, 2020)
[29] https://pmb.acdh.oeaw.ac.at/ (accessed July 20, 2020)
[30] https://www.researchspace.org/ (accessed July 20, 2020)
[31] As Wikipedia is not LOD, only links to Wikipedia articles are shown and clickable

板碑情報 【自治体番号】 04202 (宮城県) 【板碑所在地番号】 0055 【板碑番号】 0049 【文献番号】 _

【主尊】 _ 【引用文献ID】 石 － 1

【図像】 _ 【引用文献ID２】 _ － _

【法名俗名】 □縁禅尼

【造立趣旨】 _

【装飾】 _

【真言】 _

【偈】 その他（応無所住　而生其身）

【石材】 粘板岩

【年号月日】 永享２年２月２４日

【法量（全高）】 _ 【法量（現存高）】 _ 【干支】 _ 【西暦】 1430 【法量（地上高）】 92

【法量（厚さ）】 12 【法量（上幅）】 _ 【法量（下幅）】 20

【拓本】 有 【写真】 無 【記入者】 三宅宗議

【備考】 石－１、高木５６

所在情報

Day in the Japanese calendar

Year in the Western calendar

Figure 3.4: A record containing the Japanese and Western calendar[27] (above) and an example of its simplified RDF model connecting date entities (below)

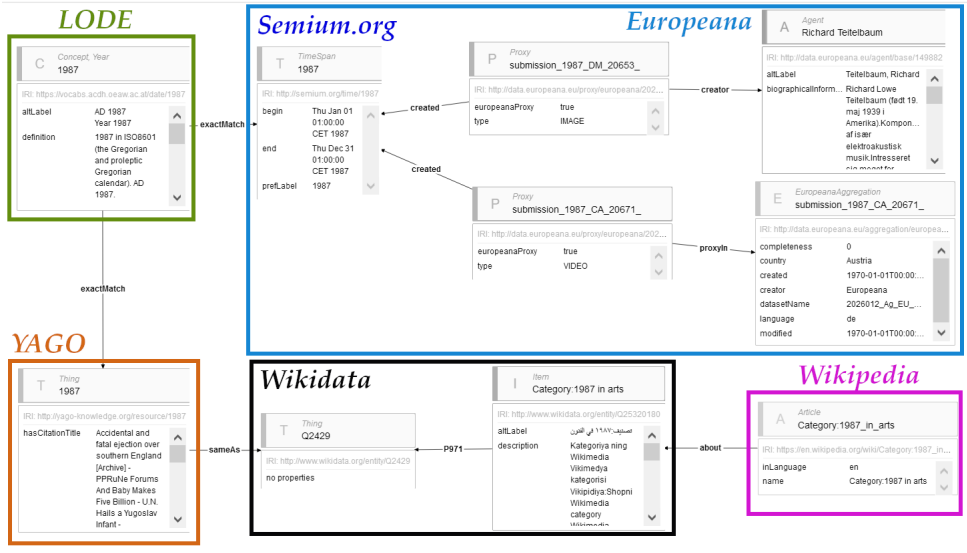

Figure 3.5: Interactive LOD exploration through node links in Research Space

3.5 FUTURE WORK AND CONCLUSION

Future work would be more case studies which the use of literals alone cannot be easily replicated. An RDF implementation of TEI[32] could bring interesting use cases by normalising and nodifying date literals in various languages, scripts, and calendars in historical texts. In addition, LODE could align with the Chinese, Islamic, Japanese, and Maya calendars, and add more information about festivities and holidays that literals cannot fully cover. Consequently, event-based analyses by SPARQL may uncover unknown connections between people, objects, concepts, and places in a global scale. It could even connect to pre-computed astronomical events at various key locations such as the visibility of planets on a specific day, with which interdisciplinary research can be performed. Further, a detailed evaluation of use cases is also needed. For instance, query performance and formulation, and usability could be measured and analysed more systematically. We are also fine-tuning the LODE model by properly modelling the concepts of instants and intervals based on the Time Ontology in OWL.

LODE attempts to solve two problems of existing LOD: a) It tries to meet the needs for greater coverage and granularity of date entities for historical research. b) By designing a simple model and suggesting a straightforward method of nodification, it helps to reduce the complexity of LOD by automatically connecting/visualising vital information in a big Web of Data. Although the research focused exclusively on cultural heritage, many science domains deal with some conception of time, and thus, this study could be an impetus to acknowledge the necessity and the impact of time entities in a broader research community. Since time is one of the most critical dimensions of datasets, we would be able to unlock more potential of LOD.

[32] https://tei-c.org/ (accessed July 20, 2020)

4

Closer Reading of RDF Generated by NLP on Wikipedia Biography: Comparative Analysis

Although Wikidata and DBpedia are closely related to Wikipedia, they often hold a small subset of its semantic information due to the specific scopes and methodologies chosen for their Linked Data (LD) construction. When we look at biographies, a large amount of RDF statements focus on the core facts about persons and many semantic narratives are not included. This hinders expert historical research on the details about the subject matter. To fill this knowledge gap, we seek a solution with Natural Language Processing (NLP). We aim to assess to what extent out-of-the-box NLP tools can semi-automatically generate new LD from the biographical articles in Wikipedia. Unlike other NLP research, we put more emphasis on the qualitative analysis of NLP ("close reading") than the statistical performance of NLP algorithms ("distant reading"). We evaluated the overlaps and gaps between Wikipedia, Wikidata, and DBpedia, as well as other biographical ontologies. We analysed the triple patterns from the NLP results in comparison with the RDF entity (instance) and ontologies. Our research revealed that we are able to capture new information about the entity that Wikidata and DBpedia do not hold. At the same time, some noise could not be easily eliminated. Our method presented a bottom-up approach for biographical ontology designing. We also briefly propose a possible solution for future work.

This chapter is partly based on *Go Sugimoto, Angel Daza, and Victor de Boer. Closer reading of RDF generated by NLP on wikipedia biography: Comparative analysis. In Emmanouel Garoufallou and Fabio Sartori, editors, Metadata and Semantic Research, pages 41–54. Springer Nature Switzerland, 2024. doi:10.1007/978-3-031-65990-4_4.* [127].

4.1 INTRODUCTION

This chapter mostly focuses on the data quality, addressing RQ3: "What are the quality gaps in biographical information between Wikipedia and Linked Data, and how can Information Extraction on Wikipedia be used to address them?" (Section 1.5). All three stakeholders play a role for evaluating the gaps and delivering solutions.

In the field of knowledge representation, there has been significant progress in recent years. Extremely large data sets have been created in the form of Knowledge Graph (KG). KGs consist of vertices and edges to semantically represent the connections between concepts and entities. KGs are highly valuable when they are published as Linked Data (LD).[1] Two well-known LD resources are DBpedia and Wikidata, which contain millions of facts about the entities covering a wide range of human knowledge. Both are closely related to Wikipedia, which is one of the primary sources of information on the web[2]. They are used not only by KG experts, but also increasingly by ordinary users.

Due to their substantial size and coverage, users expect that they can find specific relationships between entities. However, this is not always the case. For example, surprisingly Henry VIII in the English DBpedia[3] refers to the English Reformation with weak semantics. It appears only in full text in *dbo:abstract* and *rdfs:comment*, as well as *dbo:wikiPageWikiLink*. In contrast, it is in the first paragraph of the Wikipedia article on Henry VIII that describes how he started it in relation to Pope Clement VII and Catherine of Aragon[4]. Almost the same semantics are available in the inverse relation. In Wikidata, the reformation is only mentioned in the part of Anglicanism as the religion or worldview of Henry VIII, while he was referred as the founder in the inverse relation.

One of the reasons is that DBpedia, for instance, generates KG by taking advantage of the info box of Wikipedia, which is a small summary table of facts of an article. This structured data helps construct refined LD properties for the entity. In any case, those KGs do not match the information found in Wikipedia, partly because the majority is unstructured text. This hinders expert historical research on the details about the subject matter. This knowledge gap can also impact on the use of semantic queries. Biases and misinterpretations may occur. Is there a way to overcome such a shortcoming for historical research?

On the other hand, in the field of Natural Language Processing (NLP), the quality of Information Extraction (IE) has been increasing every year. In particular, there have been intensive studies on the IE for KG [75, 102, 110]. As they attempt to automate KG generation, it can solve the problem of the knowledge gap described above. However, these systems are not always easily accessible for the domain researchers with limited NLP skills.

In addition, many NLP projects aim to generate KGs without hyperlinks; and few aim to create Resource Description Framework (RDF) and publish it as LD [75][5]. The major interest of previous studies concentrates on the statistical performance of IE without analysing the

[1] https://www.w3.org/wiki/LinkedData (Accessed on 2023-07-10)

[2] As of June 2023, Wikipedia ranked as the seventh most visited site on the web: https://www.similarweb.com/en/top-websites/ (Accessed on 2023-07-10)

[3] https://dbpedia.org/page/Henry_VIII (Accessed on 2023-07-10)

[4] https://en.wikipedia.org/wiki/Henry_VIII (Accessed on 2023-07-10)

[5] According to Exner, the dataset is not publicly published despite their claim in their paper (personal communication, 2023-03-24)

details of the NLP results. Is it possible for ordinary users to generate new RDF out of texts using out-of-the-box tools without training new models?

In this chapter, we investigate the possibility of generating RDF from Wikipedia and its potential for better historical research. We aim to extract new and/or more detailed semantics and measure to what extent pre-trained NLP tools can help supplementing existing LD. In order to differentiate ourselves from the statistical overview of IE (*distant reading*), we consolidate them with a case study of a historically important person (*close reading*)[99], in our case Henry VIII. We evaluate in depth the overlaps and gaps of knowledge between those three resources by comparing the existing RDF statements of a biography with the newly generated statements. We shall call it a *closer reading*.

The emphasis is on examining the potential to arbitrate the storytelling nature of Wikipedia and the database nature of DBpedia and Wikidata. Due to rich narrations, Wikipedia articles tend to contain more anecdotes of an entity as opposed to "list of facts", which are likely to be of the interest of humanities researchers. This project attempts to fill such gaps as a LD enrichment effort. To this end, our research questions are:

- RQ3.1: Can new RDF statements be generated by out-of-the-box NLP systems on a Wikipedia article, and how the quantity and quality are, compared to DBpedia and Wikidata instances?

- RQ3.2: To what extent are they compatible with biographical ontologies?

4.2 PREVIOUS WORK

Many NLP papers explore IE and related tasks (just to name a few: [91, 100, 101]); they tend to concentrate on the methodologies and statistics about the task, ignoring what is actually extracted and the extent to which it can be useful for non-NLP researchers. This is especially noticeable with Deep Learning based systems, where there is a hyper-focus on evaluation metrics on a handful of well-curated test sets and no showcasing of their usefulness in a real-world scenario. Additionally, even when the latest models show improvements in the test sets, setting them up and using them for ad-hoc scenarios is not always possible. For this reason, we focus here on assessing services with pre-trained models.

Exner and Nugues dealt with a broad range of Wikipedia texts [75]. They scanned over 114,000 Wikipedia articles and extracted more than one million triples. Nevertheless, the most frequent ontology mapping of their result was highly biased toward biographical information, which was common triple patterns often found in DBpedia and Wikidata. Our approach contrasts with theirs in that we focus on discovering additional information not captured in most LD.

The study of Alam et al. [40] coincides with our technique. They extensively employed Semantic Role Labeling (SRL) for IE on corpus from the Penn Treebank and the Brown corpus. However, it differs from our scope of biographies. PIKES[6] provided a service to generate a KG out of texts (typically sentences). Due to the use of DBpedia URIs, the KG looks similar to LD. Although their automatic IE is sophisticated, the KG often does not make much sense to the human users.

[6]https://pikes.fbk.eu/ (Accessed on 2023-07-10)

4.3 METHODOLOGY

4.3.1 OVERVIEW OF DATA AND TASKS

The *closer reading* allows us to investigate the details about the information extracted in conjunction with the overview statistics. Henry VIII was chosen as a case study for biographies for three reasons: a) RDF example is available in BIO Vocabulary[7], which we will compare with our result, b) he is one of the most referred historical persons in Wikipedia[8], and c) rich biographical information is found in the Wikipedia article.

LD uses RDF to model triple (Subject-Predicate-Object) statements. RDF Classes and Properties in an ontology define the semantics of entities and relations. The advantage of LD is that it can connect entities on the web, using hyperlinks. The use of URI (Uniform Resource Identifier) makes it possible to disambiguate similar entities globally. Besides, as the object of a statement can become the subject of another, statements can be connected each other.

In order to generate RDF statements from Wikipedia which we can compare with existing RDF statements in DBpedia and Wikidata, a NLP pipeline was established in a Python environment together with RDFLib[9]. In particular, we performed the following tasks:

- Triple extraction, using SpacY[10], Allen NLP[11] and Flair [12]
- RDF generation by aligning the outcomes from the triple extraction with Named Entity Linking (NEL) by MediaWiki API[13] and Dandelion API[14].
- Evaluation of RDF statements against biographical ontologies and instance comparison with DBpedia and Wikidata

4.3.2 TRIPLE EXTRACTION FROM WIKIPEDIA

The first stage of NLP pipeline is divided into three steps: downloading the plain text from the English Wikipedia article, extracting sentences, detecting triples, and creating URIs. First, we obtained the plain text from Wikipedia, using the Wikipedia API. We cleaned the texts by removing empty lines and headings. Second, we used spaCy to parse the text, and detect sentences. Third, we used the SRL tagger from Allen NLP and the Semantic Frame Detector (SFD) from Flair to identify simple triple structures in each sentence. The SRL tagger identifieed predicate-argument structures using the PropBank tags [111].

To simplify our process, we focused on six argument labels: ARG0 and ARG1, which are called proto-agent and proto-patient respectively, who play the most critical roles in the sentence. We also included ARG2 (theme or secondary patient), temporal and geographical information (ARGM-TMP and ARGM-LOC), and negation (ARGM-NEG). Flair detects the semantic frame of a predicate disambiguating the predicate sense [44]. For example, succeed.02 (*take over for*) is different from succeed.01 (*win, accomplish some task, successful,*

[7] https://vocab.org/bio/ (Accessed on 2023-07-10)
[8] https://www.quantware.ups-tlse.fr/QWLIB/topwikipeople/index.html (Accessed on 2023-07-10)
[9] https://github.com/RDFLib/rdflib (Accessed on 2023-07-10)
[10] https://spacy.io/ (Accessed on 2023-07-10)
[11] https://allenai.org/allennlp (Accessed on (Accessed on 2023-07-10)
[12] https://github.com/flairNLP/flair (Accessed on 2023-07-10)
[13] https://www.mediawiki.org/wiki/Extension:TextExtracts#query+extracts (Accessed on 2023-07-10)
[14] https://dandelion.eu/ (Accessed on 2023-07-10)

accomplished). This function enables us to retrieve more specific semantics for ontology comparison.

Additionally, we created two types of URIs for the triples by detecting: a) Wikipedia's internal hyperlinks by the MediaWiki API, and b) NEL using the Entity Extraction API from Dandelion. We used the threshold of 0.3 as the confidence level for the NEL configuration to identify as many potentially relevant links as possible and to match the recognized entities.

4.3.3 RDF GENERATION

The outcomes of SRL, SFD, NEL were cross-referenced. In other words, the JSON files generated by the previous tasks were merged. Consequently, we aligned the triple structure with potential URIs to produce an RDF statement.

The detected SRL arguments were often not distinct entities compatible with other LD. For instance, an object can be a compound entity like *peace with France* instead of a single entity like *France*. As it is a challenge to match named entities with URIs, we adopted five principles to circumvent the complexity: a) If the subject and object of the triple structure contain the detected named entities from Dandelion API or the hyperlinks collected by the MediaWiki API, they were replaced with Wikipedia or DBpedia URIs. Then, the Wikipedia URIs were converted to DBpedia URIs as much as possible[15]. The objects became either bare URIs or texts/literals mixed with URIs. b) If the mapping is not possible for the object, we preserved the detected object as literals with English encoding. c) We focused on two most prominent arguments (the first two) of SRL for the object. d) The subject pronouns were replaced by the DBpedia URI of Henry VIII. e) The predicate URIs were generated by appending the detected semantic frames to the base URI: http://www.wikipedia-bio.org/predicate/.

The last two principles need more explanation. Regarding the subject, since the primary purpose of the Wikipedia article is to describe the subject of the article (i.e., Henry VIII), we set a hypothesis to simplify our process: if the subject position of the triple structure is *he* or *Henry*, there is a high possibility that it is Henry VIII himself. Therefore, we applied the DBpedia URI of Henry VIII to all such cases[16]. Regarding the predicate, it is not practical to predict all possible types. Therefore, unlike the subject, we minted own URI prefix and simply appended unchanged predicates to it.

4.3.4 QUALITY ASSESSMENT

We inspected manually the outcome of the mapping. Defining and measuring the quality of LD is a challenge. There have been some attempts to create a framework for quality matrices such as Luzzu [64]. However, many oftentimes adopt either too generic or too specific criteria on a wide array of data aspects, including accessibility, interoperability, and licensing [142]. The studies on the global quality criteria on data content and relevance are limited, mainly because the data quality is highly local and subjective in a particular domain. However, we aimed to design and apply systematic quality criteria by relevance, if not completely objective.

[15] For label readability (i.e. QID https://www.wikidata.org/wiki/Wikidata:Glossary (Accessed on 2023-07-10)) and simplicity, the URIs were mapped to Wikidata URIs on demand where necessary

[16] This hypothesis can be risky, because there is also a possibility that *he* could be a totally different person. We can also analyse the side effect of this hypothesis.

We simply classified each triple as informative or uninformative. Informative means that the RDF triple is self-explanatory and makes reasonable sense, when compared to the source sentence. While the following two RDF statements are uninformative[17]:

```
@prefix dbr: <http://dbpedia.org/resource/>
@prefix wikibio: <http://www.wikipedia-bio.org/predicate/>

dbr:Henry_VIII
wikibio:publish01 "dbr:Neoplatonism␣of␣his␣own"@en ;

wikibio:coach01 dbr:Nation_state .
```

The next statements highlight the informative RDF:

```
wikibio:sign02 dbr:Treaty_of_the_More ;

wikibio:charge05With dbr:Praemunire"@en␣.
```


Rather than forcing a quality criterion, we attempted to assess the effect of the five principles.

4.4 EVALUATION

In this section, we first outline the general outcome of NLP, and then, compare it with the entity of Henry VIII in DBpedia and Wikidata. We also compare it with biographical ontologies with a focus on generated predicates to understand to what extent our RDF is compatible with them.

4.4.1 OUTCOME OF NLP FOR WIKIPEDIA

Table 4.1 shows the key statistics about the NLP results. A substantial number of information was extracted from 514 sentences. On average, 2.2 triple structures, 3.3 hyperlinks and 5.0 named entities were detected per sentence.

We generated total 316 RDF statements (total of URI and literal)(Table 4.2). So, on average, 0.61 RDF statements were populated per sentence.

The first finding is that many sentences extracted have not only Henry VIII himself as the subject, but also many others. Consequently, many statements were dropped. We found sentences starting with *His disagreement with Pope...*, *Emperor Maximilian I had been attempting...*, and *His absence from the country*. In addition, there is a constraint in RDF that the subject must be a single entity with URI. As a result, the number of triples we can produce for RDF is less than that of KG without URIs. For example, *Henry's forces* cannot be a RDF subject.

[17] Flair somehow changed the verb: *scouted* to *coach.01*

Table 4.1: Key statistics for the NLP results on the Wikipedia article of Henry VIII

Number of Sentences (by spaCy)	514
Number of Triple Structures (by Allen NLP)	1150
Number of Internal Hyperlinks (by MediaWiki API)	1740
Number of Named Entities (by Dandelion API)	2608

Table 4.2: Quality measurements of the object of the generated RDF

URI		Literal	
total	informative	total	informative
31	29	285	253

Moreover, the automatic-mapping limits the quantity and quality of the outcomes. Although it is possible to manually define and fine-tune the mapping in order to improve the performance, the data analysis is not a trivial task and scalability is a serious issue. For instance, we found that the prepositions cause some confusion for the RDF generation. In the following two cases, the predicate *suffer from* and *consoling in* were split between the predicate and object. We used heuristics to add the preposition to the predicate, isolating the URI as the object:

```
wikibio:suffer01From dbr:Gout;

wikibio:console01In dbr:Hunting ;
```

Similarly, adjectives and possession make the triples complicated to auto-generate. In these cases, objects were preserved as literals [18]:

```
wikibio:sponsor01 "two␣lavish␣dbr:Tournament_(medieval)␣a␣year"@en ;

wikibio:have03 "an␣affair␣with␣dbr:Madge_Shelton"@en .
```

Although those triples are understandable with common-sense and we can easily generation a clean RDF from them, full semantic is sometimes not explicit in our RDF; the meaning of the triples is blurred and misleading.

Despite those caveats, we demonstrated a potential of our IE method to acquire a reasonable amount of RDF.

Out of 316 statements 31 objects are URI and 285 are literals. While 29 URIs of the objects were informative, literals contain many uninformative information. A large number of literals were caused by the URI mismatching with the SRL result. Although the correct URI was predicted, the detected object does not correspond to it. However, literals do not mean low-quality RDF. They offer rich description of facts that URI alone cannot convey. For example, there are statements like:

```
wikibio:influence01 "the␣design␣of␣rowbarges␣and␣similar␣galleys"@en;

wikibio:claim01 "descent␣from␣constantine␣the␣great␣and␣dbr:King_Arthur"@en
```

The variation of predicates is positively highly diverse. Out of 316 occurrences, 100 unique predicate senses were extracted (Table 5). Many interesting verbs concerning the life of a person were retrieved. Figure 4.1 illustrates all the predicates of the generated RDF.

4.4.2 COMPARISON WITH DBPEDIA AND WIKIDATA

Table 5 shows the overview of the predicates and classes found in biographical schemata at the instances and ontologies level, compared to our generated RDF instance. Although it is

[18] Note that the principle (a) in Section 4.3.3 is applied to the first example, which shows literals mixed with URIs. This may allow future data users to execute NLP for improving the data (after our LD is published).

[19] be.01 is cut off. It counts for 30 outside the Y-axis boundary

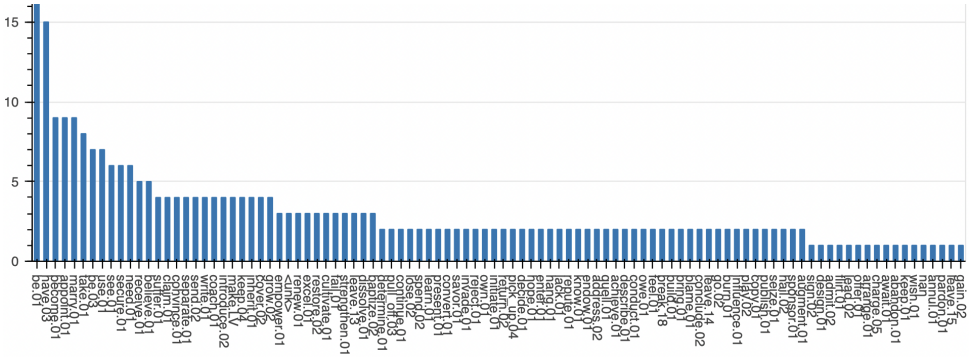

Figure 4.1: Predicate frequency of the generated RDF for Henry VIII[19]

4

not straightforward to directly compare them to each other, we can understand the sense of data size. Notice that our result is only comparable on the instance level.

While 58 out of 908 predicate occurrences in the DBpedia instance are unique, in Wikidata 74 out of 417 are unique. This means that only a small subset of predicates for person entities are used for Henry VIII. Yet, Wikidata contains a large amount of identical entities in other sources[21]. In addition, predicates are used for several subjects and objects.

Figure 4.2 lists all predicates with frequency. We can clearly observe weak semantic relations. Half of the predicates are wikiPageWikiLink, implying that the entity of Henry VIII has some kind of relation to the entities in the object, but it is implicit. We classified the properties into three types: a) 7 *is*-relation were found, including *subject*, *before*, and *predecessor*. b) 21 *has*-relation were found, including *father*, *burialPlace*, and *coronation*. c) 30 were digital specific properties and erroneous data including *wikiPageID*, *owl:sameAs*, and *depiction*.

In the Wikidata instance, 74 properties are unique. The top three predicates in Wikitata are often from inverse properties, and they do not describe Henry's life; films, TV-series, and paintings in which he was depicted. Like DBpedia, the long-tail part of the predicate list is more relevant to the biography of Henry VIII. Our classification found: a) 15 *is*-relation, including *founded by*, *creator*, *occupant*, b) 35 *has*-relation, including *award received*, *given name*, and *medical condition*. c) 24 digital properties and erroneous data, including *described by source*, *Libris-URI*, and *Commons category*. Therefore, we observed considerable reduction of information in DBpedia and Wikidata, when focusing on the stories about Henry VIII. Many properties are related to basic facts, therefore, are different from our results.

Subsequently, 28 in DBpedia and 50 in Wikidata are qualified as a source for biographical properties (or ontology) in reality, so that our results is relatively competitive; unique properties outperformed them.

Next, we analysed the overlap of our RDF with the DBpedia and Wikidata instances. As predicates cannot be easily compared, the object overlap was checked. Out of 316, seven

[20] wikiPageWikiLink in DBpedia is cut off. It counts for 454 outside the Y axis boundary

[21] See the significant difference after the properties containing *ID* in its name is removed. The ID predicates are closely related to owl:sameAs link

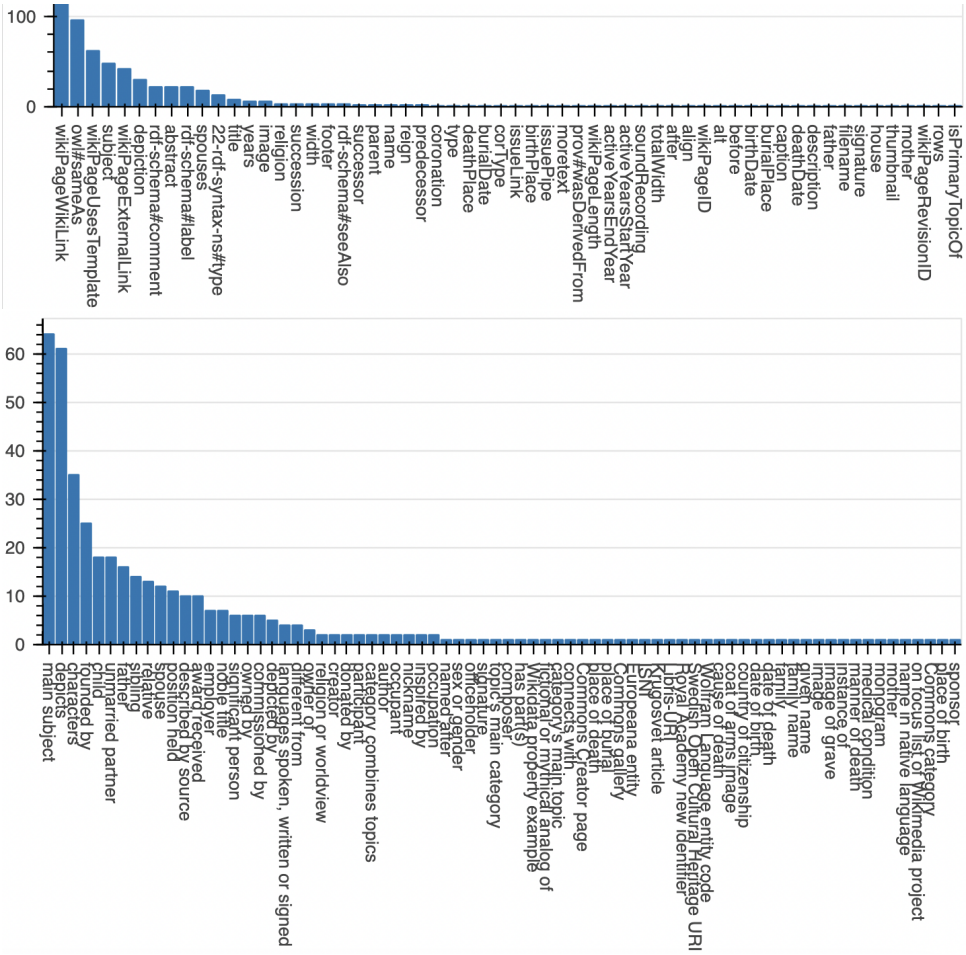

Figure 4.2: The predicate frequency of Henry VIII in DBpedia (top) and Wikidata (bottom) [20]

Table 4.3: Matching RDF objects between ours and DBpedia and Wikidata

Our Predicate	DBpedia/Wikidata Predicate	Object	Informative
wikibio:separatingFrom	dbo:wikiPageWikiLink	dbr:Church_of_England	x
wikibio:signing	dbo:wikiPageWikiLink	dbr:Treaty_of_the_More	x
wikibio:seizing	dbo:wikiPageWikiLink	dbr:Dissolution_of_the_monasteries	
wikibio:appointedAs	dbo:wikiPageWikiLink	dbr:Supreme_Head_of_the_Church_of_England	x
wikibio:secured	dbo:wikiPageWikiLink	dbr:Boulogne-sur-Mer	x
wikibio:initiate	**dbo:wikiPageWikiLink**	**dbr:English_Reformation**	**x**
wikibio:been	dbo:wikiPageWikiLink	dbr:Kell_antigen_system	x
wikibio:initiate	**wdt:P112 (founded by)**	**wd:Q1645505 (English Reformation)**	**x**

Table 4.4: Examples of triples with URIs that have no or weak semantics in DBpedia and Wikidata

Predicate	Object
wikibio:play02	dbr:Organ_(music)
wikibio:suffer01From	dbr:Gout, dbr:McLeod_syndrome
wikibio:secured	dbr:Boulogne-sur-Mer
wikibio:hang01	"2,000 http://dbpedia.org/resource/Tapestry"@en
wikibio:introduce02	dbr:Renaissance_music
wikibio:arrange01	dbr:Jousting
wikibio:augment01	dbr:Treasury

(DBpedia) and one (Wikidata) objects were matched (Table 4.3). As anticipated, loose semantic of DBpedia was verified. As described in Section 4.1, the English Reformation was indeed a match. While we managed to generate a better semantic statement than DBpedia, ours was semantically similar to Wikidata. Nevertheless, it may be slightly confusing because the original sentence describes the disagreement with Pope Clement VII as a cause, which is missed out. With our method, it is not straightforward to verify the fact without careful inspection.[22]

In addition, we compared our results with Exner and Nuges. Among the four most frequent patterns (bear.02, marry.01, locate.01, establish.01), only marry.01 was found in our list, verifying our diversity (Figure 4.1). Although the generated data size is relatively small, our method was capable of extracting more semantics. In contrast, it became clear that our method is not particularly suited to replicate the existing same RDF statements. Some interesting triples generated are shown in Table 4.4.

4.4.3 ONTOLOGY COMPARISON

The predicates of our RDF triples were scrutinised to understand the compatibility with existing ontologies. They are BIO Vocabulary, Bio CRM[23], Factoid[24], and CIDOC-CRM[25], as well as DBpedia and Wikidata. The number of classes and properties was compared (Table 5).

It is clear that the ontologies capable of representing biographies have relatively small number of properties, especially person specific ones (the right most column). Many ontolo-

[22]For example, *wikibio:written dbr:Greensleeves* is reputed.
[23]https://seco.cs.aalto.fi/projects/biographies/ (Accessed on 2023-07-10)
[24]https://www.kcl.ac.uk/factoid-prosopography/ontology (Accessed on 2023-07-10)
[25]https://www.cidoc-crm.org/ (Accessed on 2023-07-10)

Table 4.5: Comparison of classes and properties across different ontologies and instances

	Type	# of classes	# of Properties	# of Properties (human/person)
Ours	Instance	-	-	**100 (316 occurrences)**
BIO Vocabulary	Ontology	43	30	- (42 event classes)
	Instance	-	-	**10 (58 occurrences)**
Bio CRM	Ontology	26	14	8
Factoid	Ontology	38	37	4
CIDOC-CRM	Ontology	81[+]	160[+]	30
DBpedia	Ontology	1108[†]	54941[†]	382 (domain of 56 classes)
	Instance	-	-	**58 (908 occurrences)**
Wikidata	Ontology	2771888[‡]	11085[‡]	2144 (472 without ID)
	Instance	-	-	**74 (417 occurrences)**

[+] https://cidoc-crm.org/sites/default/files/cidoc_crm_v7.1.2.pdf

[†] http://78.46.100.7:9000/

[‡] https://sqid.toolforge.org/#/browse?type=classes

https://www.wikidata.org/wiki/Wikidata:List_of_properties

gies model fundamental attributes of persons such as birth, death, gender, and occupation rather than individual life events.

In CIDOC-CRM, there are 30 person-related properties such as *died in* and *has current or former residence*. It also models personal events by 27 classes including *Production*, *Joining*, and *Curation Activity*. Similarly, Factoid and Bio CRM also tend to avoid providing specific event and action types of persons. Rather, event types and person roles in event can be specified by the users, facilitating heterogeneity. Their properties include *has nationality* and *hasLifeDates*. Contrary, BIO Vocabulary has a more extended list of properties for events. An example encoding of Henry VIII uses 10 unique properties out of 58 occurrences. Their class list contains *Baptism*, *Murder*, and *Emigration*.

Our analysis highlighted the difficulties of comparison due to the diversity of data modelling principles. There is a mix use of classes and properties as well as verbs and nouns. In addition, our instance data is not easy to map to the existing ontologies, which is often designed in a top-down manner. However, our results were significantly more than BIO Vocabulary. The variety of properties were nearly twice as many as DBpedia. In this regard, our list of predicates based on a bottom-up method can be used as a starting point to create and update biographical ontologies.

4.5 DISCUSSION AND FUTURE WORK

Although our research brings new insight into LD generation by the current NLP tools, there are several limitations.

The out-of-the-box NLP tools were not able to produce as much clean RDF as expected. There are several reasons: a)SRL is not specifically tailored to extract triple structure, b)the quality of NEL by Dandelion API is limited, c)the automatic mapping is not straightforward, and d)difficulties concerning the complex sentences. As Alam et al. [40] noted, the semantics of ARG0-ARG5 is not clear, and the mapping of subject and object is not always trivial.

There is also room for developing heuristics. We deliberately conducted data cleaning as

minimum and analysed the consequence, but the precision will increase as we fine-tune the cleaning.

Since we did not manipulate the predicates generated, they are ontologically too broad. Although the senses were not sensitive in our case study, they would have some impact for an extended study. A future plan is to compile them using hypernyms on the semantic frames by dictionaries such as WordNet[26] and PropBank. A shallow hierarchy can be build using *rdfs:subPropertyOf* in order to transform the property list into a manageable level of ontology.

Oftentimes the spotted arguments cannot be simply replaced by URIs detected by NEL, because they are longer than the length of the entity. Our experiment did not have problems with our hypothesis on the subject replacement. However, careful examination is required to fully understand the mechanism how the RDF was generated from the source sentence. Sometimes, RDF does not originate from the main clause of the sentence, so misleading information may be extracted. A statement of Henry VIII can be a historian's interpretation. Thus, there is a risk to generate a RDF for debated facts. As the Wikidata data model allows users to add provenance information, as well as the nature of statement [66] [27], is is insightful for our future work.

There is a solution for the noise reduction. As we hold all information about data process including the source sentences and Wikipedia sections, we can create and publish compound RDF about it on the web. Then, the user community could use SPARQL update to improve incomplete or erroneous statements. Technologies such as RDF Star[28] and PROV-O[29] can be used to preserve sentence-graph pairs as well as provenance metadata.

We also found that our method still requires a lot of manual work, spanning from data cleaning and mapping based on data analysis, heuristics, and programming. Therefore, it is not optimal for the NLP novice users. As a result, our research was limited to one biography. Without doubt, more interesting findings can be made if we analyse more biographies. There is a high potential to explore persons in different time and space as well as gender gaps to conduct chronological, geographical, and gender studies.

Another dimension of work will be the investigation of Machine Learning (ML) predictions, based on the annotations/analysis we conducted. As the amount of data in biographies becomes beyond the capability of manual processing, ML will play a vital role. For instance, we can combine the summarisation of texts by a generative model with IE and LD generation by our method, so we will have more control over the data generation process than ML models.

4.6 CONCLUSION

In this chapter, we presented an experiment on the LD generation using out-of-the-box NLP tools. We analysed their capability by comparing gaps and overlaps among ours, DBpedia, and Wikidata, as well as other biographical ontologies. We outlined how much new information could be captured and to what extent it could improve the quality of existing LD (RQ3.1). RQ3.2 is answered by a comparison with RDF instances and biographical

[26] https://wordnet.princeton.edu/ (Accessed on 2023-07-10)
[27] https://www.wikidata.org/wiki/Property:P5102 (Accessed on 2023-07-10)
[28] https://www.w3.org/2022/08/rdf-star-wg-charter/ (Accessed on 2023-07-10)
[29] https://www.w3.org/TR/prov-o/ (Accessed on 2023-07-10)

ontologies, although the quality of LD is still hard to measure objectively. Providing the list of predicates with the semantic frames, we also demonstrated a useful bottom-up approach for ontology designing.

Our contribution lies in the fusion of *close-reading* and *distant reading* for the LG generation with NLP tools. We delivered higher level of detail about the LD as opposed to its statistical overview. At the same time, several limitations were identified in our research, especially due to the simplification of tasks, the capability of tools, and the complexity of Wikipedia articles. Noise can be eliminated by data analysis and subsequent heuristics. Although the interesting literals were retrieved and showed a potential value for researchers, more advanced IE is required. In combination with ML, we hope that the NLP workflow like ours makes it possible to supplement and enrich existing LD.

5

Wikidata Visualization for Event and Temporal Data Exploration in Digital Humanities and Cultural Heritage

Temporal data are crucial in many research activities within Digital Humanities (DH) and Cultural Heritage (CH). At the same time, there is a growing need for visualization tools to explore events in Linked Data (LD). This chapter investigates LD visualization tools that address event data in association with temporal data for research purposes in DH and CH. We analyze the availability of Wikidata visualization tools capable of handling event data for DH and CH research and propose ways to improve visualization tools to better represent temporal data in Wikidata. We identified 14 requirements based on principles from the information visualization domain, as well as an analysis of previous studies and existing tools. We design and develop a Wikidata-centric tool to meet these requirements. This tool is then evaluated through focus groups and questionnaires with DH and CH experts. The results of the evaluation show overall positive feedback and highlight the implicit need and value of visualization tools that handle events in Wikidata for research purposes in DH and CH. In addition, they indicate the improved accessibility and visualization capabilities of Wikidata through the seven time-related functionalities of the tool.

.

This chapter is partly based on *Go Sugimoto, Victor de Boer, and Jacco van Ossenbruggen* Wikidata Visualization for Event and Temporal Data Exploration in Digital Humanities and Cultural Heritage. (Submitted for review in April 2025)

5.1 INTRODUCTION

This chapter mostly focuses on the tool quality, addressing RQ4: "What are the effective designs and functionalities for Linked Data tools to support research using temporal information in Cultural Heritage and Digital Humanities?" (Section 1.5). The main stakeholders are developers. However, data consumers play an end-user role for tools. There is also an element of data quality in terms of LD data source, therefore, data consumers as well as data producers should not be forgotten.

Research domains such as the humanities and Cultural Heritage (CH) often rely on temporal data, such as dates and information based on time, for investigations in that domain. This also applies to Digital Humanities (DH), where digital technologies allow efficient management and processing of large, heterogeneous datasets at the intersection of humanities and computer science [143]. For example, Glinka et al. [81] stress the importance of the temporal dimension in cultural collections, creating a timeline-based visualization tool to explore art history focusing on drawings.

Similarly, Dörk et al. [71] explore challenges concerning large-scale visualization from diverse cultural institutions in Germany, using a timeline view among three other views. This type of project exemplifies DH in the Big Data era, where visualization plays a critical role in making sense of data that humans cannot easily consume and process. Windhager et al. [141] report an emerging number of visualization projects in DH in which 81% visually encode the temporal dimension in CH data. Chronological visualization is more specifically discussed by Davis et al. [60] in the context of arts and culture.

On the other hand, Linked Data (LD) has been at the forefront in developing new interfaces for users to experience DH and CH [72]. LD is a set of best practices for publishing and interlinking structured data (such as RDF "triples"[1]) on the Web, creating a connection between data and its contexts, which could lead to the development of intelligent search engines [112]. As such, a high volume of LD has been produced: in 2020, there are at least 202 billion triples over 1151 datasets [112]. In the same way as in DH, the need for visualization is also growing within the LD communities [53, 112]. Therefore, visualization tools that adequately handle temporal data are needed when using LD in DH and CH.

In the LD community in DH, the need to encode and create events has been identified. The concept of the event is at the heart of historical knowledge modeling in several projects [106]. Studies suggest the need for supporting events [118] as well as event gazetteers in LD [94]. For LD in history, Hyvönen et al. [94] argue that specific gazetteers for events with rich semantic structure are needed, because “they link actors, places, times, objects, and other events into larger narrative structures”. There is also increasing interest in extracting events from text for temporal visualization and exploration of archival records [118]. In this context, there will be a need for visualization tools for events in LD.

In this chapter, we explore visualization tools that deal with event data in association with temporal data for research purposes in DH and CH. To this end, we use Wikidata as a case study.

Wikidata [27] is one of the largest LD sources to date (more than100 million items), holding data relevant to DH and CH researchers: about 10 million people, 3 million places,

[1] A data structure to describe a single fact or statement in a graph, consisting of subject, predicate, and object.

and 3 million occurrences (a superclass of events) [136]. In addition, it has played an essential role in the development of LD, with many DH and CH projects using it for interlinking and semantic enrichment [143].

We can summarize the challenges and potential for LD tools in DH and CH research and translate them into the Research Question: "What are effective interface designs and functionalities for Linked Data tools to support research using temporal information in Cultural Heritage (CH) and Digital Humanities (DH)?"

To answer this Research Question, we take the following three steps. 1) We investigate the current landscape of Wikidata applications through desktop study (observation and literature). In addition, we study how events and temporal data are encoded in Wikidata. Then, we define user requirements. 2) Based on the requirements, we design a new tool called ***ReKisstory***[2] which includes functionalities for event-based temporal data in Wikidata. 3) We conduct user evaluation with humanities and CH researchers in order to measure a) their Wikidata experience in general and b) the satisfaction level of the new tool.

Section 5.2 explores the existing tools and Wikidata's event modeling. Section 5.3 describes the requirements and design of the new tool, especially time-related functionalities. Section 5.4 presents the setup of the user evaluation and the results. Section 5.5 discusses the limitations and future work. Section 5.6 concludes the article.

5.2 PREVIOUS WORK

5.2.1 LITERATURE IN DH

To address the Research Question, we first assess a systematic review of DH projects using Wikidata between 2016 and 2023 [144]. This review examines 50 projects utilizing Wikidata, and identifies three key conclusions: a) Wikidata is conceptualized as a content provider, a platform, and a technology stack; b) it is employed across various domains, including annotation and enrichment (17 projects), metadata curation (13 projects), knowledge modeling (7 projects), and Named Entity Recognition (NER) alongside other miscellaneous uses (13 projects); and c) while most projects primarily consume data from Wikidata, there is significant potential for leveraging it as a platform and technology stack for data publication and exchange. Unfortunately, tools for Wikidata are not central to the review; only a few projects primarily discuss new tools. A European survey in the arts and humanities [59] and a linked data survey for the CH domain [121] also highlight a greater focus on finding and accessing data rather than tools.

Tools for visualizing temporal data in Wikidata have not been clearly recognized. The lack of discussion of the tooling and its relation to temporal data (and event data) becomes apparent. This implies limited experience for DH and CH researchers in using Wikidata tools in general. In this sense, the need for temporal data visualization is not yet widely collected and analyzed in the domain.

5.2.2 WIKIDATA TOOLS

In this section, we investigate the tools developed by the Wikidata community. There are 49 visualization tools listed on Wikidata's official website [34]. There are many tools (potentially) using temporal data. They include Crotos [4], LOD4Culture [135], Linked

[2] `https://rekisstory.labs.vu.nl` (Accessed on 2025-03-20)

People [11], GeneaWiki [7], Entitree [6], Conzept [3], OpenArtBrowser [12], Scholia [20], Wikidata Tree Builder [31], and Archive guide to the German Colonial Past [23]. There are generic Wikidata tools like Reasonator [19]. Typically, these tools show temporal data in the visualization. However, it is mostly simple and has limited means of interacting on the time dimension, for example, querying events from a specific time period. Tools with more advanced event and temporal functionalities are limited, but we summarize them below:

Histomania [9] is a browser that visualizes, measures, and compares events in a timeline. It provides a large amount of data to the user, although the use of Wikidata is unclear, the data is overwhelming, and the design is outdated. *ViziData* [25] shows the aggregation points of spatiotemporal data on a map. However, the public demo only contains a distribution of the dates and places of death for human entities. *Wikidata tempo-spatial display* [30] shows tempo-spatial information, including a timeline of events and a map, although the interface does not necessarily provide a modern design. Only one example is provided. *Wikidata Visualization* [32] is similar to the Wikidata Query Service, which is an official service by Wikidata [28]. It provides a generic interface for switching multiple visualizations, including table, chart, image gallery, network, pivot table, and map. However, SPARQL [22] knowledge is required to query the data. *Histropedia* [10] is a tool to display multiple time intervals of the Wikipedia category articles. It shows the potential of interactive timeline visualization and crowd-sourcing. By saving and sharing customized timelines, communities can develop an array of history timeline books. *yaap!* [35] is a tool for Wikidata capable of creating timelines quickly and easily without the knowledge of SPARQL. Interestingly, the search results display both explicit and implicit events in Wikidata in a timeline (see Section 5.2.3). It is also possible to add new search results to existing ones. However, no other views are provided.

The downside of the Wikidata tool list is that it may not be academic-oriented and user evaluation is normally not included in the scope of the project. As Po et al. [112] indicate, many are experimental rather than pragmatic services for end-users. The study of Turki et al. [130] highlights the dominance of computer science research for Wikidata compared to other domains; the number of Wikidata-related research publications in arts and humanities is just over a tenth of that of computer science. Although this study is not limited to the use of tools, the potential of Wikidata for research use in DH and CH is not yet fully investigated.

Similar to Section 5.2.1, the limited landscape of LD tools tailored to temporal data and event data in DH implies limited LD experience among DH and CH researchers. The shortcomings of user evaluation for those tools also underpin this situation. Therefore, it would not be easy to collect the user needs without presenting a concrete application. For this reason, we made a decision to develop a Wikidata tool on our own first, based on the findings of previous studies. Then, we conducted a researcher's evaluation of the tool to gain "hidden" insights into user needs for the Wikidata visualization tools for handling temporal data, especially supporting event information.

5.2.3 HANDLING WIKIDATA EVENTS

This section outlines how events are encoded in Wikidata.

In Wikidata, events are not only explicitly encoded as the subject or the object of a triple/statement, but also implicitly as a part of a statement. Qualifiers in Wikidata serve as an addition to a statement to include events. There are primarily three ways to encode events

in Wikidata. We shall call them Type A, B, and C.

In Type A, an event is an independent entity. Certain properties of the entity support the inclusion of time (e.g. birth date, inception, or time of discovery[3]) and space (e.g. birth place, place of discovery). Time is typically encoded as a typed literal. In Listing 5.1 and Figure 5.1 (left) the event entity of Siege of Vienna (wd:Q7510505[4]) that appear as a subject. It holds two properties and objects: start time (wdt:P580) and its value "1485-01-29T00:00:00Z^^xsd:dateTime"), and location (wdt:P276) and its value (Vienna: wd:Q1741)

In Type B, an event is implicitly attached to an entity of other types through the time-space properties. The other types can be anything such as a person, a building, or a material. In Listing 5.2 and Figure 5.1(right), the birth event can be implicitly found in the properties of David Bowie (wd:Q5383), which is a person type. While the property of birthplace (wdt:P19) indicates a location by reference (wd:Q146690), the property of birth date (wdt:P569) provides its literal value ("1947-01-08T00:00:00Z"^^xsd:dateTime).

In Type C, an event is implicitly encoded as the time-qualifiers of objects of a statement. In Listing 5.3 and Figure 5.1 (right), the property "work location" (p: P937) defines an event by establishing an intersectional relationship between person (David Bowie: wd:Q5383), time ("1976-01-01T00:00:00Z"^^dateTime), and place (Berlin: wd:Q64). Temporal and spatial data appear as qualifiers, if available[5].

One reason why events are encoded in various ways is that Wikidata does not employ an event-centric data model. This is different from the event-centric ISO standard called CIDOC-CRM [1], which is widely used in the CH and DH domain.

[3] `https://www.wikidata.org/wiki/Help:Dates#Properties` (Accessed on 2025-03-20)

[4] Q7510505 with a namespace (wd) is the ID for “Siege of Vienna” in Wikidata. The same goes for the other IDs (Q for entities and P for properties) below

[5] For instance, the time when Bowie worked in New York City is missing in Wikidata (Figure 5.1).

Listing 5.1: Type A event for Siege of Vienna

```
wd:Q7510505
        wdt:P580 "1485-01-29T00:00:00Z"^^xsd:dateTime ;
        wdt:P276 wd:Q1741 ;
```

Listing 5.2: Type B event (birth) for David Bowie

```
    wd:Q5383
        wdt:P19 wd:Q146690 ;
        wdt:P569 "1947-01-08T00:00:00Z"^^xsd:dateTime ;
```

Listing 5.3: Type C event (work) for David Bowie

```
    wd:Q5383 p:P937 s:Q5383-E036F97C-D954-4E18-B874-EC5A1BA0866B .

    s:Q5383-E036F97C-D954-4E18-B874-EC5A1BA0866B a wikibase:Statement
        ps:P937 wd:Q64 ;
        pq:P580 "1976-01-01T00:00:00Z"^^xsd:dateTime ;
```

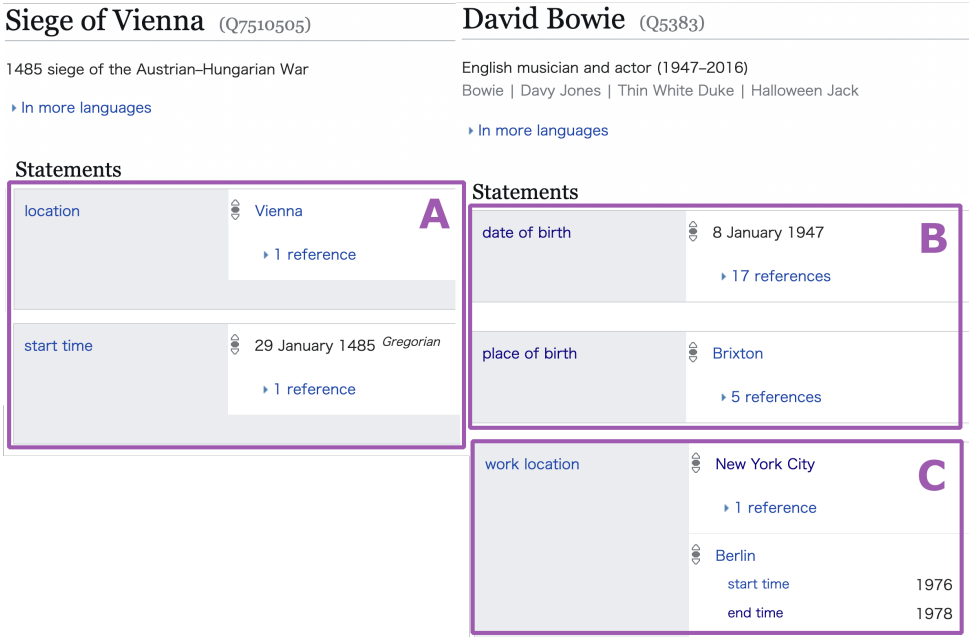

Figure 5.1: Three types of Wikidata events: type A event of Siege of Vienna (left) and Type B and C events (birth and work) of David Bowie (right) `https://www.wikidata.org/wiki/Q7510505` `https://www.wikidata.org/wiki/Q5383` (Accessed 2025-03-20 Screenshot image modified)

5.3 DESIGNING REKISSTORY

5.3.1 SCOPES AND REQUIREMENTS

Given that previous studies describing the need for the use of temporal data for Wikidata are highly scarce in DH and CH, it is a challenging task to identify requirements. However, we define them on the basis of supporting references and evidence.

We specify seven requirements (R1-R7), taken from Windhager et al. [141], who discusses various perspectives on the information visualization domain, focusing on CH. We describe an additional seven requirements (R8-R14) from our analysis of the literature and existing tools as described in Section 5.2. While the first seven requirements (R1-R7) deal with high-level conceptual or theoretical visualization requirements [141], the second seven requirements (R8-14) focus more on specific needs for (temporal) data exploration for Wikidata. In the requirements below, we refer to relevance to our tool, which is also found in the subsequent sections. We then evaluate the given functionalities of the newly developed tool aiming to address those requirements:

R1: Serendipity. This requirement is about "finding valuable or agreeable things not sought for" [141]. The tool should encourage the experience of unanticipated discoveries and outcomes while exploring data. However, as Windhager et al. [141] put it, there is no established implementation recipe for this elusive term. Therefore, this requirement is not specifically featured in our tool, although it provides flexible ways of exploring LD through various user input methods, taking advantage of semantic links in LD.

R2: Generosity. Whitelaw proposes "generous interfaces" that emphasize a more humanistic browsing experience, visual exploration, and interpretation in addition to information retrieval tasks that are typically represented by search interfaces [139]. This requirement deals with methods to provide access to data in such a way that users can understand its scale, richness, and complexity. Rather than functional needs, the concept of generosity places emphasis on process, pleasure, and thoughtful engagement to overcome narrow task- and deficiency-driven approaches to interface design. This requirement is not the focus of our tool. However, there are some elements of generosity in a broad sense. We create requirements for balance between simplification and expressivity (R8), scalable, controllable, and customizable user interface (R11), and user-oriented guided navigation (R14).

R3: Criticality. This requirement covers critical reflections and design strategies [141]. Criticality addresses both user literacy skills for end users and self-reflection for tool developers. Among the elements highlighted by Windhager et al., our tool aims to address this requirement through characteristics of LD that can provide multiple perspectives ("plurality") and provenance information ("disclosure")(F7). There are also related features in our tool to enhance criticality: grouping of items (F5), and age calculation (F2).

R4: User guidance and narration. This requirement concerns suggestions for the extension of a viewing experience as well as narratives that offer pathways of sensemaking for CH collections. For example, digital CH collections can be explored using user- and/or algorithmically derived recommendations [141]. In our tool, this is translated into a more concrete requirement from our literature observation: user-oriented guided navigation (R14), as well as the Wikipedia narrative found next to LD exploration.

R5: Remote access vs. being there. This requirement addresses the question of how to implement a system that facilitates not only remote access, but also physical access to CH

collections. The lack of narrative guidance in conventional CH interfaces has been discussed in the CH community in this regard [141]. The requirement is not directly related to our case that focuses on Wikidata and its tools on the Web.

R6: Facets of uncertainty. This requirement focuses on uncertain data that are often found in many DH fields [106]. The problem is that the lack of precision and interpretative openness has not been adequately acknowledged in the design of CH interfaces and visualizations [141]. For example, the challenges for the uncertainty on dates range from determining the exact date and different concepts of time, to deciding what date should be recorded and represented. In our tool, handling various time concepts (R13) and age calculation (F2) are relevant.

R7: Contextualization. Windhager et al. [141] specifically mention LD as a new option for "enhancing, contextualizing, linking, and reframing" cultural heritage objects. Through LD, we can identify entities including cultural objects, creators, institutions, and events, drawing typified (temporal, spatial, contextual, and conceptual) links between them. As such, LD assists the processes of sensemaking by connecting fragmented CH data and providing cross-domain representations and reasoning. In our tool, the requirement is highlighted for the focus on events in LD connecting various entities (R10), timeline overlaps (F1), and an extra timeline for period or era.

In addition to those generic visualization needs, we add more practical points (R8-R14) derived from our tooling analysis in Section 5.2.2 and additional relevant literature.

R8: Balance between simplification and expressivity. At the data level, the balance between simplification and expressivity is crucial. As seen in applications such as Histomania [9], information overload can be a problem. Simplification is vital to assist with sensemaking. In contrast, many Wikidata tools only provide a simple timeline view. More expressive and inclusiveness enhance generosity (R2). At the user level, this requirement addresses a challenge concerning the varied competence levels of end-users. Given the potentially limited experience of DH and CH researchers in using the Wikidata tools, it is sensible not to limit users by their LD skills. Novice and advanced Wikidata users should be equally targeted. In particular, Hogan [90] argues that techniques for building LD applications for novice users have not been properly explored. In addition, a set of guidance and documentation would be helpful for the users in solving this issue.

R9: Flexible user input. This requirement concerns more options for user input and interface. For example, despite the abundance of timeline visualizations, rather strangely, the input interface for the time dimension is not clearly identified in the Wikidata tools. As exemplified in [71, 81], it is convenient for users to specify time to visualize and explore data in the humanities. The flexibility of user input in the interface enhances the sense of user control and engagement. Other types of flexibility include a variety of input (search parameters) and multiple types of interface to visualise the same data (text input, drop down menu, drag and drop etc).

R10: Event exploration, especially for implicit events. This requirement corresponds to the need for event gazetteers [94] and event modeling [118] in LD. Time should be explored in relation to space (if possible), implying a retrieval of events as an intersection of time-space. As other entities, including persons, groups, objects, and buildings, can be attached to an event to provide more context about it, LD's network capability is well suited

to model events. Event-centric data models have also been discussed with regard to R7 [141]. Services like Wikidata's tempo-spatial display and yaap! also aim to meet this requirement.

R11: Scalable, controllable, and customizable user interface. This requirement relates to the user interface for the visualization of a large amount of LD. It is related to R2. Time-related information should be accessible on a scale, be easily controllable, and customizable using a good user interface. As the example of Histomania indicates, the amount of information may be overwhelming to the user, a modern designed interface is long due. yaap! has many functionalities to customize the user settings, including languages, date formats, time scales, colors, and labels.

R12: Multiple visualizations. We observed that Wikidata Query Service and Wikidata Visualization offer multiple views of SPARQL search results. They display multiple visualizations such as table, chart, network, and map view. Thus, this requirement is aligned with the growing popularity of multiple views and perspectives. The survey of Windhager et al. [141] indicates that about 80% of all interfaces in the survey use more than one non-temporal visualization method including lists, grids, maps, and networks.

R13: Handling of various time concepts. This deals with the complexity of date concepts in the humanities, including uncertain dates (which are related to R6), descriptive dates, ordering relations among time instants and intervals, and non-Gregorian calendars. The proleptic Gregorian calendar[6](which can be applied to Before Christ/Before Common Era (BC/BCE)) is relatively unexplored in LD due to the lack of awareness and element of complexity. For example, the 2006 version of OWL-Time only supports the Gregorian dates and times [58]. However, yaap! is capable of displaying (but not calculating) the dates in the proleptic Gregorian calendar in timeline. See also the limited availability of numeric date information in LD in Sugimoto [124] and Time Ontology in OWL [57].

R14: User-oriented guided navigation. This focuses on the starting point of the user interface and is related to R2 and R4. To explore tens of millions of items in Wikidata, search interface is inevitable. However, as Windhager et al. [141] point out, a search-box interface typically found on the Web the search engines is not sufficient. More user-oriented guided navigation and exploration methods should be found for a new user experience. For example, tools such as Crotos, Linked People, and OpenArtBrowser provide featured items with images next to the search box interface.

We have designed a tool called ReKisstory, which we present in Section 5.3.2 aiming to address as many requirements as possible. From now on, we refer to them when they are relevant to the functionalities of the tool.

5.3.2 CENTRAL DESIGN OF THE TOOL

In a nutshell, our application is designed to facilitate the exploration of events in Wikidata through the three types of events described above. Two major functionalities were implemented in the tool: 1) *Compare* and 2) *Find*.

The *Compare* section enables users to select and compare the events related to Wikidata items and visualize them. It offers various visualizations. For example, the biographies of Paul Gauguin and Vincent van Gogh can be compared (Figure 5.2). Users can change the language of the Wikidata-provided data labels. *Compare* aims for R1 as users can compare any combination of up to four items among 100 million items in Wikidata. A broad range

[6]It extends the Gregorian calendar backward to include dates before the adaptation of the Gregorian calendar.

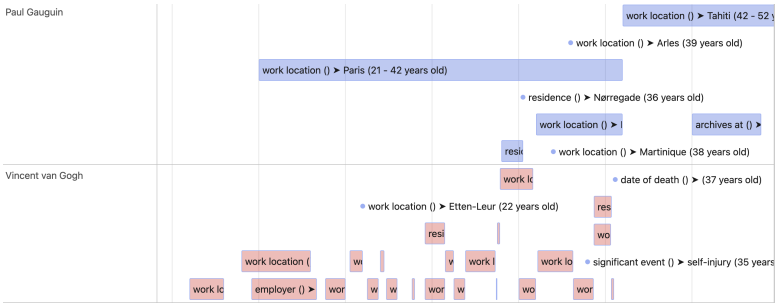

Figure 5.2: Timeline view in *Compare*. *Compare* allows users to analyze max four Wikidata items in the timeline (example of Paul Gauguin in blue and Vincent van Gogh in red. The age of an entity at a point in time is shown, if calculation is possible.)

of item types can be compared, including humans, animals, plants, products, companies, astronomical objects, events, artworks, theories, places, buildings, and chemical compounds, as long as they are associated with events. Unexpected patterns, including similarities or differences may be discovered by features described in Section 5.3.3.

The *Find* section enables users to list all Wikidata items that match a query pattern. yaap! does not offer graph pattern searches[7]. As a bare minimum, users specify the type of item to look for, for example, human (wd:Q5[8]). They can further specify characteristics of the item (sex/gender, citizenship, etc.), as well as an action and object (i.e., predicate and object in a triple). In addition, a point in time or time span of the action can be added (start and end time), dealing with R9 and R13. For example, users can search for human items whose occupation (wdt:P106[9]) is musician (wd:Q639669) and whose residence (wdt:P551) was New York City (wd:Q60) from 1960 to 1979 (Figure 5.3). This pattern is more or less the representation of events typical in historical research, often defined in terms of people doing an activity at a certain place and time [106]. Users can adjust the level of query granularity by changing the number of inputs (e.g., with or without specifying dates or object in the triple). This implies greater flexibility and customizability for users (R11).

We made a design choice to use auto-suggest for user input as much as possible (Figure 5.4), which is related to R9[10]. It is implemented by the Wikidata Reconciliation API [29], which allows users to look up and select a multilingual Wikidata item or property very easily based on a browser language. Thus, no knowledge of SPARQL is required to search, although it helps to formulate query patterns in the *Find* section. We expect that this setup should reduce the complexity of using Wikidata considerably (R8). The application processes the results of SPARQL queries on-the-fly and dynamically renders them.

The results are presented in multiple views: table, timeline, map, network, and Wikipedia (R12). Instead of hiding them (like Wikidata Query Service), we clearly provide them in five tabs by default. The Wikipedia articles can be read in the tab. This view helps to add

[7] In addition, yaap! does not have multiple views such as map (see below)
[8] wd is a name space for a Wikidata entity: http://www.wikidata.org/entity/
[9] wdt is a name space for a Wikidata property: http://www.wikidata.org/prop/direct/
[10] Drop down menus are also provided for predefined user input on other occasions

generosity (R2), narratives (R4), and context (R7) to our graph database approach.

As an RDF statement can have multiple qualifiers, it is possible to have different views or perspectives for a statement in Wikidata. For instance, if two possible dates are known for an event (e.g. historians can not specify one), they will be separate results. It allows us to address plurality, multiple perspectives, or polyvocality [119] seen in R3.

The tool is designed to be as generic yet detailed as possible to find a middle ground for a) CH and humanities experts and casual users [141], and b) LD experts and technical novice users. Two modes (simple and advanced) are provided in *Compare* and *Find* for this purpose (R8, R11), offering different input options. For instance, the context timeline option enables users to add a period or era in timeline view to provide context or background for the search results (R7). We experimented this with the periods of art movements, Popes, and US presidents.

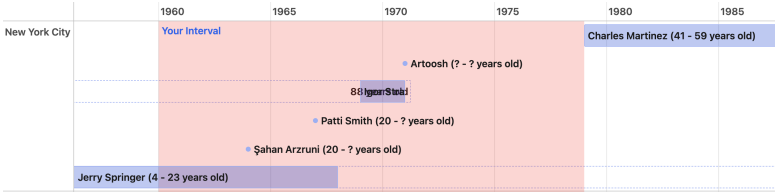

Figure 5.3: Timeline view in *Find*. *Find* allows users to analyze the Wikidata items matching the user defined pattern in the timeline (musicians who lived in New York from 1960 to 1979). The red area indicates the user-specified time (1960-1979). Overlaps among the search results (musicians in blue bars or dots) and user-specified time can be studied

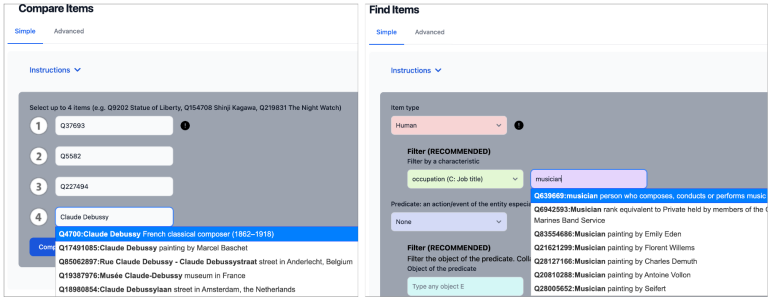

Figure 5.4: Auto-suggest-based search interface of *Compare* (left, simple input fields) and *Find* (right, more complex triple pattern input fields)

5.3.3 SEVEN TIME-RELATED FEATURES

We concentrate on examining seven time-related functionalities (F1 to F7) found in the *Compare* and *Find* sections of the tool, with which we cover as many requirements as possible.

F1 Two Time Periods Overlaps (R7, R12, R13): In the timeline result view, there are a) the solid lines representing the time of event defined by the user (Type B or C event), and b) the dotted lines representing the lifespan of an item (Figure 5.5). This becomes handy in *Find* to study how two items (e.g., persons) overlap with each other in terms of a

property (e.g. residence) and the lifespan (Figure 5.3). By providing two periods, we support contextualization (R7), and, to a certain extent, multiple visualization (R12) and handling of various time concepts (R13).

Figure 5.5: An example of F1: Two Time Periods Overlaps. The solid lines (i.e. blue bar) are the duration of residence of Princess Diana in Althorp. The dotted lines represent her life span. The red area indicates the user specified duration.

F2 Age Calculation (R3, R6, R7, R13): The tool displays the time calculation of an item in the table, map, and timeline view. The age of an item for a specific date (the start and end date of the action / prediction) is calculated by the birth and death date (Figure 5.6). This function typically makes sense for humans, but it can also show the life span of a group, object, building, or city in the form of the duration between inception and dissolution.

For Anno Domini/Common Era (AD/CE), month-level calculation is provided. We also indicate if a date is before, during, or after the lifetime, even if one or more dates are missing (start date, end date, and target date). Therefore, a certain degree of uncertainty can be represented. In addition, F2 allows us to sense time lags of items: for example, a scholarly work is written 100 years after the birth of a person (*after lifetime*) who is the subject of the work (see also the distinction between primary and secondary sources in historical research in [106]). This functionality deals with R3, R6, R7, and R13 in the sense that they provide the critical context of an item with potentially complex dates handling.

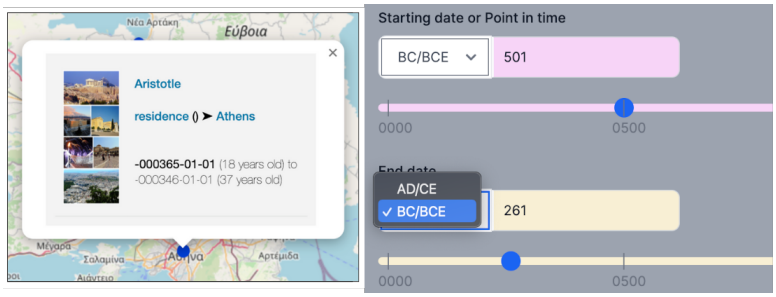

Figure 5.6: An example of F2: Age Calculation (left). The age of Aristotle when he started and ended his residence in Athens on the map view. An example of F3: BC/BCE Support (right). User can specify BC/BCE to find prehistoric items. Time slider input is partially seen in the advanced search in *Find*.

F3 BC/BCE Support (R13): It may not be common to see web search engines capable of processing non-Gregorian calendars, but we see an added value [58] providing support for R13. In particular, Before Christ/Before the Common Era (BC/BCE) is crucial (Figure 5.6). To support domains like archaeology, the tool allows users to search for items with dates in the proleptic Gregorian calendar. The need for the functionality can be verified by Histomania and Histropedia which can display items of BC/BCE in the timeline. The uniqueness of this functionality lies in arithmetic calculations based on user input Figure 5.6.

F4 Remove Items in Timeline (R8, R11): A simple yet important feature is to move or remove items in the timeline (Figure 5.7). It can minimize the problem of information overload (R8 and R11), which can be a serious problem in visualization [112]. yaap! is also equipped with this functionality.

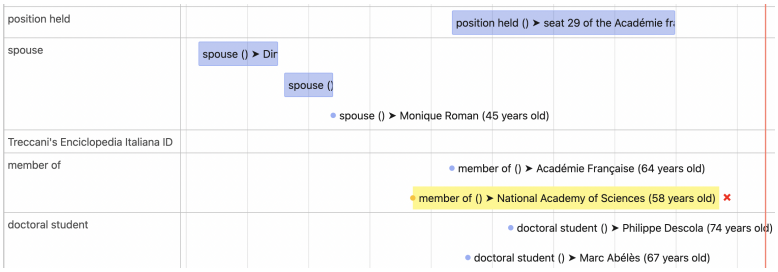

Figure 5.7: An example of F5: Grouping of Items and F4: Remove Items in Timeline. Users can change the data grouping (left column) by relation (i.e. predicates) or searched entity in timeline. This example shows the events for Niels Bohr in timeline. Data is grouped by relation (position held, spouse etc). Compare this to Figure 5.2 focusing on grouping by entity. Users can also move and remove items in the timeline by selecting an item (yellow highlighted)

F5 Grouping of Items (R3, R11): The grouping of the results in the timeline is shown based on user input (and is customizable in *Compare* (R11)). It provides a different perspective for item comparison (R3). In *Find*, temporal spatial analysis can be performed by grouping, for example, by city (the objects of the triples). Customization is either grouping by item/entity, or relations (i.e., predicate of a triple)(Figure 5.7). In yaap!, the Wikidata items are only grouped by predicates.

F6 Time Overlaps (R9, R12, R13): The tool is designed to compare and analyze the time overlap among items (Figure 5.2). It provides flexible input of time in different interfaces (R9). Users can specify a point in time (start time) or duration (both start and end time), by typing dates in ISO format (YYYY-MM-DD) or time slider (Figure 5.6 (left))

In Figure 5.3, the red area is the user-defined interval, which can be compared with the durations (and points in time) of different items. However, this functionality is very rare in the existing tools. Although interval overlapping is visualized in applications such as yaap! and Histropedia, R9 is missing for them, limiting the true capability of LD for the users. In a sense, this also concerns R12.

In relation to R13, we would also like to support queries in terms of overlap relations between temporal intervals, and this requires technical knowledge to implement in SPARQL. To provide this search functionality, four types of arithmetic calculations are required to search for time overlaps: a) a point in time against a point in time, b) a point in time against an interval, c) an interval against a point in time, and d) an interval against an interval. Figure 5.8 summarizes all patterns[11]. For a point in time, the user input (Date X) is searched against four possible time overlaps (Date x and y in A, B, C, D) (Figure 5.8 (left)). For an

[11] Patterns for an interval against a point in time are omitted in the figure, as it can be derived from an interval against an interval

interval, the user input (Date X and Y) is searched against nine patterns in four groups (Date x and y in A, B, C, D) (Figure 5.8 (right)).

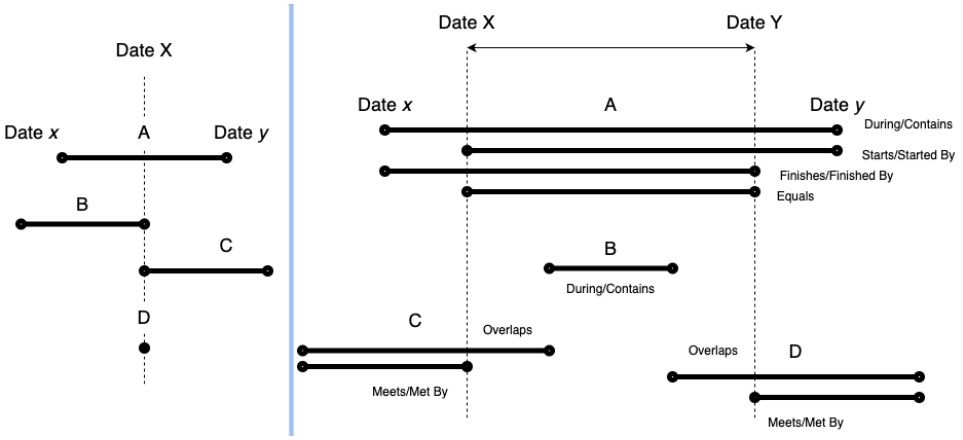

Figure 5.8: Time calculation patterns for two user input possibilities (Date X (a point in time) or Date X and Y (interval). Date X is searched against four types of temporal data (A,B,C,D) (left). Date X and Y are searched against nine types of temporal data in four groups (right)

F7 Provenance Information (R3): The significance of provenance is stressed by many [67, 103, 106] in the sense of criticality (R3). de Boer et al. [51] state that provenance information is vital to provide insight into possible data manipulation by tools and to establish the credibility of the data. A notable contribution is that our tool offers a direct hyperlink to the Wikidata statement (i.e., RDF triple) rather than a link to an item/web document (Figure 5.9). In other words, users can check exactly where an explicit or implicit event is described in a Wikidata page (see Figure 5.1). In the context of R3, this feature is related to fact-checking and transparency.

The Night Watch	locatie	Kloveniersdoelen			1642-01-01	0 years 0 months old (during lifetime)	1715-01-01	73 years 0 months old (during lifetime)	Wikidata	Check this fact
The Night Watch	datum van oprichting of creatie				1642-01-01	0 years 0 months old (during lifetime)		? years ? months old (during lifetime)	Wikidata	Check this fact
The Night Watch	locatie	Paleis op de Dam			1715-01-01	73 years 0 months old (during or after lifetime)	1808-01-01	166 years 0 months old (during or after lifetime)	Wikidata	Check this fact

Figure 5.9: An example of F7: Provenance Information. Link to the RDF statement for The Night Watch is provided as “Check this fact” (far right column) in the table view

Despite the large number of tools found in Section 5.2, there is no tool that provides all seven time-related features. Most tools deliver one of two features at most, except yaap!. These seven functionalities have been designed to address the challenging requirements

in Section 5.3. The user evaluation in the next section sheds light on the usability and usefulness of the tool.

5.3.4 IMPLEMENTATION

ReKisstory is developed in Python, JavaScript, and CSS. As it is primarily a visualization tool, it processes and renders data on-the-fly using various external APIs. This means that we do not store any data in the backend. However, the search results are processed in the frontend and can be downloaded as a CSV file.

5.4 USER EVALUATION

5.4.1 METHOD

We organized workshops as online focus groups to evaluate whether the seven time-related functionalities (F1-F7) meet user expectations in the humanities, DH, and CH communities. We used a combination of a pre-workshop survey, a series of workshops, and a post-workshop survey. We choose these methods because we would like to a) fully explain the scope of the tool and evaluation to the participants, b) assess the tool as a whole, including the seven functionalities, and c) balance qualitative and quantitative analysis.

The focus group allows us to gain detailed information about the tool, by meeting potential users and concentrating on qualitative analysis. This method is especially effective for gathering shortcomings and "early" user requirements of the tool. The pre- and post-workshop surveys are conducted through two detailed questionnaires (the first and second questionnaire [126]. They were created in Google Forms[12]. They help us to collect feedback systematically and accurately to quantify the results, which is more efficient than oral communication in the focus group.

PRE-WORKSHOP SURVEY

In May 2024, several calls for participation were sent to the mailing list of the Europeana research community, of which many CH and DH experts are members[13], as well as groups of humanities experts in the circle of the authors' colleagues via LinkedIn[14]. We called the workshop "A workshop for a new web app for Humanities and Cultural Heritage". Although we provided a brief description of the tool, we deliberately avoided mentioning LD as much as possible to welcome a broad range of potential users. There were 52 responses and one duplicate; thus, 51 responses to the questionnaire were valid for analysis (Table 5.1).

This questionnaire [126] was designed to collect information about the background and experiences of the participants. It contains questions about the demographics of the participants, experience with time-related data, Wikidata, and SPARQL. A consent form was created following the EU regulation for data protection (GDPR[15]) to ask permission of the participants to use the data for research purposes.

In the first questionnaire two questions about small research tasks were included. The two questions are: 1) How easy is it to find the answer to the question SQ1 (see below) on

[12] `https://workspace.google.com/products/forms/` (Accessed on 2025-03-20)
[13] `https://pro.europeana.eu/page/europeana-research` (Accessed on 2025-03-20)
[14] `https://www.linkedin.com/` (Accessed on 2025-03-20)
[15] `https://gdpr.eu/` (Accessed on 2025-03-20)

Table 5.1: The number of people participated in the workshops and answered two questionnaires (Q1, Q2)

	# of people
Answered Q1	**51**
Attended a workshop	16
Answered Q2	**11**

the web, and 2) How easy is it to find the answer to the question to the question SQ2 (see below) on the web.

- **SQ1**: "How old was Van Gogh when moving to Paris and how old was Paul Gauguin at that time?"
- **SQ2**: "Which musicians worked in New York between 1960 and 1979 and how their time overlap each other?".

These tasks could be effectively performed by our tool (e.g. by using F2, F5, and F6). Therefore, we asked the participants to perform these tasks during the testing of the tool in the workshop. Then, in the second questionnaire in the post-workshop survey (see below), we asked if the tool helped to complete the tasks. In this way, we can assess the change in user perceptions before and after the workshop.

These are questions about historical events we discussed earlier. They may not be exactly the research questions of researchers in this domain. However, researchers such as historians often do not have a clear research question when starting an investigation [106] and/or the question changes over time; we take them as fact-checking starting questions to be elaborated on further in the iterative research process.

WORKSHOP

Five 1.5-hour workshops targeting a group of five were held online in May and June 2024. Due to the technical limitations of an online workshop solely hosted by the first author, only a subset of the people who answered the first questionnaire could be invited. Initially, 25 people from 51 respondents were invited on a first-come-first-served basis. However, many people canceled in different stages of communication, in the end 16 people participated in the workshop.

We organized four workshops in English and one in Japanese. This setup enabled us to support diversity to a certain degree. The different languages might have an impact on the evaluation of the tool. As English is primarily used for the tool's interface, the participants had a sense of the tool in other languages only when using auto-suggest in different languages and examining search results partially in a specified language (if the results in Wikidata support the language labels). We prepared all documents and sessions in both English and Japanese. Therefore, the participants in the Japanese workshop used Japanese for all interactions. Feedback was translated into English for the assessment.

Before the workshop, several documents [126] were distributed to streamline the workshop: a) the workshop slides, b) the *Compare* section manual, c) the *Find* section manual, and d) the *Find* section example search patterns. These documents aimed to mitigate the

risk of overwhelming participants with information overload during the workshop without disturbing too much the fresh first experience of the tool. In particular, d) includes many examples of graph queries for *Find* as it can be difficult for domain experts to understand the graph database.

The workshop consisted of three parts. First, a slide presentation was given to explain the motivation and overview of the Wikidata content, as well as the overview of the main functionalities of the *Compare* and *Find* sections. Second, we asked the participants to test the *Compare* and *Find* sections as a whole without restricting to the time-related functionalities. Finally, we collected feedback in short free discussions.

After the workshop, the second questionnaire [126] asked for opinions on the tool, focusing mainly on the usefulness and usability of the seven functionalities, as well as on general remarks and room for improvement. The questionnaire was designed to measure the quality of the tool. It was possible to fill out the questionnaire during and after the workshop. Eleven responses were received.

5.4.2 THE RESULTS

This section presents and analyzes the results of the evaluation. As the raw anonymized results of the workshop can be found online [126], we focus on the key findings here.

PRE-WORKSHOP SURVEY RESULTS

In the first questionnaire, we see broad backgrounds of 51 potential users of ReKisstory. The gender balance is reasonably even. Although there is a bias toward higher age groups (over 35 years 80.4%), English speakers (19.6%), and the archives domain (21.6%), the native language and research fields are extremely diverse (Figure 5.10). The occupations vary broadly, from student, teacher, professor, consultant, and CEO, to librarian, archivist, archaeologist, Wikimedia resident, and art dealer. This is what we expected from the strategy of our call for participation. The distribution is highly suitable for our study to capture the diverse needs of users in DH. In fact, Zhao [144] reports that Wikidata is a data source for a wide range of disciplines, domains, time periods, and languages for DH projects, ranging from literary studies and history to linguistics, archaeology, and philosophy.

Among 51 respondents, 42 respondents use time-related data[16] consider it extremely important (41.5%), very important (34.1%), and moderately important (24.4%). The 42 respondents find more difficulties when they search and compare time-related data on the Web, although neutrality is the most dominant opinion (Figure 5.11).

In terms of LD experience, while 14 respondents have used SPARQL (28.0%), 28 respondents have used Wikidata (54.9%). Among those LD users, more than 75% think Wikidata and SPARQL are useful, but ease of use and learning may not be strongly agreed upon (Figure 5.12). In general, using SPARQL is more difficult than Wikidata in terms of learning and using. In particular, it is relatively clear that participants find it difficult to use SPARQL, when looking for information in LD, despite their agreement of its value (Figure 5.12). However, the reasons for neutrality is unknown. It should also be noted that SPARQL users judge their level of competence low; expert and proficient (0%), competent (28.6%),

[16]We used the term time-related data rather than temporal data in order not to exclude mixed dimensional data such as spatial temporal data.

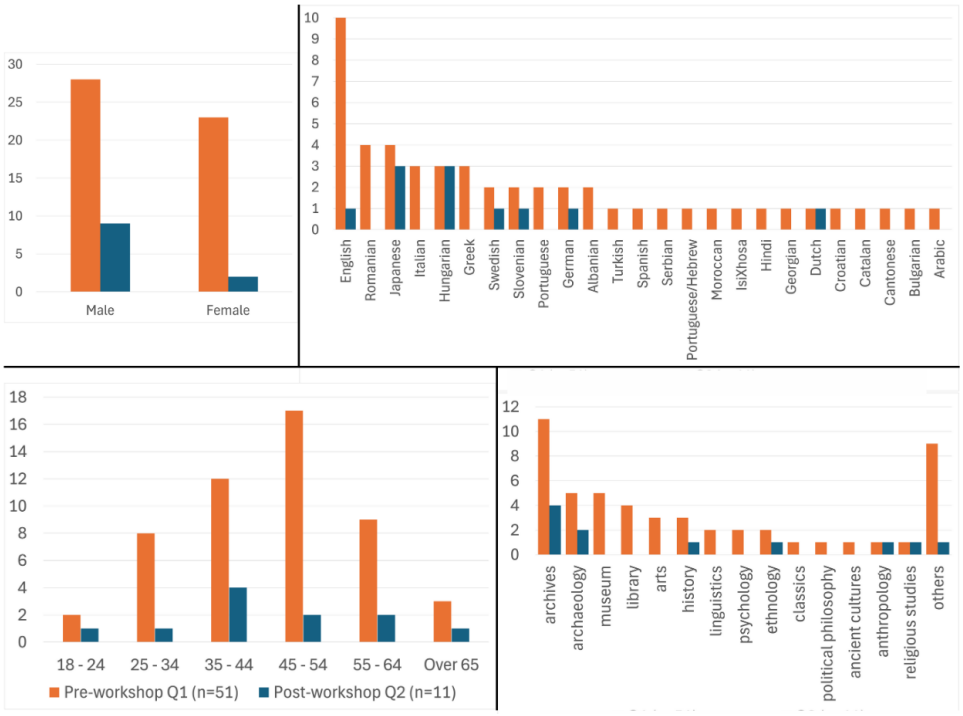

Figure 5.10: Background of respondents by gender (top left), native language (top right), age group (bottom left), and research field (bottom right) for respondents to Questionnaire 1 and 2

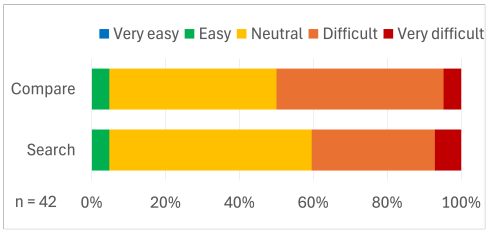

Figure 5.11: The number of answers for the question: how easy is it to search and compare time-related data on the web (among the 42 respondents using time-related data)?

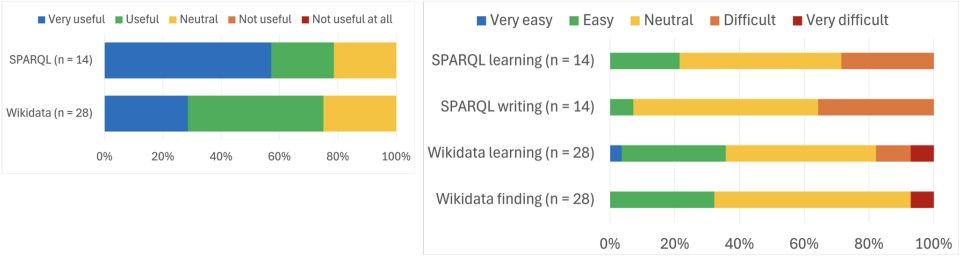

Figure 5.12: The number of answers for the questions: how useful is Wikidata and SPARQL? (left), how easy to find data and to learn to use Wikidata? (right) and how ease to write and learn SPARQL? (right)(among the 14 and 28 respondents who have used SPARQL and/or Wikidata

advanced beginner (50.0%), and novice (21.4%). This corresponds to the observation of Turki et al. on limited and narrow research scopes for Wikidata [130], which indicates the limited transfer of knowledge in DH. There could be different reasons for this: It may be due to the lack of LD and SPARQL training in DH, and/or the absence of LD tools and research, using/requiring advanced advanced SPARQL.

WORKSHOP RESULTS

The demography of the second questionnaire is similarly diverse (Figure 5.10). Although male is more dominant than the first questionnaire, the age groups are well distributed. The native languages include English, Japanese, Hungarian, Swedish, Slovenian, German, and Dutch. The respondents specialize in fields such as archives, archaeology, history, ethnology, and religious studies.

The second questionnaire focuses on the usability and usefulness of our tool for the eleven workshop participants (Figure 5.13 and Figure 5.14). The majority finds the *Compare* section of the tool useful and easy to use. In particular, the ease of use is prominent. In contrast, the *Find* functionality is considered more useful than *Compare*. However, the participants' responses regarding the usability indicate more difficulties. In terms of seven features, more than 60% of the respondents confirmed their usefulness for each feature (Figure 5.13), but the ease of use shows some diversity (Figure 5.14). In particular, F5 may not be user-friendly. Although only a few respondents have difficulties, this may be due to our requirement for the middle ground approach (R8).

We also asked why the respondents chose a specific answer. We sample and summarize their comments[17] in Figure 5.15.

It seems that negative remarks are often caused by bugs and/or limited time for testing (e.g. F5 observes mixed feedback), rather than usability, although there are occasionally difficulties in familiarizing with the interface and functionalities (e.g. interface for F3 could be improved). As there were auto-suggest malfunctions for the browser environments of some users, it was unfortunate that bugs significantly reduced user experience, impacting the user feedback. In addition, it is noticeable that the source data (Wikidata) could cause errors or confusion. Moreover, the usefulness of the functionalities highly depends on the research fields; it is not easy to provide valuable functionalities that satisfy the broad diversity of research fields in humanities and CH.

The respondents find F6 to be the most unique and F3 and F4 the least (Figure 5.16). Being asked three favorite aspects of the tool[18], they chose visualization and extra information (linking, references, provenance). Interestingly, multilingualism (provided by display labels and auto-suggest) is also voted higher than others. Data quantity and interactivity are not the primary interest.

SQ1 and SQ2 aim to measure the impact of the tool. Although 30% think it is easy to answer SQ1, a considerable number of respondents struggle to answer both questions (Figure 5.17). This tendency does not change between the workshop participants and non-participants. After the workshop, more than 72% (SQ1) and 90% (SQ2) of the participants agreed that the tool makes it easy to find the answers.

Free comments and group discussions raise some interesting questions and suggestions about the tool[19]. In general, the participants do not have strong recommendations. Positive feedback includes: "I'm very glad I got to know about this tool", "very powerful and impressive tools with great potential for making it more efficient to work with this kind of data for research". Negative feedback includes "the program has quite often failed with the error "504 Gateway Time-out"", "browser issue"[20].

The respondents suggest minor and concrete improvement such as more customization options (see R9)(e.g. selecting properties to be displayed, changing the amount of results), clearing input fields at once, alphabetically sorting drop down menu, providing drop down menu also in the advanced mode, and sharing of used SPARQL query. Positive comments include good documentation and low barriers for end users of LD (R4 and R8).

There are interesting requests for new functionalities: visually presenting the changes of relationships over time (see R1, R3, R7, and R12). In addition, a use case is proposed: the users can "input research information that previously had to be provided individually and build a unique database." by updating Wikidata by themselves (see R7).

The participants commented on *Compare* as follows: "sometimes it might be a bit difficult to understand the logic of the comparison and what it means", "I was not able to take advantage of nice functionalities because I was not familiar with the input method" (see R9 and R14), "comparisons is a core functionality in the historical research". Their comments on *Find* include the following: "could not get it to work very much to test full

[17] Errors such as typos in comments are modified. Japanese answers are translated
[18] Two respondents selected four, but all responses are included here
[19] Errors such as typos in comments are modified
[20] A problem with auto-complete

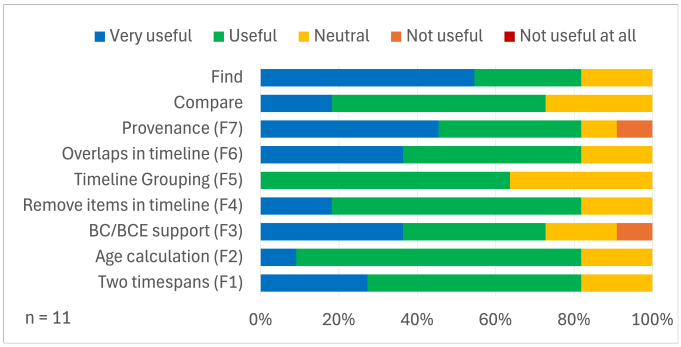

Figure 5.13: The number of answers for the usefulness of the seven features, *Compare*, and *Find*

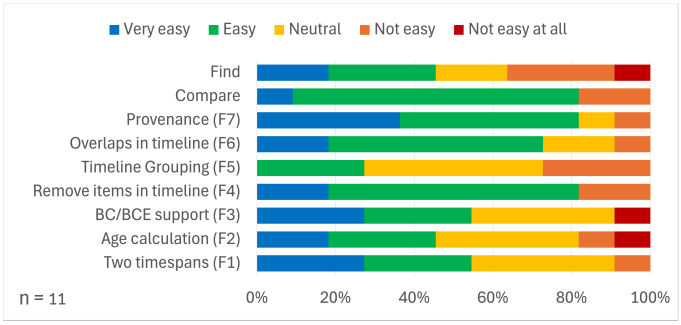

Figure 5.14: The number of answers for the ease-of-use of the seven features, *Compare*, and *Find*

F1: "works very well", "useful, but not sure about the added value", "birth and death information is fundamentally important", "the comparison offers useful insights how a specific timespan is connected with earlier and later time periods"
F2: "could not get it to work", "not easy to understand why certain places were marked on the map", "important to examine space and time at the same time on map", "easy, but the usefulness is of course related to the research questions", "the functionality can't be used a lot of times for some reason", "the different times might need to be seperated more by titles, for example. Titles of occupation close to relevant date"
F3: "it is clearly described, but I still ended up seting it up in the wrong way, with the later time first, thus getting no results", "would be good with feedback that the reason for no result if for incorrectly definition of start/end date", "hard to use the Simple search function of the Find section when looking for people who lived before the birth of Christ", "always good to have some guidance about how to enter timespans", "cannot judge, because my research focuses on the time after the 4th century AD" "
F4: "intuitive and impressive", "able to preserve the relevant data", "not yet fully grasped the input method that would allow me to fully utilize the functions", "since the items are links to external pages, selecting it can be a bit tricky", "uncertainty about the best usage".
F5: "lack of data for the interest of the user", "limited use case for the research field", "very good function and easy to use, but potential to be extremely useful", "this funcion doesn't seem to work right now", "it didn't work, so I couldn't try", "we often compare objects from the same point of view", "easy to find information"
F6: "able to draw unexpected conclusions by find out what is just outside the scope of your data range", "very powerful function", "very usefull to see intersections between objects"
F7: "it's not research if there are no sources", "easy to understand", "it would be nice to be able to choose how many rows are displayed per page", "finding source information is fundamentally important", "easy to search for specific information, also I can order the results by any specific order".

Figure 5.15: Samples of the participants' comments for the seven functionalities (F1-F7)

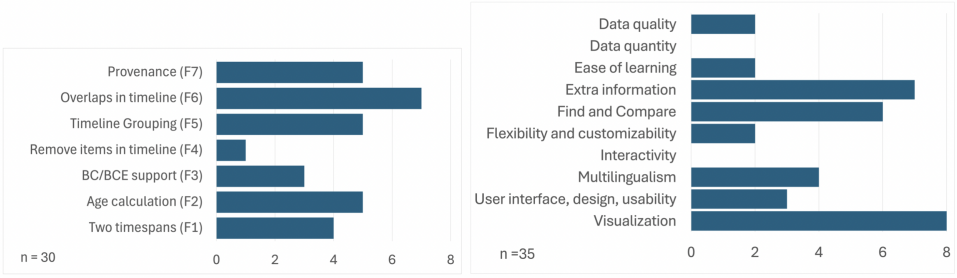

Figure 5.16: The number of answers for the uniqueness of seven features (left) and favorite aspects of the tool (right)(Note multiple choice is possible)

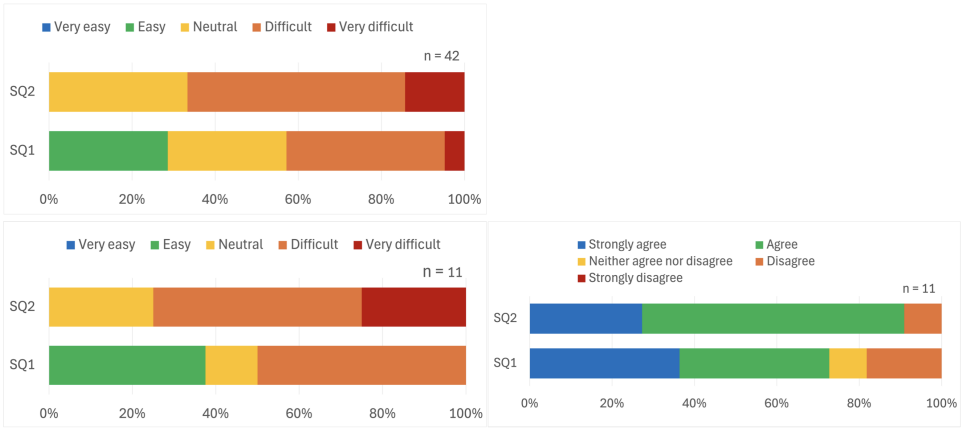

Figure 5.17: The number of answers for the difficulty to answer SQ1 ("How old was Van Gogh when moving to Paris and how old was Paul Gauguin at that time?") and SQ2 ("Which musicians worked in New York between 1960 and 1979 and how their time overlap each other?") on the web for non-participants (top left) and participants before the workshop (bottom left) and the agreement of assistance by the tool to answer them after the workshop (right)

functionality", "straight forward and very easy to use and interpret results" (see R8). Quite a few like the approach and see the potential once bugs are fixed: "a bit buggy at the current time, once those get solved it'll be a solid tool to use", "if the auto-complete works it is very useful and lately it worked great".

A few remarks concern the particularities of Wikidata (ontology, data quality). This may be related to knowledge acquisition issues, because our Wikidata-centric approach is contrary to many engineering projects seen in Section 5.2, which use Wikidata as secondary content. There are also general comments that the participants require more time to familiarize themselves with the tool (see R4 and R8): "it is difficult to use, but, once I get used to the search methods, I can access the information I look for". The lack of available data for a specific research field (religious studies) is also pointed out, leading to the question of immediate usefulness. Similarly, a respondent states: "I liked the approach overall. It is a great improvement on trying to type SPARQL queries! But afraid my feedback is rather limited by lack of return results" (see R8)

Some questions are also raised for time-related features: "time slider was a bit random in my limited experience" (see R9), "Would it make sense to separate past/present/future/searches?" (see R9 and R13)[21], "different interface approaches to different search requirements whether the user wants to search a time range or perhaps just return dates" (see R9 and R11), "problems with navigating the timeline table. More exact, with enhancing and scaling out the found timeline results" (see R11).

A use case to use the timeline view as a research presentation is suggested. Synergies with PeriodO [15] and CIDOC-CRM are proposed.

5.5 DISCUSSION

In general, our tool provided solutions for eight requirements out of fourteen (R3, R6, R7, R8, R9, R11, R12, R13) with different levels of technical maturity. Some functionalities are more favored by users than others. The time-related functionalities cover eight requirements out of fourteen (R3, R6, R7, R8, R9, R11, R12, R13), although there are a few functionalities that split opinions of the users.

Although the overall positive reactions in the user evaluation are encouraging, it is clear that the tool was unable to address all the requirements. For example, we focused on the search engine interface; therefore, we did not provide a new user experience such as user-oriented guided navigation and exploration (R14). It is a tremendous challenge for LD developers to provide meaningful visualization overviews and generous exploratory interfaces for such large and diverse LD sources as Wikidata, as opposed to relatively smaller-size and more focused CH collections, such as image collections in a museum. We hardly delivered any solution for R2. It is possible to provide an overview of the Wikidata statistics. However it will not let users to understand the scale and complexity of data easily. In addition, given the volume of Wikidata, data rendering will be a challenge to avoid performance issues. Similarly, R5 was not tackled. We did not concentrate on this requirement, because the tool primarily targeted online visualization. It could be updated

[21] The time slider has a range of 5000 years from BC/BCE 2500 to AD/CE 2500. It is implemented this way to support cases like dates found in science fiction

in the future to add mobile functionality, including GPS, Virtual Reality, and Augmented Reality.

Critical discourses in humanities [67] including uncertainty (R6) [106, 141] are only addressed to a small extent in this study (F2). Although Wikidata employs some techniques[22] to cope with them [67], we did not delve into them. In particular, we did not deal with various time concepts (R13) such as uncertain dates and non-Gregorian calendars.

Figure 5.14 indicates that some users struggled to search for the data in *Find*. This is understandable, because graph queries may be new to ordinary web users (see also [98] about the burden for LD interpretation without context). As Section 5.4.1 suggests, the level of SPARQL competence is relatively low even among those who have experience with SPARQL. However, this result is caused not only by the difficulties of graph queries, but also by the lack of data in Wikidata. It turned out that the availability of time in the Wikidata qualifiers is still sparse. For example, neither Gustav Klimt [8] nor Egon Schiele [5] have dates for work locations. Although Klimt has dates for his education, Schiele has no dates for it. This situation confuses the users as they do not know the comprehensiveness of the query results. In this sense, R2 is increasingly critical. In addition, this is related to the argument of Hyvönen et al. [94] about the imminent need for event gazetteers.

The participants struggled with query timeouts. Scalability and performance are often the bottleneck of LD applications; web communities need fundamental hardware and software updates in the future to solve them. For example, a warning of importing Wikidata dumps to a local installation is reported [28]. Wikidata plans to split scholarly articles from other data, because they have a serious impact on query performance [33].

For the same performance reasons, we deliberately disabled the inference function for the SPARQL queries in the Find section [90]. This is a well-known technical problem for LD implementers. Because of this, the degree of serendipity (R1) is limited. In addition, it is highly challenging to achieve R3. It is necessary to clarify it and provide solutions for a bespoke tool based on ReKisstory for each use case.

We did not address a solution for mass visualization (R11) [112]. For this study, we restricted ourselves to the visualization of a maximum of 100 results for *Find*. Although we have a simple solution for the table view (pagination) and map view (aggregation of nearby places), timeline will face a serious issue. We have no mechanism to display items by vertical immersion (zoom in) and abstraction (zoom out)[141]. Improved techniques for LD visualization need to be developed.

Desirably, we would like to achieve an interactive exploration of time and space in sync (changes over time and space are discussed in [106]) which one respondent requested in the questionnaire. InTaVia [24] partially presents a visualization of this type in the biographical domain.

Due to our focus on temporal data, we did not explore the feedback for all requirements. Future studies could drill down each requirement and conduct user evaluation. At the same time, our tool resembles yaap! in terms of this focus. This has pros and cons. While the uniqueness of our tool is reduced, our evaluation results indicated that our approach as well as yaap!'s are valid and our tool provides different functionalities that yaap! does not have, especially, the *Find* section, multiple views, and the combination of various functionalities.

[22] https://www.wikidata.org/wiki/Help:Dates#Qualifiers (Accessed on 2024-06-24)

It is not easy to identify and form a group of the best target users for evaluation. This is partly because of our strategies (R8) for middle-ground positioning. In a way, we inevitably both benefit and suffer from this broad scope of the tool. In this regard, it would be interesting to assess the tool in different ways by collecting opinions from LD experts and/or casual users and reorganizing our evaluation to understand generation gaps, technical competencies, and geographical differences in depth.

There could be different use cases of the tool. For instance, thanks to the availability of provenance information, the tool can be used for error detection and update/improvement of Wikidata. Integration with Wikidata editing tools would be useful. For the time being, our tool only provides a link to a statement on the Wikidata website, leaving the possibility of content update to the users within Wikidata.

5.6 CONCLUSION

Our literature review showed the limitation of the available LD tools and user evaluation. Although our list of previous studies is not exhaustive, Wikidata-centric visualization tools addressing the challenges of event data (as well as temporal data) in DH and CH are probably highly scarce. As such, studies that include user evaluation are hardly available.

From the user evaluation, we learn many insights into the user experience and needs for Wikidata tools. We found that there is still room for recognition to adapt and use Wikidata and SPARQL. This matches the evidence of increasing interest in, yet immaturity of Wikidata tools in DH [144]. However, the overall positive feedback for our tool confirms our concept and approach.

We discussed the (implicit) need and value of visualization tools that deal with events in LD for research purposes in DH and CH. In addition, we highlight how temporal data can be explored through events in Wikidata. Our study demonstrated that the tool uniquely addresses the user needs of LD visualization in the domain. In particular, the evaluation results indicate the improved access and visualization of LD through the seven time-related functionalities. Furthermore, it became evident that the Wikidata-centric approach to visualize events is valuable in the DH and CH domain, as opposed to the tools using Wikidata as the secondary content.

At the same time, the abstract nature of the requirements may have resulted in some mixed reactions in the user evaluation. Nevertheless, our research can be used as a starting point to define and refine requirements and user specifications for the development of this type of tool.

6

Conclusion

This chapter provides the general conclusion of the thesis. First, we revisit the four research questions (RQs) specified in the introduction. Second, we address the main research question of our study. Finally, we summarise the limitations of our research and future work.

6.1 Addressing Research Questions

RQ1: How is the quality of Linked Data instances in Cultural Heritage, particularly in terms of their connectivity?

Many previous studies focus on the quantitative analysis of the full spectrum of LD datasets. There have been numerous discussions about the quality of the owl:sameAs network because it is critical for LD data integration: to connect as many datasets as possible through this property. Although this approach is necessary, domain-specific and qualitative details were largely missing. Chapter 2 addresses this specific area of LD quality. To narrow our scope to the foundation of CH data, five categories are defined: agents, events, dates, places, and objects and concepts. We sampled the most representative 100 instances (20 instances per category) from eleven LD sources commonly used in CH. We analysed the interlinking of LD source datasets at the instance level. In this process, we scrutinised not only owl:sameAs but also three standardised properties: skos:exactMatch, rdfs:seeAlso, and schema:sameAs. This setup allowed us to gain a more comprehensive view of LD connectivity. As a result, we uncovered important linkage issues in the eleven LD data sources, which play a central role in LD applications in CH and the humanities.

A large proportion of LOD sources are unevenly interlinked, with limited reciprocal links. The interlinking network is condensed within a few data sources, particularly generic knowledge bases. This centrality is also observable in the data content. Generally, these findings support the outcomes of previous studies. Additionally, we identified that one of the major obstacles to high-level interoperability and automated data processing is the use of proprietary properties in data sources. We also report on issues with respect to data duplication in the context of data aggregation.

In conclusion, quality is insufficient for such tasks as data integration —including NEL — to process "linked" data efficiently in CH. When traversing LD within the eleven sources, some sources are not easily reachable. Currently, data consumers need a good understanding

of LD connectivity and careful strategies to traverse the LD network to find useful extra information. This is especially problematic in execution of (semi-)automated data integration and subsequent intelligent search for a substantial amount of data.

To mitigate this obstacle, our traversal maps help data consumers perform data integration and semantic enrichment more effectively. It is possible that limited discussion and coordination within the LD community, especially between data consumers and producers, have led to the current situation. As a first step to fill this gap, we recommend that LD data producers generate more reciprocal links to improve LD connectivity. By removing quality barriers, LOD traversing and data integration will become more feasible for data consumers with the help of automated tools.

LD quality can be increased, not only by improving existing LD sources, but also by creating new LD sources. When investigating LD connectivity for CH and DH in Chapter 2, the lack of numeric date entities became evident. As the user surveys in Chapter 5 clearly suggest, dates are critical information to conduct research in CH and DH. Therefore, it is highly valuable to examine the current state of these entities in more detail, taking the perspective of data consumers into account. Chapter 3 delves into this "weakest link" of LD quality, which answers RQ2.

RQ2: How can Linked Data connectivity for date entities be improved?

As a baseline, we confirmed through our brief observation of LD in CH data in Section 3.2 in Chapter 3 that a) descriptive time entities (or historical periods) such as the Neolithic and the Sui Dynasty are already available, but b) numeric dates are mostly limited to years for a certain time range, with little to no entities for specific months or days of a particular year. These phenomena are clearly identified in DBpedia and Wikidata, datasets which are often used in CH as a target reference in NEL tasks. This limitation is significant, especially for end users such as historians interested in the day-to-day reconstruction of events.

In contrast, we found that some ontologies are designed to model dates in a flexible manner, allowing data modellers and producers to represent numeric dates in RDF. Specifically, the Time Ontology in OWL and Wikidata provides sophisticated relationships between classes and properties of temporal information. Furthermore, there are some examples of temporal entities in LD implementations related to data enrichment and entity linking in CH. This reveals a gap between the potential to model temporal entities and what is actually published or exercised in current LD implementations.

To demonstrate the potential of numeric date entities, we designed and produced LODE, an experimental RDF dataset by reusing the data structure of the Time Ontology in OWL and the Wikidata ontology. The choice of these ontologies was mainly due to the potential for high interoperability, as well as to avoid inventing the wheel that data producers have already worked on. We explained the primary reasons for "nodifying" source data and proposed a typical workflow for generating LODE. As numeric dates in LD are typically encoded as literals, the "nodification" of literals is key to connecting temporal entities with other entities. This ensures an easy-to-enrich practical solution for future data producers. Nodification preserves original literals, resulting in no loss of information. Arithmetic calculations for date types are therefore still possible.

We developed a lookup application using the SKOSMOS software to provide access to

LODE at the most granular level — a single day at a specific point in time — for a span of 6,000 years. This configuration enables users to connect a significant number of LD sources through numeric date entities, covering the critical time span after prehistory. Additionally, our use cases demonstrated how LODE can serve as a catalyst for connecting heterogeneous data on the Web in the CH and DH domains. We argue that this approach has the potential to open new avenues of research in CH and DH. Although our focus was on improving the connectivity of numeric date entities in LD, we argue that this type of contribution will support the overall improvement of LD connectivity in the long run.

Chapter 3 demonstrated how to improve LD quality by creating a new LD source. Chapter 4 explored a case of generating LD by means of Information Extraction, a Natural Language Processing technique. We evaluated the process of extracting information from Wikipedia to create more detailed LD than existing LD sources such as DBpedia and Wikidata.

RQ3: What are the quality gaps in biographical information between Wikipedia and Linked Data, and how can Information Extraction on Wikipedia be used to address them?

Due to the high volume and coverage of DBpedia and Wikidata, data consumers may expect a reasonable level of data quality from these sources. However, this may not always be the case. Biography is one of the emerging areas of study in CH and DH [24, 94]. We performed a case study, analyzing biographical information in a well-known LD source to get insights into LD quality.

In Chapter 4, we used Henry VIII as a case study to compare the quality of two LD datasets (Wikidata and DBpedia) with Wikipedia (a resource closely related to the two LDs). His biography is one of the most referenced articles on Wikipedia. We found important data missing in both sources.

Given that Henry VIII is one of the most referenced biographical articles on Wikipedia, we assumed that similar quality gaps could be found in many lesser-referenced Wikipedia articles and corresponding LD entities. We therefore explored methodologies for creating LD using IE from the Wikipedia article on Henry VIII. If we can extract valuable information from Wikipedia and generate LD of reasonable quality and quantity, missing data in DBpedia and Wikidata could be supplemented, reducing information gaps between Wikipedia and other LD sources.

We conducted triple extraction and RDF generation with NEL, deploying state-of-the-art NLP models to examine the quality of the generated LD. Several techniques, including SRL (Semantic Role Labeling) and SFD (Semantic Frame Detector), were employed. We evaluated the performance of pre-trained models and compared the generated RDF with the RDF currently available in DBpedia and Wikidata. Additionally, we analysed the compatibility of the generated data with biographical ontologies.

Although no normalisation was applied to the extracted properties and some data mapping noise remains, a significant amount of new RDF statements at a detailed level was successfully extracted. By establishing a comprehensive pipeline from text data ingestion to LD publication, semi-automatic NLP could create sizeable new LD datasets from Wikipedia articles. These datasets could provide more detailed semantics for LD entities and supplement Wikidata and DBpedia. Our method showed potential for addressing data quality

gaps. As we also compared the properties of the generated triples with existing biographical ontologies, we realised that they could be used to develop or improve biographical ontologies. This implies that LD could also be enriched through the improvement of ontologies, although we did not intend to scrutinise ontological aspects of LD quality in the beginning. In a way, we developed a bottom-up approach for ontology design and enrichment. Our study also suggested that close collaboration between developers and data producers is necessary to identify data quality issues, establish methodologies, and maintain generated data.

Our research has demonstrated that all stakeholders should be involved in the LD ecosystem. However, developers have not been the primary focus. Chapter 4 suggested that we may need to place greater emphasis on communication between developers and data consumers. As a result, all stakeholders make different, yet valuable contributions. In the next question, we explore the cross-dimension of developers and data consumers in terms of tooling quality.

RQ4: What are the effective designs and functionalities for Linked Data tools to support research using temporal information in Cultural Heritage (CH) and Digital Humanities (DH)?

It is a challenge to define effective designs and functionalities, given the diversity of data consumers in CH and DH.Chapter 5 attempted to partially address that challenge.

As found in Section 5.1, there is a growing need for visualisation tools to explore events in LD. We explore the potential of LD visualisation tools that address event data in association with temporal data for research purposes in DH and CH. Wikidata was chosen as a case study due to its importance in the current practice of LD in DH and CH.

However, our research in Section 5.2 revealed the limited availability, as well as the limited user evaluation, of LD and Wikidata tools that handle events and time. Therefore, it was anticipated that it would not be easy to collect the user needs without presenting a concrete application to the potential users. Consequently, we made a strategic decision to develop a Wikidata tool first, based on the findings of the previous studies. We then conducted a user evaluation for the tool to gain "hidden needs", which could eventually identify effective design and functionalities for a LD visualisation tool.

We introduced 14 potential user requirements based on principles from the information visualisation domain, as well as an analysis of previous studies and existing tools. A newly developed tool, called ReKisstory, designed to meet these requirements was evaluated through five focus groups and questionnaires with DH and CH experts.

Despite the potentially broad needs of experts in diverse fields, positive responses from the participants provided evidence that our tool met many of the requirements for research involving temporal information in the humanities. We implemented nine requirements out of fourteen with different levels of technical maturity. In addition, the time-related functionalities deal with eight requirements out of fourteen. Although there are a few functionalities that users perceived differently, our survey revealed that the seven time-related functionalities of our tool offered valuable user experience for the exploration of time in research within CH and DH. To our knowledge, no other tool covers all seven functionalities.

Our pre-workshop survey indicated that some experts find it hard to compare time-related information on the web and write SPARQL queries. After the workshop, the participants

rated both the usefulness and ease of use of the seven functionalities high. Their interest in the availability of extra information in the tool was related to the linking capability of LD (see also RQ1). In addition, the user evaluation showed that ReKisstory is helpful with respect to two research tasks that are difficult to complete without it. Moreover, it became clear that the Wikidata-centric approach to visualise events is valuable in the DH and CH domain.

We obtained a considerable amount of details regarding the users' opinions about their Wikidata experience and the Wikidata tool handling events and temporal data. We proved the implicit needs of LD visualisation tools dealing with events in DH and CH. Our research can serve as a starting point for identifying effective design and functionalities for the development of this type of tools in the future.

How can the quality of data and tools for Linked Data in Cultural Heritage and Digital Humanities be enhanced?

During our research, it became clear that the quality issues in LD are complex and multifaceted. Admittedly, there is no single solution to address them all. Nevertheless, to answer this main question, we provide overall conclusions for: (a) our research methods, and (b) the analyses of LD quality. These lead to eight strategies (four for improving data and four for improving the tools).

(a) Research methodologies

To enhance the quality of data and tools for LD in CH and DH, it is essential to analyse the current quality. The question is how to analyse it. From a methodological point of view, we conclude that our strategy to emphasise more on the qualitative analysis (Chapter 2, Chapter 3, Chapter 4, and Chapter 5) than quantitative analysis provided more user-centric understanding of LD quality, compared to previous studies. The results of our analyses shed light on ways in which to fill LD quality gaps when producing and consuming LD data using available tools. This strategy helped to formulate solutions to improve the quality of LD for as many stakeholders as possible. In order to properly measure LD quality, future research in the LD community could increase attention to qualitative analysis.

(b) Analysis and Strategies

We look at enhancing LD quality in two dimensions: data and tool quality. We showed two primary ways to improve quality: one is to enrich existing resourcesand the other is to create new high-quality resources. Therefore, the quality improvement of both scenarios can be seen in our conclusions below.

Specifically, we suggest four strategies for improving data quality (related chapters are indicated). In ideal stakeholder communication scenarios, all stakeholders should play a role in all strategies. However, we specify the most relevant targeted stakeholders for each strategy:

1. Provide more links, especially reciprocal links [Data producers, Developers] (Chapter 2)

2. Perform coordinated analyses on the LD quality and productions of data [Data producers, Data consumers] (Chapter 2, Chapter 3)

3. Generate new LD as well as enriching it to address gaps in existing LD sources [Data producers] (Chapter 3)

4. Design, develop, and enrich ontologies, taking more insight from bottom-up approaches such as NLP on text corpora [Data producers, Data consumers, Developers] (Chapter 4)

The first three points are derived from Chapter 2, Chapter 3, and Chapter 4. The importance of interlinking is reaffirmed in Chapter 2. However, it is challenging for LD producers to maintain an overview of incoming and outgoing links in their datasets. Without effective communication and collaboration among LD producers, creating and maintaining reciprocal links remains a difficult task. In this context, greater coordination between LD producers is essential when analysing the data quality and producing the data. Chapter 3 and Chapter 4 address issues related to missing data resources. New LD datasets such as LODE for numeric date entities can be created in areas where existing LD sources fall short. Numeric date entities and more detailed semantics for the biographical domain are two examples.

The fourth point stems from our work in Chapter 4. Techniques such as NLP can identify missing vocabularies in existing ontologies, which are often constructed top-down based on expert knowledge. By analysing textual descriptions of entities (in our case, biographies), Chapter 4 demonstrated that refined vocabularies can enhance domain ontologies.

The four strategies to improve tool quality are as follows (mostly derived from Chapter 5). The relevant targeted stakeholders are also indicated:

1. Accelerate the development of LD tools in CH and DH [Developers]
2. Incorporate user evaluations during LD tool development and publicly publish the results [Data consumers, Developers]
3. Develop features related to key concepts in CH and DH, such as temporal data and provenance [Data consumers, Developers]
4. Aim for second- or third-generation systems as suggested by Hyvönen [93] [Developers]

As Chapter 5 indicates, user evaluations for LD tools are still largely absent in CH and DH. One reason for this is the lack of focus on tooling in research. Therefore, we suggest intensifying the development of LD tools for these domains. Although the volume of tools does not automatically guarantee the quality, Chapter 5 suggests that "hidden needs" tend to remain unnoticed until the tools are broadly disseminated. The visibility of LD tools would increase the awareness and experience of data consumers, which gradually leads to the improvement of tool quality in general, when more user evaluations are executed (point two).

User evaluation should be systematically planned when developing new tools, particularly in an academic research setting. It is highly recommended for developers to publish the results of such user evaluation. The subsequent development of tools can then benefit from the results. Furthermore, tool development should prioritise features relevant to CH and DH research. For instance, although ReKisstory received positive feedback, some users expressed uncertainty about its usefulness for their specific research needs. This highlights the importance of conducting user evaluations that focus on functionalities of interest to researchers.

We also observed a shift toward second- and third-generation LD systems due to advancements in LD and the evolving needs of the CH and DH domains. New tools should account for these trends to meet the expectations of data consumers. Chapter 5 showed the "hidden" needs or requirements for visualization tools, therefore, we foresee room for discovering new types of tools for different purposes.

6.2 DISCUSSIONS AND FUTURE WORK

6.2.1 LIMITATIONS

As our investigation focused on the Cultural Heritage (CH) and Digital Humanities (DH) domains, we did not provide insights into the overall quality of LD sources and tools on the Web. As interdisciplinary research is becoming more common, interdisciplinary studies on LD quality are also becoming increasingly important.

Two approaches are pragmatic solutions for data integration in CH and DH. One is the interlinking of LD in CH and DH through generic LD sources. This is why NEL is frequently performed in these domains. We investigated this approach in this thesis.

The other approach, which we did not explore in the thesis, is LD data modeling, using other ontologies than the Wikidata ontologies. CIDOC-CRM is one of the most well-known standards for heterogeneous data integration. While facilitating the construction of an overarching and expressive knowledge base for CH documentation, the CIDOC-CRM community also publishes compatible models for specialised subdomains [2]. Eleven compatible models are listed on the website. For instance, while LRMoo is a library reference model, CRMsci and CRMtex are designed for scientific observation and ancient texts, respectively. In terms of practical data integration using CIDOC-CRM, the ARIADNE Ontology is worth mentioning [77, 114]. Built upon this CIDOC-CRM-based ontology, the ARIADNE Portal [26] is a research infrastructure which aggregates heterogenous archaeological data from different sources across Europe. These ontologies and their applications should be investigated in future research.

Currently, examples of combining the two approaches are limited to relatively small-scale local datasets. We assume that this is due to the time-consuming tasks of NEL and data mapping and encoding. In our future work, the challenge of applying these approaches to large data sets should be considered to achieve a high level of quality in terms of data and tools. Still, such heterogeneous interdisciplinary data integration scenarios could be complex, making comparative analysis a challenging task.

With regard to the relationship between data and tooling, we focused on the "second generation" LD tools, namely tools specializing in DH analyses (including visualization tools for the analyses) in relation to the existing data (Chapter 5). However, the gap between data production and data consumption we identified in Chapter 2 is more related to the tools for data production, authoring, and publication, which may well be the first generation systems. Although they are equally important, we have not evaluated them in this thesis.

Quality assessment is not a trivial task. Due to various constraints, we did not explore different methodologies and techniques for evaluation in depth. In particular, there is room for further user evaluation. More case studies will allow us to draw stronger conclusions about LD quality. Formal methodologies in design science need to be applied. In terms of LD datasets, Chapter 2, Chapter 3 and Chapter 4 could be extended to conduct user surveys

on data publication in relation to use cases in specialised research projects. In terms of tooling, we consider Chapter 5 as a baseline study for specific functionalities. As LD can be used in many ways, further evaluations are needed to discover its potential uses. For instance, it would be worthwhile to evaluate whether a generic tool like ReKisstory could be valuable across different CH and DH research themes and projects. In both cases, embedding a user evaluation within a larger research project, where target users are carefully defined and feedback is systematically gathered from a significant number of users, would be a valuable contribution.

6.2.2 DISCUSSIONS

This thesis undertook the challenge of investigating and addressing the quality of LD. Although we managed to answer most of the research questions, there are still some areas that require further discussion.

In terms of methodology, much of our research placed particular emphasis on the qualitative analysis of LD quality. The main reason is that the analysis of two specific domains (CH and DH) requires examining use cases and individual instances. We conducted our research with the awareness that qualitative details are often missing in previous studies. In Chapter 2, we highlighted this contrast using the analogy of the wood and the forest. Similarly, we referred to the difference between close reading and distant reading in the DH field in Chapter 4.

This situation itself reflects the gaps between LD data producers, data consumers, and developers. The quality analysis of LD, primarily produced by data producers and developers, is often conducted using quantitative methods. This is natural since available LD datasets are often too large to be manually assessed in their entirety. While these methods provide an essential overview of LD sources as a whole, they may not meet the needs of data consumers who require a more granular view of individual data for their specific purposes. If our study has helped shift the attention of data producers and developers to this issue, then the gaps may already be closing. These gaps are not necessarily due to misunderstanding or technical problems, but rather to a lack of awareness or a slightly different focus. Therefore, closing the gaps may not be difficult if more contributions like ours are made. In this regard, sharing research outcomes with a wider audience is crucial.

Likewise, humanities scholars may confront a dilemma of positioning themselves somewhere between the "traditional" close reading (qualitative) and "new" distant reading in DH (quantitative) [52]. Indeed, DH scholars are well aware that close reading should not be neglected, even if distant reading provides brand new insights into the humanities research. In this sense, the gap between close and distant reading needs to be filled by the research community in such a way that research facilitates multiple perspectives and interpretations, which are the core values of the humanities.

In terms of CH and DH, we used two popular LD datasets — Wikidata and DBpedia — in all chapters. We demonstrated that they play an important role in LD applications for CH and DH. However, we should also remember that these datasets contain cross-domain encyclopedic data. Thus, they are not specifically designed for CH and DH, and their content coverage is quite diverse. At the same time, Chapter 2 also points out that some other well-known LD datasets in CH and DH tend to serve merely as global identifiers rather than new sources of information. In this sense, there is no single "LD" for CH and DH. The

complexity of the current LD landscape implies challenges in defining the needs of data producers, consumers, and developers, as well as in closing the gaps between them.

At the same time, this issue can also be viewed from another angle. In distributed systems like the Web, it is normal to expect data integration from heterogeneous sources, as one source is often insufficient. There is then a chance of serendipity in integrating data from various sources through interlinks. As Chapter 2 suggests, generic LD sources such as Wikidata and DBpedia can be connected to many local LD sources that were not discussed in this thesis. Furthermore, using generic LD hubs, local-to-local (L2L) data integration becomes feasible. Local sources often hold previously unknown data, enabling interesting data integration opportunities. This is especially the case in the CH sector. Being a part of the Open Data landscape, GLAM institutions, especially public organisations, have a keen interest in leveraging previously "hidden" collections for the public. The synergies from such L2L connections could lead to impactful outcomes in CH and DH. Although examples of highly international, interdisciplinary, multilingual, and heterogeneous data integration have been limited, there are initiatives such as ARIADNE to facilitate L2L data integration through CIDOC-CRM-based ontology. Going beyond the domain of archaeology, generic LD sources could serve as a pivotal connection resource in the broader context of CH and DH. Improving LD and its tools in such scenarios should be considered for future work.

These challenges can also be rephrased in the context of LD development phases, roughly aligning with the six stages of LOD development [72] and generations of LD systems mentioned by Hyvönen [93] as discussed in Chapter 1. The first phase of LD is the generation and publication of LD for more generic purposes. Generic LD sources like Wikidata and DBpedia serve as foundational resources for broader communities. The second phase is connecting generic LD to more specialised local LD that has emerged in parallel. NEL has been used to achieve this (Chapter 2), allowing for data integration. The third phase would involve increasing L2L connections, where more data reuse and knowledge discoveries may occur. This phase may be happening now as the need for LD grows. In such knowledge discovery scenarios, user interfaces (especially visualization and analytical tools we discussed in Chapter 5) play a vital role to deal with large datasets, potentially reshaping our knowledge. Since the L2L interlinking may involve data integration across domains, more interdisciplinary studies could become possible. As L2L data integration becomes more common, new potentials for reshaping CH (e.g. expanding the definition of CH) would emerge. At the same time, new challenges in assessing LD quality may arise. The quality of highly interdisciplinary data integration is still largely unknown and requires further investigation.

There can be more discussions about the distributed nature of LD. Chapter 2 found many overlaps in data across LD sources. On the one hand, these overlaps reduce redundant traversing; on the other hand, they increase duplicate information. This situation is common, as data aggregation is a typical activity in web-based data integration. The examples of VIAF, YAGO, BabelNet, and Europeana illustrate this phenomenon. This discussion also relates to the implementation of provenance information, as discussed in Chapter 4 and Chapter 5. Moreover, the aforementioned data integration scenarios for international, interdisciplinary, multilingual, and heterogeneous data aggregation would complicate discussions of LD quality. By evaluating the pros and cons of LD's distributed approach, we can find a way to effectively manage and utilize the LD ecosystem as a whole.

*

At the beginning of this thesis, we introduced the concept of the "Missing Links". Our investigation uncovered several critical issues of the LD quality, highlighting our current position in the evolution of the Web. This research addressed these challenges from multiple perspectives, using various techniques. We also outlined potential future developments. While Berners-Lee's second dream is still unfolding, we believe this thesis contributes to advancing it.

APPENDIX A: ENTITY COVERAGE PER DATA SOURCE

This appendix contains a table created for Chapter 2, showing an overview of the 100 entitiees in 11 data sources.

Table A.1: The occurrences of 100 entities in 11 data sources (A to K) (Zero indicates absence. More than one means duplicate entities)

		A YAGO	B Worldcat	C Wikidata	D VIAF	E LoC	F Getty	G GeoNames	H Europeana	I DBpedia	J BabelNet	K Wikipedia	SUM	Occurence SUM
1	Carl Linnaeus	1	1	1	1	1	1	0	0	1	1	1	9	9
2	Jesus	1	1	1	1	1	0	0	0	1	1	1	8	8
3	Aristotle	1	1	1	2	1	1	0	1	1	1	1	11	10
4	Napoleon	1	1	1	1	1	1	0	0	1	1	1	9	9
5	Adolf Hitler	1	1	1	1	1	1	0	1	1	1	1	10	10
6	Julius Caesar	1	1	1	2	1	1	0	0	1	1	1	10	9
7	Plato	1	1	1	1	1	1	0	1	1	1	1	10	10
8	William Shakespeare	1	1	1	1	1	1	0	1	1	1	1	10	10
9	Albert Einstein	1	1	1	1	1	1	0	0	1	1	1	9	9
10	Elizabeth II	1	1	1	1	1	1	0	0	1	1	1	9	9
11	Michael Jackson	1	1	1	1	1	1	0	0	1	1	1	9	9
12	Madonna (entertainer)	1	1	1	1	1	0	0	1	1	1	1	9	9
13	Ludwig van Beethoven	1	1	1	1	1	1	0	1	1	1	1	10	10
14	Wolfgang Amadeus Mozart	1	1	1	1	1	1	0	1	1	1	1	10	10
15	Pope Benedict XVI	1	1	1	1	1	0	0	0	1	1	1	8	8
16	Alexander the Great	1	1	1	2	1	1	0	0	1	1	1	10	9
17	Charles Darwin	1	1	1	1	1	1	0	0	1	1	1	9	9
18	Barack Obama	1	1	1	1	1	1	0	0	1	1	1	9	9
19	Mary (mother of Jesus)	1	1	1	1	1	0	0	0	1	1	1	8	8
20	Queen Victoria	1	1	1	1	1	1	0	0	1	1	1	9	9
1	World War II	1	1	1	0	1	0	0	0	1	1	1	7	7
2	World War I	1	1	1	0	1	0	0	1	1	1	1	8	8
3	American Civil War	1	1	1	0	1	0	0	0	1	1	1	7	7
4	FA Cup	1	1	1	0	1	0	0	0	1	1	1	7	7
5	Vietnam War	1	1	1	0	1	0	0	0	1	1	1	7	7
6	Academy Awards	1	1	1	0	1	0	0	0	1	1	1	7	7
7	Cold War	1	1	1	0	1	0	0	0	1	1	1	7	7
8	Korean War	1	1	1	0	1	0	0	0	1	1	1	7	7
9	American Revolutionary War	1	1	1	0	1	0	0	0	1	1	1	7	7
10	UEFA Champions League	1	1	1	0	1	0	0	0	1	1	1	7	7
11	UEFA Europa League	1	0	1	0	0	0	0	0	1	1	1	5	5
12	Olympic Games	2	1	1	0	1	0	0	0	1	1	1	8	7
13	Stanley Cup	1	1	1	0	1	0	0	0	1	1	1	7	7
14	Super Bowl	1	1	1	0	1	0	0	0	1	1	1	7	7
15	Iraq War	1	1	1	0	1	0	0	0	1	1	1	7	7
16	War of 1812	1	1	1	0	1	0	0	0	1	1	1	7	7
17	Gulf War	1	1	1	0	1	0	0	0	1	1	1	7	7
18	Spanish Civil War	1	1	1	0	1	0	0	0	1	0	1	6	6
19	World Series	1	1	1	0	1	0	0	0	1	1	1	7	7
20	EFL Cup	1	1	1	0	1	0	0	0	1	1	1	7	7
1	1987	1	1	1	0	1	0	0	0	1	1	1	7	7
2	1986	1	1	1	0	1	0	0	0	1	1	1	7	7
3	1985	1	1	1	0	1	0	0	0	1	1	1	7	7
4	1984	1	1	1	0	1	0	0	0	1	1	1	7	7
5	1983	1	1	1	0	1	0	0	0	1	1	1	7	7
6	1982	1	1	1	0	1	0	0	0	1	1	1	7	7
7	1981	1	1	1	0	1	0	0	0	1	1	1	7	7
8	1980	1	1	1	0	1	0	0	0	1	1	1	7	7
9	1979	1	0	1	0	0	0	0	0	1	1	1	5	5
10	1978	0	1	1	0	1	0	0	0	1	1	1	6	6
11	1977	1	1	1	0	1	0	0	0	1	1	1	7	7
12	1976	1	0	1	0	0	0	0	0	1	1	1	5	5
13	1975	1	1	1	0	1	0	0	0	1	1	1	7	7
14	1969	1	1	1	0	1	0	0	0	1	1	1	7	7
15	1968	1	1	1	0	1	0	0	0	1	1	1	7	7
16	1967	1	1	1	0	1	0	0	0	1	1	1	7	7
17	1966	1	1	1	0	1	0	0	0	1	1	1	7	7
18	1965	1	1	1	0	1	0	0	0	1	1	1	7	7
19	1964	1	1	1	0	1	0	0	0	1	1	1	7	7
20	1960	1	1	1	0	1	0	0	0	1	1	1	7	7
1	United States	1	1	1	1	1	1	1	1	1	1	1	11	11
2	United Kingdom	1	1	1	1	1	1	1	1	1	1	1	11	11
3	France	1	1	1	1	1	1	1	1	1	1	1	11	11
4	England	1	1	1	1	1	1	1	1	1	1	1	11	11
5	Germany	1	1	1	1	1	1	1	1	1	1	1	11	11
6	Canada	1	1	1	1	1	1	1	1	1	1	1	11	11
7	Australia	1	1	1	2	1	1	1	1	1	1	1	12	11
8	Japan	1	1	1	1	1	1	1	1	1	1	1	11	11
9	Italy	1	1	1	1	1	1	1	1	1	1	1	11	11
10	Poland	1	1	1	1	1	1	1	1	1	1	1	11	11
11	India	2	1	1	1	1	1	1	1	1	1	1	12	11
12	Spain	2	1	1	1	1	1	1	1	1	1	1	12	11
13	London	2	1	1	1	1	1	1	1	1	1	1	12	11
14	Russia	2	1	1	1	1	1	1	1	1	1	1	12	11
15	New York City	1	0	1	1	1	1	1	0	1	1	1	9	9
16	Brazil	2	1	1	1	1	1	1	0	1	1	1	11	10
17	California	2	1	1	1	1	1	1	0	1	1	1	11	10
18	New York	2	1	1	1	1	1	1	1	1	1	1	12	11
19	The Netherlands	2	1	1	1	1	1	1	1	1	1	1	12	11
20	Sweden	2	1	1	1	1	1	1	1	1	1	1	12	11
1	Book of Kells	1	1	1	1	1	0	0	0	1	1	1	8	8
2	Vasa	1	1	1	1	1	0	0	0	1	1	1	8	8
3	The Garden of Earthly Delights	1	1	1	1	1	0	0	0	1	1	1	8	8
4	Rosetta Stone	1	1	1	1	1	0	0	0	1	1	1	8	8
5	Palazzo Pitti	1	1	1	2	1	1	1	0	1	1	1	11	10
6	Boeing 747	1	1	1	0	1	0	0	0	1	1	1	7	7
7	Sgt. Pepper's	1	1	1	2	1	0	0	0	1	1	1	9	8
8	Tosca	1	1	1	1	1	0	0	0	1	1	1	8	8
9	Blade Runner	1	1	1	1	1	0	0	0	1	1	1	8	8
10	Uncle Tom's Cabin	1	1	1	1	1	0	0	0	1	1	1	8	8
11	Ming Dynasty	1	0	1	1	1	1	0	0	1	1	1	8	8
12	Ukiyo-e	1	1	1	0	1	1	0	0	1	1	1	8	8
13	Angkor Wat	1	1	1	1	1	1	1	0	1	1	1	10	10
14	Toraja	1	1	1	0	1	0	0	0	1	1	1	7	7
15	Byzantine Empire	1	1	1	2	1	1	0	0	1	1	1	10	9
16	Mars	1	1	1	1	1	1	0	0	1	1	1	9	9
17	Tamil language	2	1	1	0	3	1	0	0	1	1	1	11	8
18	Influenza	1	1	1	0	1	1	0	0	1	1	1	8	8
19	The King and I	1	1	1	3	2	0	0	0	1	1	1	11	8
20	Like a Rolling Stone	1	1	1	1	1	0	0	0	1	1	1	8	8
	SUM	110	95	100	64	100	44	22	25	100	99	100	859	836
	Occurence SUM	99	95	100	55	97	44	22	25	100	99	100	836	

Appendix B: Source Matrix Data

This appendix contains tables created for Chapter 2. The tables are raw data matrices used to generate traversal maps in each category (agents, events, dates, places, objects and concepts).

Table B.1: Matrix data with inverse (Figure 1 left)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	98	0	20	0	0	0	0	0	0	0	0	118
LoC	93	43	52	13	0	0	0	0	0	0	0	201
VIAF	58	59	0	0	0	16	0	0	0	16	0	149
Getty	0	0	18	56	0	0	0	0	0	0	0	74
Wikidata	1	0	43	0	192	100	0	0	0	98	8	442
DBpedia	0	0	38	0	0	5599	108	23	0	1397	870	8035
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	22	20	0	0	20	17	87
Wikipedia	1	0	0	0	0	0	0	0	0	1094	0	1095
YAGO	0	0	0	0	0	95	82	0	0	88	8	273
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	259	102	171	69	192	5832	210	23	0	2713	903	10474

Table B.2: Matrix data without inverse (Figure 1 right)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	20	0	0	0	0	0	0	0	0	21
LoC	93	43	52	13	0	0	0	0	0	0	0	201
VIAF	58	59	0	0	0	16	0	0	0	16	0	149
Getty	0	0	18	56	0	0	0	0	0	0	0	74
Wikidata	1	0	43	0	0	100	0	0	0	98	8	250
DBpedia	0	0	38	0	0	1581	108	23	0	1397	870	4017
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	22	20	0	0	20	17	87
Wikipedia	1	0	0	0	0	0	0	0	0	1094	0	1095
YAGO	0	0	0	0	0	94	82	0	0	88	8	272
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	162	102	171	69	0	1813	210	23	0	2713	903	6166

Table B.3: skos:exactMatch traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	3	0	13	0	0	0	0	0	0	0	16
VIAF	0	59	0	0	0	0	0	0	0	0	0	59
Getty	0	0	0	12	0	0	0	0	0	0	0	12
Wikidata	0	0	0	0	0	0	0	0	0	0	1	1
DBpedia	0	0	0	0	0	0	108	0	0	0	109	217
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	20	0	0	0	0	20
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	82	0	0	0	1	83
Europeana	0	0	0	0	0	0	0	0	0	0	1	1
SUM	0	62	0	25	0	0	210	0	0	0	112	409

Table B.4: rdfs:seeAlso traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	97	0	0	0	0	0	0	0	0	0	0	97
LoC	0	1	0	0	0	0	0	0	0	0	0	1
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	44	0	0	0	0	0	0	0	44
Wikidata	1	0	0	0	0	0	0	0	0	0	0	1
DBpedia	0	0	0	0	0	4100	0	23	0	261	0	4384
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	1	0	0	0	0	0	0	0	0	10	0	11
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	99	1	0	44	0	4100	0	23	0	271	0	4538

Table B.5: owl:sameAs traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	39	0	0	0	0	0	0	0	0	0	39
VIAF	0	0	0	0	0	16	0	0	0	16	0	32
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	192	100	0	0	0	98	7	397
DBpedia	0	0	0	0	0	1469	0	0	0	1132	761	3362
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	22	0	0	0	20	17	59
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	95	0	0	0	87	7	189
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	39	0	0	192	1702	0	0	0	1353	792	4078

Table B.6: schema:sameAs traversal map

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	0	20	0	0	0	0	0	0	0	0	21
LoC	93	0	52	0	0	0	0	0	0	0	0	145
VIAF	58	0	0	0	0	0	0	0	0	0	0	58
Getty	0	0	18	0	0	0	0	0	0	0	0	18
Wikidata	0	0	43	0	0	0	0	0	0	0	0	43
DBpedia	0	0	38	0	0	0	0	0	0	0	0	38
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	0	0	0	0	0	0	8
Wikipedia	0	0	0	0	0	0	0	0	0	1084	0	1084
YAGO	0	0	0	0	0	0	0	0	0	1	0	1
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	160	0	171	0	0	0	0	0	0	1085	0	1416

Table B.7: Matrix data which generated the traversal map for agents (Figure 4)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	20	0	0	0	0	0	0	0	0	0	0	20
LoC	20	0	20	12	0	0	0	0	0	0	0	52
VIAF	20	20	0	0	0	16	0	0	0	16	0	72
Getty	0	0	15	28	0	0	0	0	0	0	0	43
Wikidata	0	0	20	0	55	20	0	0	0	19	7	121
DBpedia	0	0	19	0	0	408	24	0	0	343	761	1555
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	26	23	0	0	26	7	82
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	60	20	74	40	55	470	47	0	0	404	775	1945

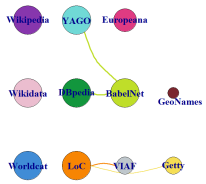

Table B.8: skos:exactMatch traversal map for agents

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	12	0	0	0	0	0	0	0	12
VIAF	0	20	0	0	0	0	0	0	0	0	0	20
Getty	0	0	0	12	0	0	0	0	0	0	0	12
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	24	0	0	0	0	24
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	23	0	0	0	0	23
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	20	0	24	0	0	47	0	0	0	0	91

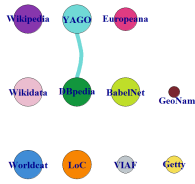

Table B.9: rdfs:seeAlso traversal map for agents

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	20	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**20**
LoC	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	16	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**16**
Wikidata	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
DBpedia	**0**	**0**	**0**	**0**	**0**	104	**0**	**0**	**0**	77	**0**	**181**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**20**	**0**	**0**	**16**	**0**	**104**	**0**	**0**	**0**	**77**	**0**	**217**

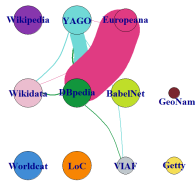

Table B.10: owl:sameAs traversal map for agents

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LoC	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
VIAF	**0**	**0**	0	**0**	**0**	16	**0**	**0**	**0**	16	**0**	**32**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	55	**20**	**0**	**0**	**0**	**19**	**7**	**101**
DBpedia	**0**	**0**	**0**	**0**	**0**	360	**0**	**0**	**0**	**266**	**761**	**1327**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**26**	**0**	**0**	**0**	25	**7**	**59**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**0**	**0**	**0**	**0**	**55**	**362**	**0**	**0**	**0**	**327**	**775**	**1519**

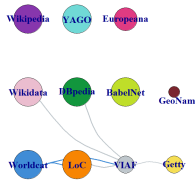

Table B.11: schema:sameAs traversal map for agents

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LoC	**20**	0	**20**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**40**
VIAF	**20**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**20**
Getty	**0**	**0**	**15**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**15**
Wikidata	**0**	**0**	**20**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**20**
DBpedia	**0**	**0**	**19**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**19**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**40**	**0**	**74**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**114**

Table B.12: Matrix data which generated the traversal map for events (Figure 6)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**19**
LoC	**18**	12	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**30**
VIAF	**7**	**7**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**14**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	36	**20**	**0**	**0**	**0**	**20**	**1**	**77**
DBpedia	**0**	**0**	**0**	**0**	**0**	563	**23**	**0**	**0**	**361**	**109**	**1056**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**1**	**0**	**1**
YAGO	**0**	**0**	**0**	**0**	**0**	**20**	**21**	**0**	**0**	21	**1**	**63**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**44**	**19**	**0**	**0**	**36**	**603**	**44**	**0**	**0**	**403**	**111**	**1260**

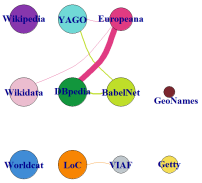

Table B.13: skos:exactMatch traversal map for events

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LoC	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
VIAF	**0**	**7**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**7**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**1**	**1**
DBpedia	**0**	**0**	**0**	**0**	**0**	0	**23**	**0**	**0**	**0**	**109**	**132**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**21**	**0**	**0**	0	**1**	**22**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1	**1**
SUM	**0**	**7**	**0**	**0**	**0**	**0**	**44**	**0**	**0**	**0**	**112**	**163**

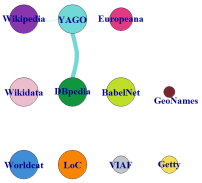

Table B.14: rdfs:seeAlso traversal map for events

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**19**
LoC	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
DBpedia	**0**	**0**	**0**	**0**	**0**	263	**0**	**0**	**0**	**82**	**0**	**345**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**1**	**0**	**1**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**19**	**0**	**0**	**0**	**0**	**263**	**0**	**0**	**0**	**83**	**0**	**365**

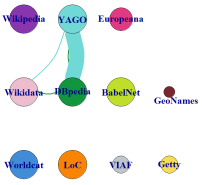

Table B.15: owl:sameAs traversal map for events

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LoC	**0**	12	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**12**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	36	**20**	**0**	**0**	**0**	**20**	**0**	**76**
DBpedia	**0**	**0**	**0**	**0**	**0**	298	**0**	**0**	**0**	**279**	**0**	**577**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**20**	**0**	**0**	**0**	21	**0**	**41**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**0**	**12**	**0**	**0**	**36**	**338**	**0**	**0**	**0**	**320**	**0**	**706**

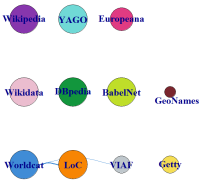

Table B.16: schema:sameAs traversal map for events

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	18	0	0	0	0	0	0	0	0	0	0	18
VIAF	7	0	0	0	0	0	0	0	0	0	0	7
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	0	0	0	0	0	0
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	25	0	0	0	0	0	0	0	0	0	0	25

Table B.17: Matrix data which generated the traversal map for dates (Figure 8)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	17	18	0	0	0	0	0	0	0	0	0	35
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	60	20	0	0	0	19	0	99
DBpedia	0	0	0	0	0	462	20	0	0	228	0	710
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	36	18	0	0	60	482	20	0	0	247	0	863

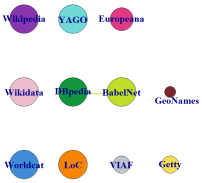

Table B.18: skos:exactMatch traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	20	0	0	0	0	20
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	0	0	0	0	0	20	0	0	0	0	20

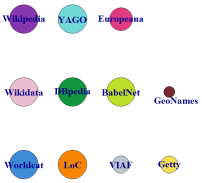

Table B.19: rdfs:seeAlso traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	18	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**18**
LoC	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
DBpedia	**0**	**0**	**0**	**0**	**0**	162	**0**	**0**	**0**	**0**	**0**	**162**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**18**	**0**	**0**	**0**	**0**	**162**	**0**	**0**	**0**	**0**	**0**	**180**

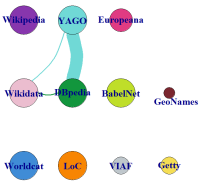

Table B.20: owl:sameAs traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
LoC	**0**	18	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**18**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	60	**20**	**0**	**0**	**0**	**19**	**0**	**99**
DBpedia	**0**	**0**	**0**	**0**	**0**	300	**0**	**0**	**0**	**228**	**0**	**528**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**0**	**18**	**0**	**0**	**60**	**320**	**0**	**0**	**0**	**247**	**0**	**645**

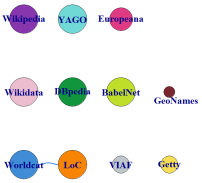

Table B.21: schema:sameAs traversal map for dates

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	1	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
LoC	**17**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**17**
VIAF	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Getty	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wikidata	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**
DBpedia	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**	**0**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**
YAGO	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**18**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**18**

Table B.22: Matrix data which generated the traversal map for places (Figure 10)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	**0**	**18**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**37**
LoC	**19**	1	**19**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**39**
VIAF	**19**	**20**	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**39**
Getty	**0**	**0**	**1**	20	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**21**
Wikidata	**0**	**0**	**21**	**0**	41	**20**	**0**	**0**	**0**	**20**	**0**	**102**
DBpedia	**0**	**0**	**17**	**0**	**0**	3854	**21**	**20**	**0**	**390**	**0**	**4302**
BabelNet	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**	**0**	**0**	**0**	**0**
GeoNames	**8**	**0**	**0**	**0**	**0**	**20**	**18**	0	**0**	**20**	**17**	**83**
Wikipedia	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**9**	**0**	**9**
YAGO	**0**	**0**	**0**	**0**	**0**	**26**	**20**	**0**	**0**	35	**0**	**81**
Europeana	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0	**0**
SUM	**65**	**21**	**76**	**20**	**41**	**3920**	**59**	**20**	**0**	**474**	**17**	**4713**

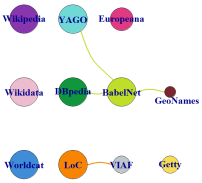

Table B.23: skos:exactMatch traversal map for places

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	20	0	0	0	0	0	0	0	0	0	20
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	21	0	0	0	0	21
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	18	0	0	0	0	18
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	20	0	0	0	0	20
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	20	0	0	0	0	59	0	0	0	0	79

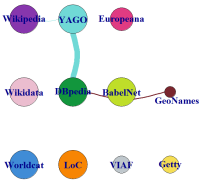

Table B.24: rdfs:seeAlso traversal map for places

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	0	1	0	0	0	0	0	0	0	0	0	1
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	20	0	0	0	0	0	0	0	20
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	3537	0	20	0	101	0	3658
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	7	0	7
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	19	1	0	20	0	3537	0	20	0	108	0	3705

Table B.25: owl:sameAs traversal map for places

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	41	20	0	0	0	20	0	81
DBpedia	0	0	0	0	0	300	0	0	0	287	0	587
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	20	0	0	0	20	16	56
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	26	0	0	0	33	0	59
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	0	0	0	41	366	0	0	0	360	16	783

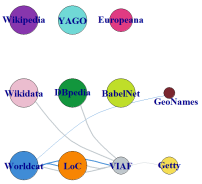

Table B.26: schema:sameAs traversal map for places

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	18	0	0	0	0	0	0	0	0	18
LoC	19	0	19	0	0	0	0	0	0	0	0	38
VIAF	19	0	0	0	0	0	0	0	0	0	0	19
Getty	0	0	1	0	0	0	0	0	0	0	0	1
Wikidata	0	0	21	0	0	0	0	0	0	0	0	21
DBpedia	0	0	17	0	0	0	0	0	0	0	0	17
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	8	0	0	0	0	0	0	0	0	0	0	8
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	46	0	76	0	0	0	0	0	0	0	0	122

Table B.27: Matrix data which generated the traversal map for objects and concepts (Figure 12)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	2	0	0	0	0	0	0	0	0	21
LoC	19	12	13	1	0	0	0	0	0	0	0	45
VIAF	12	12	0	0	0	0	0	0	0	0	0	24
Getty	0	0	2	8	0	0	0	0	0	0	0	10
Wikidata	1	0	2	0	0	20	0	0	0	20	0	43
DBpedia	0	0	2	0	0	312	20	3	0	75	0	412
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	2	2	0	0	0	0	4
Wikipedia	1	0	0	0	0	0	0	0	0	1084	0	1085
YAGO	0	0	0	0	0	23	18	0	0	6	0	47
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	52	24	21	9	0	357	40	3	0	1185	0	1691

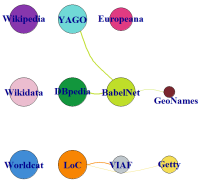

Table B.28: skos:exactMatch traversal map for objects and concepts

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	3	0	1	0	0	0	0	0	0	0	4
VIAF	0	12	0	0	0	0	0	0	0	0	0	12
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	0	0	0	0	0	0	0
DBpedia	0	0	0	0	0	0	20	0	0	0	0	20
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	2	0	0	0	0	2
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	0	18	0	0	0	0	18
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	15	0	1	0	0	40	0	0	0	0	56

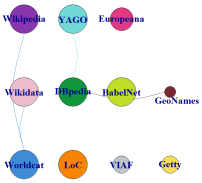

Table B.29: rdfs:seeAlso traversal map for objects and concepts

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	19	0	0	0	0	0	0	0	0	0	0	19
LoC	0	0	0	0	0	0	0	0	0	0	0	0
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	8	0	0	0	0	0	0	0	8
Wikidata	1	0	0	0	0	0	0	0	0	0	0	1
DBpedia	0	0	0	0	0	34	0	3	0	1	0	38
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	1	0	0	0	0	0	0	0	0	0	0	1
YAGO	0	0	0	0	0	0	0	0	0	0	0	0
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	21	0	0	8	0	34	0	3	0	1	0	67

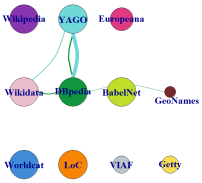

Table B.30: owl:sameAs traversal map for objects and concepts

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	0	0	0	0	0	0	0	0	0	0
LoC	0	9	0	0	0	0	0	0	0	0	0	9
VIAF	0	0	0	0	0	0	0	0	0	0	0	0
Getty	0	0	0	0	0	0	0	0	0	0	0	0
Wikidata	0	0	0	0	0	20	0	0	0	20	0	40
DBpedia	0	0	0	0	0	271	0	0	0	70	0	341
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	2	0	0	0	0	0	2
Wikipedia	0	0	0	0	0	0	0	0	0	0	0	0
YAGO	0	0	0	0	0	23	0	0	0	5	0	28
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	0	9	0	0	0	316	0	0	0	95	0	420

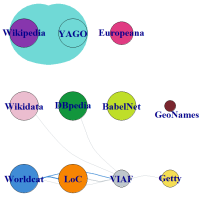

Table B.31: schema:sameAs traversal map for objects and concepts

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	Wikipedia	YAGO	Europeana	SUM
WorldCat	0	0	2	0	0	0	0	0	0	0	0	2
LoC	19	0	13	0	0	0	0	0	0	0	0	32
VIAF	12	0	0	0	0	0	0	0	0	0	0	12
Getty	0	0	2	0	0	0	0	0	0	0	0	2
Wikidata	0	0	2	0	0	0	0	0	0	0	0	2
DBpedia	0	0	2	0	0	0	0	0	0	0	0	2
BabelNet	0	0	0	0	0	0	0	0	0	0	0	0
GeoNames	0	0	0	0	0	0	0	0	0	0	0	0
Wikipedia	0	0	0	0	0	0	0	0	0	1084	0	1084
YAGO	0	0	0	0	0	0	0	0	0	1	0	1
Europeana	0	0	0	0	0	0	0	0	0	0	0	0
SUM	31	0	21	0	0	0	0	0	0	1085	0	1137

APPENDIX C: PYTHON ANALYSIS DETAILS

This appendix contains tables and figures created for Chapter 2. The tables are created by Python to analyze the number of concent in each category (agents, events, dates, places, objects and concepts).

[4]New York (state) might be low, due to its less popular concept, compared to countries and big cities. Somehow USA stands out, with a lot of unexpected contribution from Getty TGN.

	WorldCat	Full Coverage	Library of Congress	VIAF	Getty ULAN	Wikidata	Dbpedia	BabelNet	GeoNames	YAGO	Europeana
0	29136	29136				29136					
1	Daerwen, 1809-1882	Daerwen, 1809-1882	Daerwen, 1809-1882								
2	Daerwen, 1809-1882	Daerwen, 1809-1882	Daerwen, 1809-1882								
3	Darvin, Charl'z, 1809-1882	Darvin, Charl'z, 1809-1882	Darvin, Charl'z, 1809-1882								
4	Darvin, Charl'z, 1809-1882	Darvin, Charl'z, 1809-1882	Darvin, Charl'z, 1809-1882								
5	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882								
6	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882	Darvin, Tsharlz, 1809-1882								
7	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882								
8	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882	Darvin, Čarls, 1809-1882								
9	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882								
10	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882	Darwin, Carlos R., 1809-1882								
11	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882								
12	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882	Darwin, Charles Robert, 1809-1882								
13	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882			Darwin, Charles, 1809-1882					
14	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882	Darwin, Charles, 1809-1882			Darwin, Charles, 1809-1882					
15	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882								
16	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882	Darwin, Karol, 1809-1882								
17	Dāwin, 1809-1882	Dāwin, 1809-1882	Dāwin, 1809-1882								
18	Dāwin, 1809-1882	Dāwin, 1809-1882	Dāwin, 1809-1882								
19	Sdar-win, 1809-1882	Sdar-win, 1809-1882	Sdar-win, 1809-1882								
20	Sdar-win, 1809-1882	Sdar-win, 1809-1882	Sdar-win, 1809-1882								
21	Sdar-win, Char-le-si Ro-sbe-thi, 180	Sdar-win, Char-le-si Ro-sbe-thi, 180	Sdar-win, Char-le-si Ro-sbe-thi, 1809-1882								
22	Sdar-win, Char-le-si Ro-sbe-thi, 180	Sdar-win, Char-le-si Ro-sbe-thi, 180	Sdar-win, Char-le-si Ro-sbe-thi, 1809-1882								
23	Tārvin, 1809-1882	Tārvin, 1809-1882	Tārvin, 1809-1882								
24	Tārvin, 1809-1882	Tārvin, 1809-1882	Tārvin, 1809-1882								
25	Tārvin, Cārlas, 1809-1882	Tārvin, Cārlas, 1809-1882	Tārvin, Cārlas, 1809-1882								
26	Tārvin, Cārlas, 1809-1882	Tārvin, Cārlas, 1809-1882	Tārvin, Cārlas, 1809-1882								
27	http://id.loc.gov/authorities/name	http://id.loc.gov/authorities/name	http://id.loc.gov/authorities/name	http://id.loc.go	http://id.loc.gov/authorities/names/n78095637						
28	https://viaf.org/viaf/27063124	https://viaf.org/viaf/27063124									
29	schema:Person	schema:Person		schema:Person			schema:Person				
30	http://en.wikipedia.org/wiki/Charl	http://en.wikipedia.org/wiki/Charles_Darwin					http://en.wikipedia.org/wiki/Ch	http://en.wikipedia.or			
31	http://id.worldcat.org/fast/ontolog	http://id.worldcat.org/fast/ontology/1.0/#facet-Personal									
32	http://id.worldcat.org/fast/ontolog	http://id.worldcat.org/fast/ontology/1.0/#fast									
33	http://id.worldcat.org/fast/29136	http://id.worldcat.org/fast/29136	http://id.worldcat.org/fast/29136			http://id.worldcat.org/fast/29136					
34		madsrdf:Authority	madsrdf:Authority								
35		madsrdf:PersonalName	madsrdf:PersonalName								
36		skos:Concept	skos:Concept		skos:Concept			skos:Concept			

Figure C.1: Python scripts generate EXCEL files to show the content overlaps across 11 dataset. Content in the same row is overlap (example of Charles Darwin)

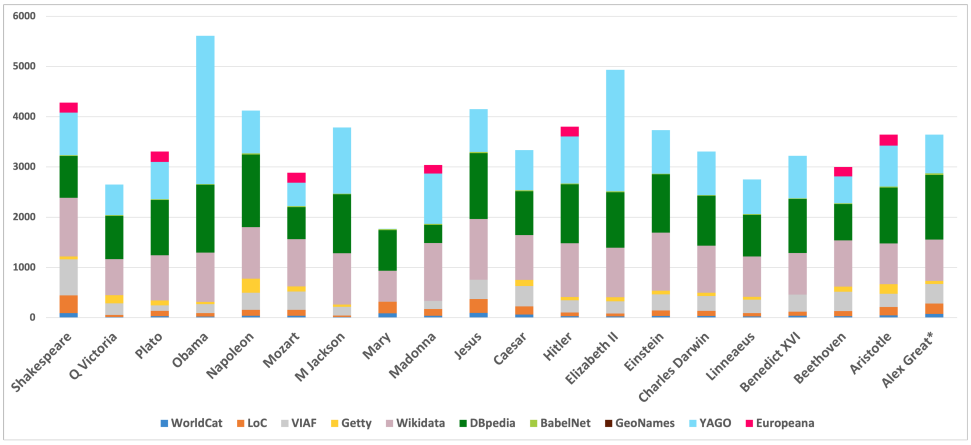

Figure C.2: The number of content in agents entities per data source[1]

Table C.1: The number of content in agents entities per data source

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
Shakespeare	92	353	718	57	1167	830	19	0	846	199	**3532**
Q Victoria	14	40	230	165	718	858	18	0	608	0	**2026**
Plato	30	104	113	96	901	1097	17	0	742	208	**2546**
Obama	24	67	178	44	984	1350	10	0	2954	0	**4764**
Napoleon	39	118	340	281	1025	1444	22	0	852	0	**3263**
Mozart	40	117	366	101	939	639	18	0	466	198	**2268**
M Jackson	14	29	173	46	1022	1170	15	0	1317	0	**2965**
Mary	89	229	6	0	609	808	20	0	8	0	**1643**
Madonna	40	133	161	0	1152	363	19	0	1002	169	**2615**
Jesus	97	274	386	0	1209	1307	26	0	853	0	**3152**
Caesar	64	163	404	122	893	870	20	0	801	0	**2570**
Hitler	29	74	245	64	1069	1172	17	0	938	195	**2872**
Elizabeth II	25	60	240	84	983	1105	19	0	2418	0	**4078**
Einstein	36	109	321	72	1154	1161	14	0	867	0	**2965**
Charles Darwin	34	100	297	65	936	994	13	0	867	0	**2579**
Linneaeus	28	63	269	56	802	828	15	0	690	0	**2084**
Benedict XVI	37	84	339	0	827	1077	10	0	850	0	**2331**
Beethoven	32	99	387	99	921	723	16	0	535	188	**2423**
Aristotle	48	164	263	191	811	1113	21	0	814	218	**2791**
Alex Great*	74	207	390	62	822	1287	30	0	773	0	**2732**
SUM	**886**	**2587**	**5826**	**1605**	**18944**	**20196**	**359**	**0**	**19201**	**1375**	**56199**

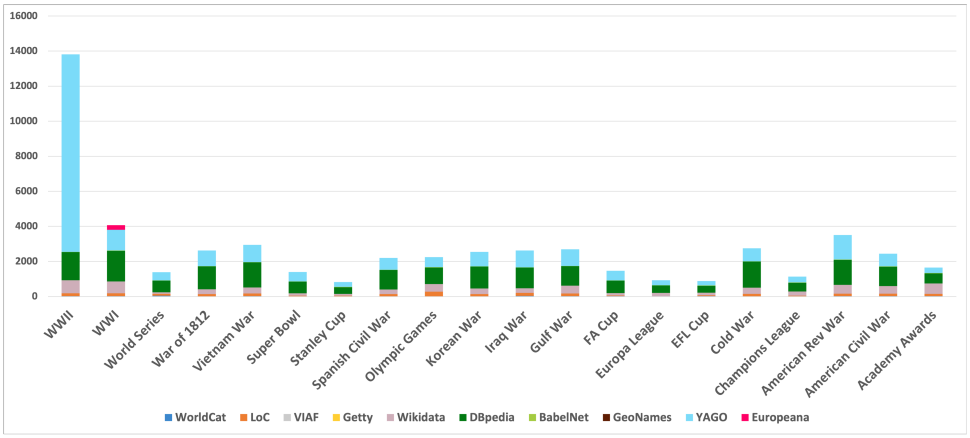

Figure C.3: The number of content in events entities per data source[2]

Table C.2: The number of content in events entities per data source

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
WWII	26	151	0	0	749	1601	15	0	11273	0	12811
WWI	26	155	0	0	677	1740	21	0	1189	256	2939
World Series	39	111	0	0	89	662	16	0	471	0	924
War of 1812	12	128	0	0	266	1308	18	0	897	0	2156
Vietnam War	19	147	0	0	350	1426	15	0	991	0	2288
Super Bowl	11	28	0	0	143	663	18	0	539	0	1016
Stanley Cup	11	30	0	0	109	378	17	0	278	0	571
Spanish Civil War	10	124	0	0	262	1113	13	0	679	0	1656
Olympic Games	10	261	0	0	436	941	28	0	566	0	1749
Korean War	11	125	0	0	326	1245	14	0	817	0	1922
Iraq War	33	176	0	0	255	1186	9	0	965	0	1946
Gulf War	26	143	0	0	447	1115	12	0	949	0	2013
FA Cup	17	41	0	0	130	715	16	0	551	0	1036
Europa League	0	0	0	0	202	421	12	0	290	0	723
EFL Cup	21	45	0	0	148	395	14	0	253	0	665
Cold War	11	132	0	0	364	1484	16	0	740	0	2235
Champions League	14	33	0	0	233	498	18	0	339	0	908
American Rev War	16	146	0	0	498	1421	29	0	1401	0	3140
American Civil War	17	147	0	0	431	1095	27	0	729	0	1953
Academy Awards	16	130	0	0	593	582	17	0	310	0	1393
SUM	**346**	**2253**	**0**	**0**	**6708**	**19989**	**345**	**0**	**24227**	**256**	**44044**

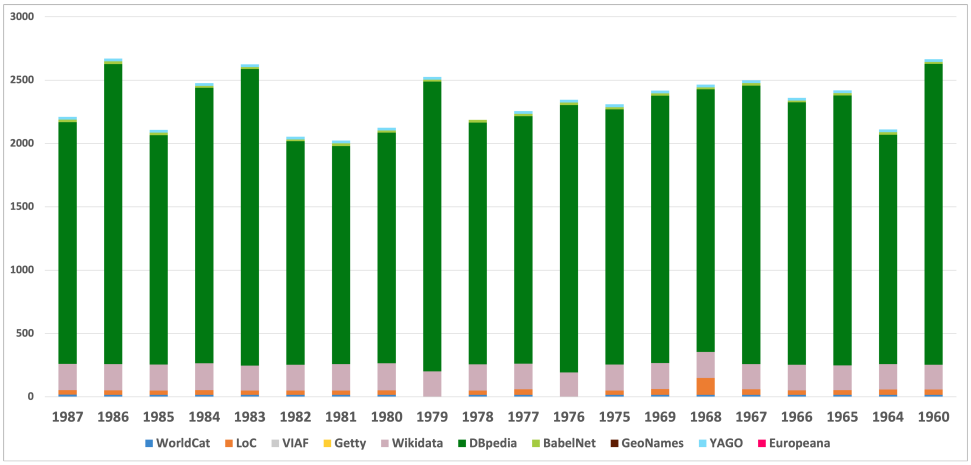

Figure C.4: The number of content in dates entities per data source[3]

Table C.3: The number of content in dates entities per data source

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
1987	18	35	0	0	207	1909	20	0	22	0	2190
1986	16	35	0	0	207	2368	22	0	22	0	2650
1985	16	34	0	0	205	1810	20	0	22	0	2087
1984	15	37	0	0	213	2175	15	0	22	0	2458
1983	15	34	0	0	198	2341	15	0	22	0	2606
1982	15	34	0	0	205	1764	14	0	22	0	2035
1981	15	34	0	0	209	1722	21	0	22	0	2004
1980	15	36	0	0	214	1820	17	0	22	0	2105
1979	0	0	0	0	201	2288	15	0	22	0	2519
1978	15	35	0	0	207	1907	22	0	0	0	2169
1977	16	44	0	0	202	1953	19	0	22	0	2236
1976	0	0	0	0	193	2110	21	0	22	0	2339
1975	15	35	0	0	205	2014	18	0	22	0	2289
1969	15	46	0	0	205	2111	18	0	22	0	2398
1968	15	133	0	0	207	2073	15	0	22	0	2446
1967	15	44	0	0	200	2198	19	0	22	0	2479
1966	15	36	0	0	203	2071	14	0	22	0	2341
1965	15	37	0	0	197	2130	18	0	22	0	2399
1964	15	43	0	0	200	1811	20	0	22	0	2092
1960	15	42	0	0	196	2376	15	0	22	0	2647
SUM	**276**	**774**	**0**	**0**	**4074**	**40951**	**358**	**0**	**418**	**0**	**46489**

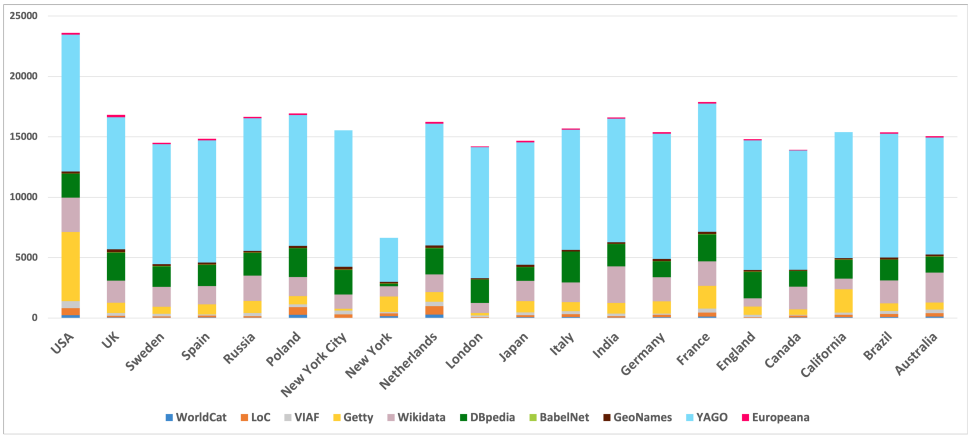

Figure C.5: The number of content in places entities per data source[4]

Table C.4: The number of content in places entities per data source

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
USA	241	580	579	5725	2847	1979	12	176	11329	150	21351
UK	58	134	227	852	1828	2318	33	237	10942	192	15022
Sweden	34	100	219	585	1647	1698	27	159	9925	120	12829
Spain	56	153	93	828	1512	1758	22	174	10113	140	13283
Russia	30	126	265	996	2106	1874	25	139	10979	114	15208
Poland	273	627	228	687	1584	2371	22	177	10832	139	14840
New York City	0	307	309	151	1191	2026	36	233	11289	0	13958
New York	153	230	128	1266	860	203	34	131	3631	0	6278
Netherlands	285	697	363	804	1458	2177	22	204	10081	155	13845
London	21	48	159	188	832	1928	28	101	10836	71	12932
Japan	63	167	240	937	1668	1131	19	193	10125	140	13334
Italy	84	235	251	744	1636	2555	0	134	9954	105	13914
India	37	110	227	876	3035	1838	24	136	10219	110	15229
Germany	74	202	100	1003	2006	1303	27	188	10351	146	13909
France	127	334	329	1870	2030	2241	26	180	10612	143	15502
England	21	57	194	679	689	2172	30	139	10730	100	12831
Canada	50	147	78	441	1878	1298	24	81	9882	57	12963
California	64	208	175	1934	880	1549	35	110	10446	0	14117
Brazil	92	242	245	635	1904	1708	24	155	10256	118	13777
Australia	113	293	293	582	2477	1324	27	153	9683	115	13685
SUM	**1876**	**4997**	**4702**	**21783**	**34068**	**35451**	**497**	**3200**	**202215**	**2115**	**278807**

Table C.5: The number of content in objects and concepts entities per data source (see Figure 15 in the main text)

	WorldCat	LoC	VIAF	Getty	Wikidata	DBpedia	BabelNet	GeoNames	YAGO	Europeana	Full Coverage
Vasa	11	24	46	0	146	386	11	0	107	0	**626**
Uncle Tom's Cabin	13	25	193	0	226	516	0	0	113	0	**1015**
Ukiyo-e	24	152	0	118	151	609	19	0	137	0	**1096**
Tosca	11	53	21	0	162	413	11	0	97	0	**718**
Toraja	17	56	0	0	87	23	11	0	71	0	**215**
Tamil Language	15	162	0	49	403	655	14	0	481	0	**1392**
Sgt. Pepper's	11	24	22	0	246	1007	14	0	666	0	**1479**
Rosetta Stone	19	46	44	0	249	592	11	0	188	0	**985**
Like a Rolling Stone	9	20	11	0	71	383	19	0	75	0	**547**
Palazzo Pitti	16	37	21	30	178	367	22	47	134	0	**703**
Ming Dynasty	0	120	6	77	369	963	24	0	269	0	**1659**
Mars	12	164	34	28	668	957	21	0	767	0	**2118**
King and I*	9	46	48	0	95	482	18	0	375	0	**718**
Book of Kells	15	47	43	0	147	406	11	0	134	0	**669**
Influenza	23	159	0	48	418	433	20	0	512	0	**1440**
Garden of E Delights	17	36	32	0	253	395	18	0	126	0	**776**
Byzantine Empire	27	63	221	91	626	1310	17	0	393	0	**2282**
Boeing 747	12	41	0	0	164	560	18	0	143	0	**863**
Blade Runner	9	42	29	0	490	639	11	0	191	0	**1298**
Angkor Wat	13	72	62	65	240	569	17	37	253	0	**1100**
SUM	**283**	**1389**	**833**	**506**	**5389**	**11665**	**307**	**84**	**5232**	**0**	**21699**

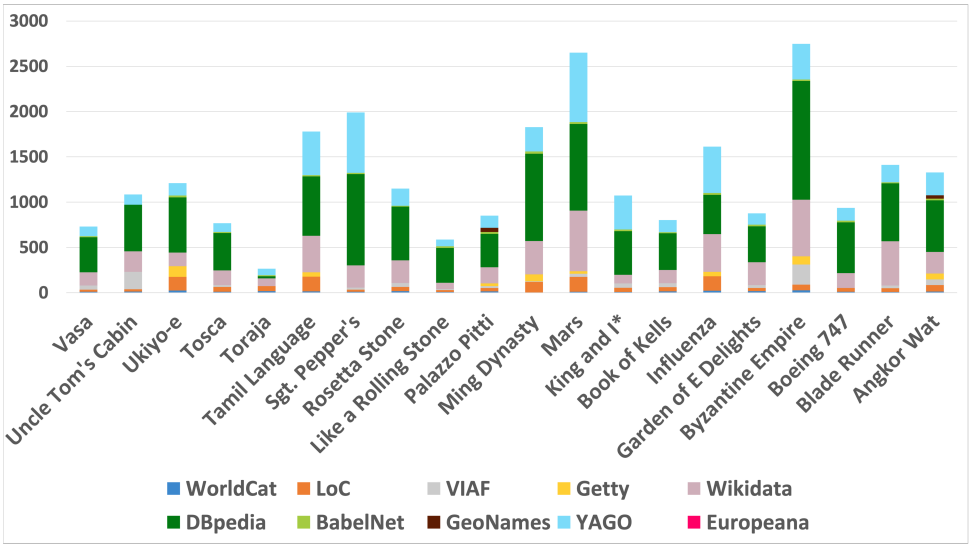

Figure C.6: The number of content in objects and concepts entities per data source (see Figure 15 in the main text)

[4]YAGO contains a large amount of content for WWII.

BIBLIOGRAPHY

REFERENCES

[1] CIDOC CRM. Accessed on 2024-07-08. URL: https://www.cidoc-crm.org/.

[2] Compatible models & collaborations | CIDOC CRM. Accessed on 2024-10-14. URL: https://cidoc-crm.org/collaborations.

[3] Conzept encyclopedia. Accessed on 2024-07-08. URL: https://conze.pt/explore.

[4] Crotos. Accessed on 2024-07-08. URL: https://zone47.com/crotos/.

[5] Egon schiele. Accessed on 2024-07-08. URL: https://www.wikidata.org/wiki/Q44032.

[6] EntiTree. Accessed on 2024-07-08. URL: https://www.entitree.com/.

[7] Geneawiki. Accessed on 2024-07-08. URL: https://magnus-toolserver.toolforge.org/ts2/geneawiki/.

[8] Gustav klimt. Accessed on 2024-07-08. URL: https://www.wikidata.org/wiki/Q34661.

[9] Histomania. Accessed on 2024-07-08. URL: https://histomania.com/.

[10] Histropedia - the timeline of everything. Accessed on 2024-07-08. URL: https://histropedia.com/timeline-everything.

[11] Linked people. Accessed on 2024-07-08. URL: https://linkedpeople.net/.

[12] open art browser. Accessed on 2024-07-08. URL: https://openartbrowser.org/en/.

[13] Our mission. Accessed on 2024-11-07. URL: https://www.w3.org/mission/.

[14] OWL - semantic web standards. Accessed on 2024-08-20. URL: https://www.w3.org/OWL/.

[15] PeriodO – periods, organized. Accessed on 2024-07-08. URL: https://perio.do/en/.

[16] RDF - semantic web standards. Accessed on 2024-08-20. URL: https://www.w3.org/RDF/.

[17] RDF schema 1.1. Accessed on 2024-08-21. URL: https://www.w3.org/TR/rdf11-schema/.

[18] RdfSyntax - w3c wiki. Accessed on 2024-08-20. URL: https://www.w3.org/wiki/RdfSyntax.

[19] Reasonator. Accessed on 2024-07-08. URL: https://reasonator.toolforge.org/.

[20] Scholia. Accessed on 2024-07-08. URL: https://scholia.toolforge.org/.

[21] SPARQL 1.1 query language. Accessed on 2024-08-21. URL: https://www.w3.org/TR/sparql11-query/.

[22] SPARQL 1.1 query language. Accessed on 2024-07-08. URL: https://www.w3.org/TR/sparql11-query/.

[23] Thesaurus - archivführer deutsche kolonialgeschichte. Accessed on 2024-07-08. URL: https://archivfuehrer-kolonialzeit.de/thesaurus.

[24] Visual analysis, curation & communication for in/tangible european heritage. Accessed on 2024-07-08. URL: https://intavia.acdh-dev.oeaw.ac.at.

[25] ViziData. Accessed on 2024-07-08. URL: https://sylum.lima-city.de/viziData/.

[26] Welcome - Ariadne portal. https://portal.ariadne-infrastructure.eu/. Accessed on 2025-07-30. URL: https://portal.ariadne-infrastructure.eu/.

[27] Wikidata. Accessed on 2024-09-17. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page.

[28] Wikidata query service/user manual. Accessed on 2024-07-08. URL: https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual.

[29] Wikidata reconciliation for OpenRefine. Accessed on 2024-07-08. URL: https://wikidata.reconci.link/en/api.

[30] Wikidata tempo-spatial display. Accessed on 2024-07-08. URL: https://wikidata-todo.toolforge.org/tempo_spatial_display.html.

[31] Wikidata tree builder. Accessed on 2024-07-08. URL: https://wikioverdata.toolforge.org/wikitree/public/.

[32] Wikidata visualization. Accessed on 2024-07-08. URL: https://dataviz.toolforge.org/.

[33] Wikidata:SPARQL query service/WDQS graph split - wikidata. Accessed on 2024-07-08. URL: https://m.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_graph_split.

[34] Wikidata:tools - wikidata. Accessed on 2024-07-08. URL: https://www.wikidata.org/wiki/Wikidata:Tools.

[35] yaap! explore history with a timeline. Accessed on 2024-07-08. URL: https://yaap.ch/.

[36] Data on the web best practices, 2017. Accessed on 2024-08-20. URL: https://www.w3.org/TR/dwbp/.

[37] Manel Achichi, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy, and Jean Delahousse. DOREMUS: A graph of linked musical works. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, Lecture Notes in Computer Science, pages 3–19. Springer International Publishing, 2018. doi:10.1007/978-3-030-00668-6_1.

[38] Eneko Agirre, A. Barrena, Oier Lopez de Lacalle, A. Soroa, Samuel Fernando, and Mark Stevenson. Matching cultural heritage items to wikipedia. In *LREC*, 2012.

[39] Dirk Ahlers. Linkage quality analysis of GeoNames in the semantic web. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*, GIR'17, pages 10:1–10:2. ACM, 2017. Accessed on 2018-10-14. URL: http://doi.acm.org/10.1145/3155902.3155904, doi:10.1145/3155902.3155904.

[40] Mehwish Alam, Aldo Gangemi, Valentina Presutti, and Diego Reforgiato Recupero. Semantic role labeling for knowledge graph extraction from text. 10(3):309–320, 2021. Accessed on 2023-07-12. doi:10.1007/s13748-021-00241-7.

[41] James F. Allen. Maintaining knowledge about temporal intervals. 26(11):832–843, 1983. Accessed on 2020-10-08. URL: https://dl.acm.org/doi/10.1145/182.358434, doi:10.1145/182.358434.

[42] James F. Allen and George Ferguson. Actions and events in interval temporal logic. In Oliviero Stock, editor, *Spatial and Temporal Reasoning*, pages 205–245. Springer Netherlands, 1997. Accessed on 2020-03-13. URL: http://link.springer.com/10.1007/978-0-585-28322-7_7, doi:10.1007/978-0-585-28322-7_7.

[43] A. B. Antopolsky. Linked open data in the digital humanities (review of publications). 49(2):119–126, 2022. Accessed on 2024-06-30. doi:10.3103/S014768822202006X.

[44] Collin F. Baker. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5. Association for Computational Linguistics, 2014. Accessed on 2023-07-14. URL: https://aclanthology.org/W14-3001, doi:10.3115/v1/W14-3001.

[45] Wouter Beek, Joe Raad, Jan Wielemaker, and Frank van Harmelen. sameAs.cc: The closure of 500m owl:sameAs statements. In Aldo Gangemi, Roberto Navigli,

Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 65–80. Springer International Publishing, 2018.

[46] Mike Belshe, Roberto Peon, and Martin Thomson. Hypertext transfer protocol version 2 (HTTP/2), 2015. Accessed on 2024-09-17. URL: https://datatracker.ietf.org/doc/rfc7540, doi:10.17487/RFC7540.

[47] Tim Berners-Lee. Semantic web roadmap, 1998. Accessed on 2024-09-16. URL: https://www.w3.org/DesignIssues/Semantic.html.

[48] Tim Berners-Lee. Linked data - design issues, 2009. Accessed on 2018-04-24. URL: https://www.w3.org/DesignIssues/LinkedData.html.

[49] Tim Berners-Lee, Roy T. Fielding, and Larry M. Masinter. Uniform resource identifier (URI): Generic syntax, 2005. Accessed on 2024-09-17. URL: https://datatracker.ietf.org/doc/rfc3986, doi:10.17487/RFC3986.

[50] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[51] V. de Boer, A. Merono Penuela, and C. J. Ockeloen. Linked data for digital history: Lessons learned from three case studies. (4):139–162, 2016. Accessed on 2024-07-08. URL: https://research.vu.nl/en/publications/linked-data-for-digital-history-lessons-learned-from-three-case-s.

[52] Victor de Boer and Lise Stork. Hybrid intelligence for digital humanities, 2024. Accessed on 2025-01-29. URL: http://arxiv.org/abs/2406.15374, arXiv:2406.15374[cs], doi:10.48550/arXiv.2406.15374.

[53] Josep Maria Brunetti, Sören Auer, and Roberto García. The linked data visualization model. 2012.

[54] Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. A linked open data framework to enhance the discoverability and impact of culture heritage. 45(6):756–766, 2019. Accessed on 2020-12-22. doi:10.1177/0165551518812658.

[55] Gustavo Candela, Pilar Escobar, Rafael C Carrasco, and Manuel Marco-Such. Evaluating the quality of linked open data in digital libraries. page 0165551520930951, 2020. Accessed on 2020-12-22. doi:10.1177/0165551520930951.

[56] Gianluca Correndo, Antonio Penta, Nicholas Gibbins, and Nigel Shadbolt. Statistical analysis of the owl:sameAs network for aligning concepts in the linking open data cloud. In Stephen W. Liddle, Klaus-Dieter Schewe, A. Min Tjoa, and Xiaofang Zhou, editors, *Database and Expert Systems Applications*, Lecture Notes in Computer Science, pages 215–230. Springer Berlin Heidelberg, 2012.

[57] Simon Cox, Chris Little, J. Hobbs, and Feng Pan. Time ontology in OWL, 2017. Accessed on 2024-07-08. URL: https://www.w3.org/TR/2017/REC-owl-time-20171019/.

[58] Simon J.D. Cox. Time ontology extended for non-gregorian calendar applications. 7(2):201–209, 2016. Accessed on 2020-03-23. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-150187, doi:10.3233/SW-150187.

[59] Costis Dallas, Nephelie Chatzidiakou, Agiatis Benardou, Michael Bender, Aurélien Berra, Claire Clivaz, John Cunningham, Meredith Dabek, Patricia Garrido, Elena Gonzalez-Blanco, Jurij Hadalin, Lorna Hughes, Beat Immenhauser, Anne Joly, Ingrida Kelpšienė, Michał Kozak, Koraljka Kuzman, Marko Lukin, Irena Marinski, Maciej Maryl, Robert Owain, Eliza Papaki, Gerlinde Schneider, Walter Scholger, Susan Schreibman, Zoe Schubert, Toma Tasovac, Manfred Thaller, Piotr Wciślik, Marcin Werla, and Tvrtko Zebec. European survey on scholarly practices and digital needs in the arts and humanities - highlights report, 2017. Accessed on 2024-07-08. URL: https://zenodo.org/records/260101, doi:10.5281/zenodo.260101.

[60] Stephen Boyd Davis, Emma Bevan, and Aleksei Kudikov. Just in time: Defining historical chronographics. In Jonathan P. Bowen, Suzanne Keene, and Kia Ng, editors, *Electronic Visualisation in Arts and Culture*, pages 243–257. Springer, 2013. Accessed on 2024-07-07. doi:10.1007/978-1-4471-5406-8_17.

[61] Viktor de Boer, Maarten van Someren, and Bob J. Wielinga. Extracting historical time periods from the web. 61(9):1888–1908, 2010. Accessed on 2020-03-11. URL: http://doi.wiley.com/10.1002/asi.21378, doi:10.1002/asi.21378.

[62] Max De Wilde, Max De Wilde, and Simon Hengchen. Semantic enrichment of a multilingual archive with linked open data. 011(4), 2018.

[63] Jeremy Debattista, Eamon Clinton, and Rob Brennan. Assessing the quality of geospatial linked data –experiences from ordnance survey ireland (OSi). Vol-2198, 2018. Accessed on 2018-09-12. URL: http://ceur-ws.org/Vol-2198/paper_94.pdf.

[64] Jeremy Debattista, Christoph Lange, and Sören Auer. Luzzu - a framework for linked data quality assessment. 2014. Accessed on 2018-10-18. URL: http://arxiv.org/abs/1412.3750, arXiv:1412.3750.

[65] Jeremy Debattista, Christoph Lange, Sören Auer, and Dominic Cortis. Evaluating the quality of the lod cloud: An empirical investigation. 9:859–901, 2018. Accessed on 2024-07-08. doi:10.3233/SW-180306.

[66] Alessio Di Pasquale, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali. On assessing weaker logical status claims in wikidata cultural heritage records. Accessed on 2023-07-10. URL: https://github.com/alessiodipasquale/Wikidata_WLS.

[67] Alessio Di Pasquale, Valentina Pasqual, Francesca Tomasi, and Fabio Vitali. Representation of critical discourses in the humanities within wikidata. Accessed on 2024-07-11, 2023. URL: https://zenodo.org/records/8107888, doi:10.5281/zenodo.8107888.

[68] Li Ding, Joshua Shinavier, Timothy W. Finin, and Deborah L. McGuinness. owl:sameas and linked data: An empirical study. 2010. Accessed on 2024-07-08. doi:10.13016/M2XP6V734.

[69] Li Ding, Joshua Shinavier, Zhenning Shangguan, and Deborah L. McGuinness. SameAs networks and beyond: Analyzing deployment status and implications of owl:sameAs in linked data. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web – ISWC 2010*, Lecture Notes in Computer Science, pages 145–160. Springer Berlin Heidelberg, 2010.

[70] Martin Doerr, Richard Light, and Gerald Hiebel. Implementing the CIDOC conceptual reference model in RDF version 1.1, 2020.

[71] Marian Dörk, Christopher Pietsch, and Gabriel Credico. One view is not enough: High-level visualizations of a large cultural collection. *Information Design Journal*, 23(1):39–47, January 2017. Accessed on 2024-11-20. URL: https://www.jbe-platform.com/content/journals/10.1075/idj.23.1.06dor, doi:10.1075/idj.23.1.06dor.

[72] Jeffrey Edelstein, Carolyn Li-Madeo, Noreen Whysel, and Marden, Marden. Linked open data for cultural heritage: evolution of an information technology. In *Proceedings of the 31st ACM international conference on Design of communication*, SIGDOC '13, pages 107–112. Association for Computing Machinery, 2013. Accessed on 2024-08-14. doi:10.1145/2507065.2507103.

[73] Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L. Shepelyansky. Interactions of cultures and top people of wikipedia from ranking of 24 language editions. 10(3):e0114825, 2015. Accessed on 2019-06-21. URL: http://arxiv.org/abs/1405.7183, arXiv:1405.7183, doi:10.1371/journal.pone.0114825.

[74] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. Introducing wikidata to the linked data web. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014*, volume 8796, pages 50–65. Springer International Publishing, 2014. Accessed on 2021-01-26. URL: http://link.springer.com/10.1007/978-3-319-11964-9_4, doi:10.1007/978-3-319-11964-9_4.

[75] Peter Exner and Pierre Nugues. Entity extraction: From unstructured text to DBpedia RDF triples. 2012. Accessed on 2024-07-08. URL: https://ceur-ws.org/Vol-906/paper7.pdf.

[76] Michael Farag. Entity matching and disambiguation across multiple knowledge graphs, 2019. Accessed on 2021-01-26. URL: https://uwspace.uwaterloo.ca/handle/10012/14750.

[77] Achille Felicetti. D4.3 – Final report on dataset integration. August 2022. Accessed on 2025-07-30. doi:10.5281/zenodo.7612672.

[78] Erwin Folmer and Jack Verhoosel. State of the Art on Semantic IS Standardization, Interoperability & Quality. 2011. URL: https://ris.utwente.nl/ws/portalfiles/portal/5132324/Folmer-SOTA_web.pdf.

[79] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. 9(1):77–129, 2017. Accessed on 2019-12-06. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-170275, doi:10.3233/SW-170275.

[80] Haris Georgiadis, Agathi Papanoti, Maria Paschou, Alexandra Roubani, Despina Hardouveli, and Evi Sachini. The semantic enrichment strategy for types, chronologies and historical periods in searchculture.gr. In Emmanouel Garoufallou, Sirje Virkus, Rania Siatri, and Damiana Koutsomiha, editors, *Metadata and Semantic Research*, Communications in Computer and Information Science, pages 211–223. Springer International Publishing, 2017. Accessed on 2024-07-08. doi:10.1007/978-3-319-70863-8_20.

[81] Katrin Glinka, Christopher Pietsch, and Marian Dörk. Past visions and reconciling views: Visualizing time, texture and themes in cultural collections. 11(2), 2017. Accessed on 2024-06-28. URL: https://www.digitalhumanities.org/dhq/vol/11/2/000290/000290.html#drucker2011.

[82] Sugimoto Go. Open data empowerment of digital humanities by wikipedia/DBpedia gamification and crowd curation –WiQiZi's challenges with APIs and SPARQL. Accessed on 2024-09-25, 2019. URL: https://zenodo.org/records/3465654, doi:10.5281/zenodo.3465654.

[83] Thomas Gottron, Ansgar Scherp, Bastian Krayer, and Arne Peters. Get the google feeling: Supporting users in finding relevant sources of linked open data at web-scale. 2012. Accessed on 2024-07-08. URL: https://pdfs.semanticscholar.org/43a9/670c57fec2a4d05ff72429886e457a88e59f.pdf.

[84] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 87–102. Springer, 2012. Accessed on 2024-07-08. doi:10.1007/978-3-642-30284-8_13.

[85] Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. When owl:sameAs isn't the same: An analysis of identity in linked data. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei

Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web – ISWC 2010*, Lecture Notes in Computer Science, pages 305–320. Springer Berlin Heidelberg, 2010.

[86] Olaf Hartig. SQUIN: a traversal based query execution system for the web of linked data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 1081–1084. ACM, 2013.

[87] Olaf Hartig and M. Tamer Özsu. Walking without a map: Ranking-based traversal for querying linked data. In *International Semantic Web Conference*, pages 305–324. Springer, 2016.

[88] Ashleigh Hawkins. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. 22(3):319–344, 2022. Accessed on 2024-09-11. doi:10.1007/s10502-021-09381-0.

[89] Jerry R. Hobbs and Feng Pan. An ontology of time for the semantic web. 3(1):66–85, 2004. Accessed on 2020-03-23. URL: http://dl.acm.org/doi/10.1145/1017068.1017073, doi:10.1145/1017068.1017073.

[90] Aidan Hogan. *The Web of Data*. Springer Cham, 2020. Accessed on 2024-09-11. URL: https://doi.org/10.1007/978-3-030-51580-5.

[91] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *EMNLP 2021*, pages 2370–2381. Association for Computational Linguistics, 2021. Accessed on 2024-07-08. doi:10.18653/v1/2021.findings-emnlp.204.

[92] Eero Hyvönen. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on Data, Semantics, and Knowledge. Springer International Publishing, 2012. Accessed on 2024-09-24. URL: https://link.springer.com/10.1007/978-3-031-79438-4, doi:10.1007/978-3-031-79438-4.

[93] Eero Hyvönen. Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. 11(1):187–193, 2020. Accessed on 2024-09-12. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-190386, doi:10.3233/SW-190386.

[94] Eero Hyvönen, Thea Lindquist, Juha Törnroos, and Eetu Mäkelä. History on the semantic web as linked data. 2012. Accessed on 2024-07-08. URL: https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/hyvonen.pdf.

[95] Al Idrissou, Frank van Harmelen, and Peter van den Besselaar. Network metrics for assessing the quality of entity resolution between multiple datasets. 12(1):21–40, 2021. Accessed on 2021-01-24. URL: https://content.iospress.com/articles/semantic-web/sw200410, doi:10.3233/SW-200410.

[96] Afraz Jaffri, Hugh Glaser, and Ian Millard. Managing URI synonymity to enable consistent reference on the semantic web. IRSW2008 - Identity and Reference on the Semantic Web 2008. Accessed on 2024-07-08, 2008. URL: https://eprints.soton.ac.uk/265614/.

[97] Afraz Jaffri, Hugh Glaser, and Ian Millard. URI disambiguation in the context of linked data. Linked Data on the Web (LDOW2008). Accessed on 2024-07-08, 2008. URL: https://eprints.soton.ac.uk/265181/.

[98] Krzysztof Janowicz. The role of space and time for knowledge organization on the semantic web. 1(1):25–32, 2010. Accessed on 2024-07-07. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-2010-0001, doi:10.3233/SW-2010-0001.

[99] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. 2015. Accessed on 2020-02-27. URL: https://diglib.eg.org:443/xmlui/handle/10.2312/eurovisstar.20151113.083-103, doi:10.2312/eurovisstar.20151113.

[100] Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. OpenIE6: Iterative grid labeling and coordination analysis for open information extraction. In *EMNLP*, pages 3748–3761. Association for Computational Linguistics, 2020. Accessed on 2024-07-08. doi:10.18653/v1/2020.emnlp-main.306.

[101] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009. Association for Computational Linguistics, 2020. Accessed on 2024-07-08. doi:10.18653/v1/2020.acl-main.713.

[102] Jose L. Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: A survey. 11(2):255–335, 2020. Accessed on 2023-03-01. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-180333, doi:10.3233/SW-180333.

[103] Arcangelo Massari, Silvio Peroni, Francesca Tomasi, and Ivan Heibi. Representing provenance and track changes of cultural heritage metadata in RDF: a survey of existing approaches, 2023. Accessed on 2024-07-11. URL: http://arxiv.org/abs/2305.08477, arXiv:2305.08477[cs], doi:10.48550/arXiv.2305.08477.

[104] R. Maturana, M. Ortega, Susana López-Sola, María Elena Alvarado, and M. J. Ibáñez. Mismuseos.net: Art after technology. putting cultural data to work in a linked data platform. In *Veni@OKCon*, 2013.

[105] Lucy McKenna, Christophe Debruyne, and Declan O'Sullivan. Understanding the position of information professionals with regards to linked data: A survey of libraries, archives and museums. In *Proceedings of the 18th ACM/IEEE on Joint Conference*

on Digital Libraries, JCDL '18, pages 7–16. Association for Computing Machinery, 2018. Accessed on 2024-09-24. doi:10.1145/3197026.3197041.

[106] Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. Semantic technologies for historical research: A survey. 6(6):539–564, 2014. Accessed on 2024-07-08. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-140158, doi:10.3233/SW-140158.

[107] David Milne and Ian Witten. Learning to link with wikipedia. 2008. Accessed on 2024-07-08. doi:10.1145/1458082.1458150.

[108] Franco Moretti. *Distant Reading*. Verso, 2013. OCLC: 813931586.

[109] Michalis Mountantonakis and Yannis Tzitzikas. High performance methods for linked open data connectivity analytics. 9:134, 2018. Accessed on 2024-07-08. doi:10.3390/info9060134.

[110] Dat P.T. Nguyen, Yutaka Matsuo, and Mitsuru Ishizuka. Subtree mining for relation extraction from wikipedia. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 125–128. Association for Computational Linguistics, 2007. Accessed on 2023-03-01. URL: https://aclanthology.org/N07-2032.

[111] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. 31(1):71–106, 2005. Accessed on 2023-07-14. URL: https://direct.mit.edu/coli/article/31/1/71-106/1861, doi:10.1162/0891201053630264.

[112] Laura Po, Nikos Bikakis, Federico Desimoni, and George Papastefanatos. *Linked Data Visualization: Techniques, Tools, and Big Data*. Synthesis Lectures on Data, Semantics, and Knowledge. Springer International Publishing, 2020. Accessed on 2024-06-25. URL: https://link.springer.com/10.1007/978-3-031-79490-2, doi:10.1007/978-3-031-79490-2.

[113] Joe Raad, Wouter Beek, Frank van Harmelen, Nathalie Pernelle, and Fatiha Saïs. Detecting erroneous identity links on the web using network metrics. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*, volume 11136, pages 391–407. Springer International Publishing, 2018. Accessed on 2019-04-01. URL: http://link.springer.com/10.1007/978-3-030-00671-6_23, doi:10.1007/978-3-030-00671-6_23.

[114] Julian Richards, Achille Felicetti, Carlo Meghini, and Maria Theodoridou. D4.4 – Final report on ontology implementation. December 2022. Accessed on 2025-07-30. doi:10.5281/zenodo.7636720.

[115] Anisa Rula, Andrea Maurino, and Carlo Batini. Data quality issues in linked open data. In Carlo Batini and Monica Scannapieco, editors, *Data and Information Quality: Dimensions, Principles and Techniques*, Data-Centric Systems and Applications, pages 87–112. Springer International Publishing, 2016. Accessed on 2018-10-18. doi:10.1007/978-3-319-24106-7_4.

[116] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *International Semantic Web Conference*, 2014. Accessed on 2024-07-08. doi:10.1007/978-3-319-11964-9_16.

[117] Susan Schreibman, Ray Siemens, and John Unsworth. *A New Companion to Digital Humanities*. John Wiley & Sons, 2016.

[118] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. LODE: Linking open descriptions of events. In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *The Semantic Web*, volume 5926, pages 153–167. Springer Berlin Heidelberg, 2009. Accessed on 2024-12-03. URL: http://link.springer.com/10.1007/978-3-642-10871-6_11, doi:10.1007/978-3-642-10871-6_11.

[119] Sarah Binta Alam Shoilee, Victor de Boer, and Jacco van Ossenbruggen. Polyvocal knowledge modelling for ethnographic heritage object provenance. In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, pages 127–143. IOS Press, 2023. Accessed on 2024-07-09. URL: https://ebooks.iospress.nl/doi/10.3233/SSW230010, doi:10.3233/SSW230010.

[120] Agnès Simon, Daniel Vila Suero, Eero Hyvönen, Esther Guggenheim, Lars G Svensson, Nuno Freire, Rainer Simon, Rodolphe Bailly, Roxanne Wyns, Seth van Hooland, Shenghui Wang, Vladimir Alexiev, Juliane Stiller, Antoine Isaac, and Vivien Petras. EuropeanaTech task force on a multilingual and semantic enrichment strategy: final report, 2014. Accessed on 2024-07-08. URL: https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/MultilingualSemanticEnrichment/Multilingual%20Semantic%20Enrichment%20report.pdf.

[121] Karen Smith-Yoshimura. Analysis of 2018 international linked data survey for implementers. (42), 2018. Accessed on 2024-07-08. URL: https://journal.code4lib.org/articles/13867.

[122] Juliane Stiller, Vivien Petras, Maria Gäde, and Antoine Isaac. Automatic enrichments with controlled vocabularies in europeana: Challenges and consequences. In Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, Lecture Notes in Computer Science, pages 238–247. Springer International Publishing, 2014. Accessed on 2024-07-08. doi:10.1007/978-3-319-13695-0_23.

[123] Go Sugimoto. Who is open data for and why could it be hard to use it in the digital humanities? federated application programming interfaces for interdisciplinary research.

12(4):204, 2017. Accessed on 2018-10-02. URL: http://www.inderscience.com/link.php?id=10014806, doi:10.1504/IJMSO.2017.10014806.

[124] Go Sugimoto. Building linked open date entities for historical research. In Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, editors, *Metadata and Semantic Research*, Communications in Computer and Information Science, pages 323–335. Springer International Publishing, 2021. Accessed on 2024-07-08. doi:10.1007/978-3-030-71903-6_30.

[125] Go Sugimoto. Instance level analysis on linked open data connectivity for cultural heritage entity linking and data integration. *Semantic Web*, 14(1):55–100, November 2022. Accessed on 2024-07-08. doi:10.3233/SW-223026.

[126] Go Sugimoto. User evaluation results for a wikidata-centric tool for temporal data in humanities and cultural heritage (june 2024): raw tabular data for two questionnaires from five online focus group workshops, 2024. Accessed on 2025-02-02. URL: https://zenodo.org/records/12693161, doi:10.5281/zenodo.12693161.

[127] Go Sugimoto, Angel Daza, and Victor de Boer. Closer reading of RDF generated by NLP on wikipedia biography: Comparative analysis. In Emmanouel Garoufallou and Fabio Sartori, editors, *Metadata and Semantic Research*, pages 41–54. Springer Nature Switzerland, 2024. Accessed on 2025-02-02. doi:10.1007/978-3-031-65990-4_4.

[128] Karim Tharani. Much more than a mere technology: A systematic review of wikidata in libraries. 47(2):102326, 2021. Accessed on 2024-07-08. URL: https://www.sciencedirect.com/science/article/pii/S0099133321000173, doi:10.1016/j.acalib.2021.102326.

[129] Dominik Tomaszuk and David Hyland-Wood. RDF 1.1: Knowledge representation and data integration language for the web. 12:84, 2020. Accessed on 2024-07-08. doi:10.3390/sym12010084.

[130] Houcemeddine Turki, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Lane Rasberry, and Daniel Mietchen. Ten years of wikidata: A bibliometric study. 2023. Accessed on 2024-07-08. https://wikidataworkshop.github.io/2023/.

[131] Jürgen Umbrich, Aidan Hogan, Axel Polleres, and Stefan Decker. Link traversal querying for a diverse web of data. 6(6):585–624, 2015.

[132] S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. 30(2):262–279, 2015. Accessed on 2021-01-24. URL: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqt067, doi:10.1093/llc/fqt067.

[133] Annelies van Nispen. Ehri vocabularies and linked open data: An enrichment? 106:117–122, 2019. Accessed on 2020-03-13. URL: https://hal.archives-ouvertes.fr/hal-02125036.

[134] Theo van Veen, Juliette Lonij, and Willem Jan Faber. Linking named entities in dutch historical newspapers. In Emmanouel Garoufallou, Imma Subirats Coll, Armando Stellato, and Jane Greenberg, editors, *Metadata and Semantics Research*, Communications in Computer and Information Science, pages 205–210. Springer International Publishing, 2016. Accessed on 2024-07-08. doi:10.1007/978-3-319-49157-8_18.

[135] Guillermo Vega-Gorgojo. LOD4culture. Accessed on 2024-07-08. URL: https://lod4culture.gsic.uva.es/.

[136] VIGNERON. Content on wikidata, 2024. Accessed on 2024-07-08. URL: https://commons.wikimedia.org/wiki/File:Wikidata_content_2024.svg.

[137] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009*, volume 5823, pages 650–665. Springer Berlin Heidelberg, 2009. Accessed on 2019-03-29. URL: http://link.springer.com/10.1007/978-3-642-04930-9_41, doi:10.1007/978-3-642-04930-9_41.

[138] Andra Waagmeester, Egon Willighagen, Nuria Queralt Rosinach, Elvira Mitraka, Sebastian Burgstaller-Muehlbacher, Tim E Putman, Julia Turner, Lynn M Schriml, Paul Pavlidis, Andrew I Su, and Benjamin M Good. Linking wikidata to the rest of the semantic web. page 2, 2017.

[139] Mitchell Whitelaw. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, Vol 9(1), 2015. Accessed on 2024-07-08. URL: https://researchsystem.canberra.edu.au/ws/portalfiles/portal/8448811/1296682.pdf.

[140] Roelf J. Wieringa and Johannes M. G. Heerkens. The methodological soundness of requirements engineering papers: a conceptual framework and two case studies. 11(4):295–307, 2006. Accessed on 2024-06-16. URL: https://research.utwente.nl/en/publications/the-methodological-soundness-of-requirements-engineering-papers-a, doi:10.1007/s00766-006-0037-6.

[141] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. 25:2311–2330, 2019. Accessed on 2024-07-08. doi:10.1109/TVCG.2018.2830759.

[142] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey: A systematic literature review and conceptual framework. 7(1):63–93, 2015. Accessed on 2019-04-09. URL: http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-150175, doi:10.3233/SW-150175.

[143] Marcia Lei Zeng. Semantic enrichment for enhancing LAM data and supporting digital humanities. review article. 28(1), 2019. Accessed on 2020-03-12. URL: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2019.ene.03, doi:10.3145/epi.2019.ene.03.

[144] Fudie Zhao. A systematic review of wikidata in digital humanities projects. 38(2):852–874, 2023. Accessed on 2024-06-24. URL: https://dx.doi.org/10.1093/llc/fqac083, doi:10.1093/llc/fqac083.

[145] Qing Zou and Eun G. Park. Modelling ancient chinese time ontology. 37(3):332–341, 2011. Accessed on 2020-10-08. URL: http://journals.sagepub.com/doi/10.1177/0165551511406063, doi:10.1177/0165551511406063.

SUMMARY

Tim Berners-Lee, the inventor of the World Wide Web (WWW), conceived the concept of the Semantic Web, or the Web of Data, as early as 1989. Linked Data (LD) principles were proposed as a building block for this approach. Linked Open Data (LOD) is a term used for LD published under an open license. Despite significant progress in LD, LD quality issues have been identified in the research community. We focus on LD quality in Cultural Heritage (CH) and Digital Humanities (DH). The primary Research Question of this thesis is as follows: "How can the quality of data and tools for Linked Data (LD) in Cultural Heritage and Digital Humanities be enhanced?"

Chapter 2 investigates the quality of LD data frequently used in cultural heritage. It addresses RQ1: "How is the quality of Linked Data instances in Cultural Heritage, particularly in terms of their connectivity?". The main stakeholders are data producers and data consumers. We analysed the quality of the eleven major LOD sources used for NEL in cultural heritage. We performed qualitative analysis on instance-level connectivity and graph traversals. Our outcomes suggest that a very limited number of links are found for major LOD datasets, with the exception of DBpedia. As a result, the LOD graph is centrally condensed and not fully interconnected. The centrality can be observed for not only linkages, but also for data content. Quantity and quality are unbalanced in 11 sources. These findings result in identified challenges for automatically identifying, accessing, and integrating known and unknown datasets. This implies the need for LOD improvement, as well as improving the NEL strategies to maximize data integration.

Chapter 3 mostly focusses on data quality. We address RQ2: "How can Linked Data connectivity for date entities be improved?" In historical research, time plays a crucial role in many investigations. However, discussions on the diversity of time information in LD have been limited in the community. To improve this situation, we built an RDF model and lookup service based on SKOSMOS for numeric dates at the lowest granularity level of a single day at a specific point in time, for the duration of 6000 years. The project, Linked Open Date Entities (LODE), generated stable URIs for more than 2.2 million entities, which include essential information and links to other LOD resources. It is based on the Wikidata data model and Time Ontology in OWL to facilitate entity linking by providing a detailed level of numeric time information in LD. The value of the date entities is discussed in two use cases with existing datasets. The first use case demonstrates the possibility to consolidate information from two calendar systems (Japanese and Western calendars). The second use case connects resources via LODE (Europeana/Semium.org, YAGO, Wikidata (Wikipedia)). We described how LODE can facilitate improved access and connectivity to unlock the potential for data integration in interdisciplinary research.

In Chapter 4, we address RQ3: "What are the quality gaps in biographical information between Wikipedia and Linked Data, and how can Information Extraction on Wikipedia be used to address them?"

Wikidata and DBpedia are valuable LD resources closely related to Wikipedia. However, they often hold a small subset of its semantic information due to the specific scopes and methodologies chosen for their LD construction. To fill this knowledge gap, we deployed Information Extraction (IE) used Natural Language Processing (NLP). We aimed to assess to what extent out-of-the-box NLP tools can semiautomatically generate new LD from biographical articles in Wikipedia. We evaluated the overlaps and gaps between Wikipedia, Wikidata, and DBpedia, as well as other biographical ontologies. We analysed the triple patterns from the NLP results in comparison with the RDF entity (instance) and ontologies. Our research revealed that we are able to capture new information about the entity that Wikidata and DBpedia do not hold. At the same time, our approach was not particularly suited to generate the same LD as Wikdiata and DBpedia. In addition, it turned out that some noise cannot be easily eliminated. Our method also presented the possibility for a bottom-up approach to design a biographical ontology from textual data.

In Chapter 5, we address RQ4: "What are the effective designs and functionalities for Linked Data tools to support research using temporal information in Cultural Heritage and Digital Humanities?" We investigated LD visualization tools that address event data in association with temporal data for research purposes in DH and CH. We analyzed the availability of Wikidata visualization tools capable of handling event data for DH and CH research and proposed ways to improve them to better represent temporal data in Wikidata. We identified 14 requirements based on principles from the information visualization domain, as well as an analysis of previous studies and existing tools. We designed and developed a Wikidata-centric tool to meet these requirements. This tool was then evaluated through focus groups and questionnaires with DH and CH experts. The results of the evaluation showed overall positive feedback and highlight the implicit need for and value of visualization tools that handle events in Wikidata for research purposes in DH and CH. Additionally, they indicated the improved accessibility and visualization capabilities of Wikidata through the tool's seven time-related functionalities.

Chapter 6 revisits the overall RQ: "How can the quality of data and tools for Linked Data in Cultural Heritage and Digital Humanities be enhanced?" During our research, it became clear that quality issues in LD are complex and multifaceted. Admittedly, there is no single solution to address them all. However, we are able to draw conclusions with respect to (a) research methodologies and (b) analysis and strategies. (a) We conclude that our strategy to emphasize more on qualitative analysis than quantitative analysis provides a more user-centric understanding of LD quality, compared to previous studies. This strategy helps to formulate solutions to improve the LD quality for as many stakeholders as possible. (b) We believe that there are two primary ways to enhance the quality: one is to enrich existing resources and the other is to create new high-quality resources. In particular, ten specific strategies are provided (five for improving data and additional five for improving the tools). Additionally, we shortlisted specific potential social activities for each stakeholder to enhance LD quality.

There are a few areas of the thesis that require further discussion and investigation. We faced the same issues that humanities scholars may confront: the dilemma of positioning themselves somewhere between the "traditional" close reading (qualitative) and "new" distant reading in DH (quantitative). In addition, the quality of highly interdisciplinary data integration is still largely unknown.

List of Publications

1. *Go Sugimoto.* Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data Integration, 1 Jan. 2023, the Semantic Web, 1, 55 - 100 (**Chapter 2**)

2. *Go Sugimoto.* Building linked open date entities for historical research. In Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, editors, Metadata and Semantic Research, Communications in Computer and Information Science, pages 323–335. Springer International Publishing, 2021. doi:10.1007/978-3-030-71903-6_30. (**Chapter 3**)

3. *Go Sugimoto, Angel Daza, and Victor de Boer.* Closer reading of RDF generated by NLP on wikipedia biography: Comparative analysis. In Emmanouel Garoufallou and Fabio Sartori, editors, Metadata and Semantic Research, pages 41–54. Springer Nature Switzerland, 2024. doi:10.1007/978-3-031-65990-4_4. (**Chapter 4**)

4. *Go Sugimoto, Victor de Boer, and Jacco van Ossenbruggen* Wikidata Visualization for Event and Temporal Data Exploration in Digital Humanities and Cultural Heritage. (Submitted for review in April 2025 (**Chapter 5**)

SIKS Dissertations

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017
01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs

	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
31 Gongjin Lan (VUA), Learning better – From Baby to Better
32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach

10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline