A Primer by the Open Data Policy Lab

Data Sandboxes:

Managing the Open Data Spectrum

Authors:
Uma Kalkar, Sampriti Saxena and Stefaan Verhulst

September 2023

SMIT

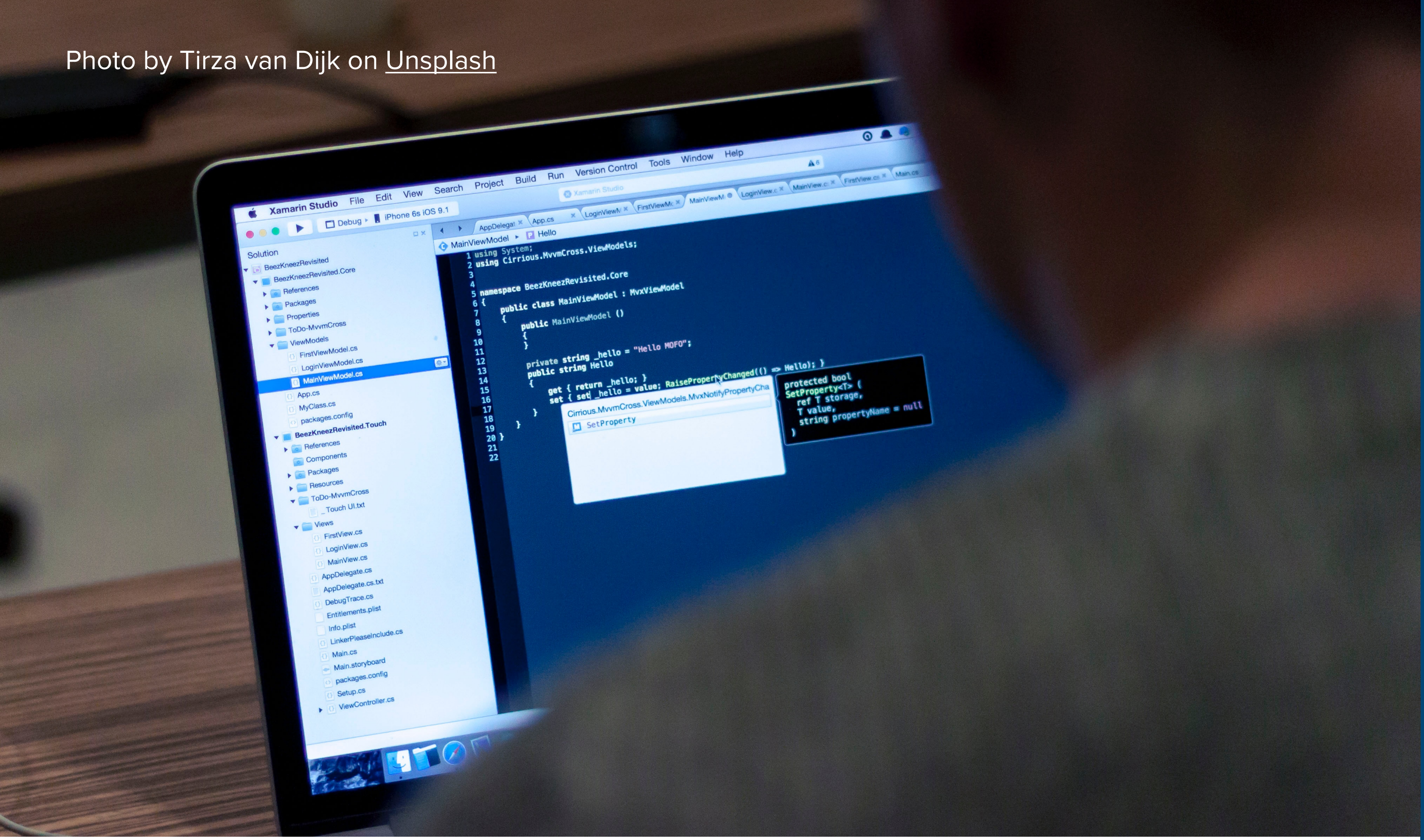
Vlaanderen
verbeelding werkt

VUB VRIJE UNIVERSITEIT BRUSSEL

imec

Open.Data.Policy.Lab

TABLE OF CONTENTS



```
using System;
using Cirrious.MvvmCross.ViewModels;


namespace BeezKneezRevisited.Core
{
    public class MainViewModel : MvxViewModel
    {
        public MainViewModel ()
        {
        }

        private string _hello = "Hello MOFO";
        public string Hello
        {
            get { return _hello; }
            set { set _hello = value; RaisePropertyChanged(() => Hello); }
        }
    }
}
```

```
protected bool
SetProperty<T> (
  ref T storage,
  T value,
  string propertyName = null
)
```

Photo by Tirza van Dijk on Unsplash

Cover photo by fabio on Unsplash

ACKNOWLEDGEMENTS

This work would not have been possible without the support and guidance of many individuals. In particular, we are grateful to Karl-Filip Coenegrachts, who provided us with the initial idea and was instrumental in securing the resources to make this primer possible. We would like to thank imec, SMIT, Vrije Universiteit Brussel and the Smart Flanders program of the Government of Flanders (Agency of Home Affairs, Urban Policy Team) for supporting our work on the subject. We are also grateful to our peer reviewers: Marjiana Novak (Circle Economy) and Justin Bryant (The Datasphere Initiative), for their valuable feedback, as well as all our colleagues at The GovLab for their review and input. Finally, the authors are extremely grateful to the following individuals for their generosity in sharing their time and valuable insights with us: Bart Rousseau (Digital Flanders), Laust Hvas Mortenson (DST's Data Science Lab), Juliana Outes Velarde (Government Outcomes Lab), and Ruth Meßner (EIT Health).

EXECUTIVE SUMMARY

Open data offers opportunities to enhance governance, elevate public services, empower individuals, and bolster public well-being. However, achieving the delicate balance between open data access and the responsible use of sensitive and valuable information presents complex challenges. Data sandboxes are an emerging approach to balancing these needs.

In this white paper, we seek to answer the following questions surrounding data sandboxes: **What are data sandboxes? How can data sandboxes empower decision-makers to unlock the potential of open data while maintaining the necessary safeguards for data privacy and security? Can data sandboxes help decision-makers overcome barriers to data access and promote purposeful, informed data (re-)use?** By investigating these questions, we seek to help policymakers consider the sandbox methodology's potential to advance data collaboration between multiple stakeholders, evaluate the social value of opening data, and better articulate the running costs of data-driven initiatives.

The GovLab defines a data sandbox as *a temporary, controlled, and secure virtual space where different actors can access, process, (re-)use, and analyze different types of data (often with the intent to determine how to open data, if at all).*

By evaluating a series of case studies, we identified the following key findings:

- Data sandboxes present six unique characteristics that make them a strong tool for facilitating open data and data re-use. These six characteristics are: controlled, secure, multi-sectoral and collaborative, high computing environments, temporal in nature, adaptable, and scalable.

- Data sandboxes can be used for: pre-engagement assessment, data mesh enablement, rapid prototyping, familiarization, quality and privacy assurance, experimentation and ideation, white labeling and minimization, and maturing data insights.

- There are many benefits to implementing data sandboxes. We found ten value propositions, such as: decreasing risk in accessing more sensitive data; enhancing data capacity; and fostering greater experimentation and innovation, to name a few.

- When looking to implement a data sandbox, decision-makers should consider how they will attract and obtain high-quality, relevant data, keep the data fresh for accurate re-use, manage risks of data (re-)use, and translate and scale up sandbox solutions in real markets.

- Advances in the use of the Internet of Things and Privacy Enhancing Technologies could help improve the creation, preparation, analysis, and security of data in a data sandbox. The development of these technologies, in parallel with European legislative measures such as the Digital Markets Act, the Data Act and the Data Governance Act, can improve the way data is unlocked in a data sandbox, improving trust and encouraging data (re-)use initiatives.

When it comes to implementing and governing data sandboxes, we offer nine operational considerations to carry out when setting up a data sandbox, including:

- Appoint a Data Sandbox Facilitator;
- Define a clear purpose and problem;
- Identify the necessary skills and expertise, along with roles and responsibilities for stakeholders;
- Establish data and technological infrastructure requirements;
- Establish the role of a data steward;
- Categorize the sandbox's data assets using a collaborative methodology;
- Provide strong metadata about the data assets available;
- Define data access and use parameters; and
- Building dispute resolution processes to proactively handle conflict.

Further, we offer 10 principles to consider when designing a data sandbox, as well as a series of processes and practices that can operationalize these governance requirements.

We conclude that data sandboxes offer a secure and controlled environment for data access and exploration that provides a low-risk and high-reward way of trialing data sharing and (re-)use. This model of data collaboration can help promote data openness among new players, improving the availability and access to data to address public problems.

INTRODUCTION

The transformative potential of open data is immense. Open data offers opportunities to enhance governance, elevate public services, empower individuals, and bolster public well-being.[1] However, achieving the delicate balance between open data access and the responsible use of sensitive and valuable information presents complex challenges. In some cases, a secure and controlled environment for data access and exploration is essential to assess the feasibility and implications of open data and data re-use.

Over the past few years, the popularity of data sandboxes to support data-driven innovation has grown. For example, in the U.S., 11 'Mobility on Demand' sandboxes were piloted for government officials and private sector players to assess and improve smartphone applications and integrated fare payments for smarter public transportation.[2] In Europe, the European Institute of Innovation & Technology launched a Digital Sandbox for health and medical technology companies to open and simplify access to biobanks and registries to leverage public health data for innovation.[3]

The central issue that data sandboxes aim to address is the need to reconcile the competing demands of open data access and the protection of sensitive information. They seek to provide a safe, controlled space where decision-makers can responsibly explore, de-risk, share, and re-use data to evaluate the potential as well as the risks of broader data access.

Important questions remain to be answered. **How can data sandboxes empower decision-makers to unlock the potential of open data while maintaining the necessary safeguards for data privacy and security? Can data sandboxes help decision-makers overcome barriers to data access and promote purposeful, informed data (re-)use?**

[1] Young, Andrew, and Stefaan Verhulst. "The Global Impact of Open Data." *O'Reilly Media*, September 2016. https://odimpact.org/files/open-data-impact-key-findings.pdf; Young, Andrew, and Stefaan Verhulst. "Open Data Impact: When Demand and Supply Meet: Key Findings of the Open Data Impact Case Studies." *Open Data Impact, March 2016*. https://odimpact.org/files/open-data-impact-key-findings.pdf

[2] Patel, Ronik Ketankumar, Roya Etminani-Ghasrodashti, Sharareh Kermanshachi, Jay Michael Rosenberger, and Ann Foss. "Mobility-on-demand (MOD) Projects: A study of the best practices adopted in United States." *Transportation Research Interdisciplinary Perspectives* 14 (2022): 100601. https://doi.org/10.1016/j.trip.2022.100601

[3] "The Digital Sandbox Accelerator." *EIT*, February 2020. https://eit.europa.eu/our-activities/opportunities/digital-sandbox-accelerator

This white paper seeks to help policymakers consider the sandbox methodology's potential to advance data collaboration between multiple stakeholders, evaluate the social value of opening data, and better articulate the running costs of data-driven initiatives.

1. WHAT ARE DATA SANDBOXES?

1.1. Definition of a Data Sandbox

In this white paper, we focus on **data sandboxes**, which The GovLab defines as *a temporary, controlled, and secure virtual space where different actors can access, process, (re-)use, and analyze different types of data (often with the intent to determine how to open data, if at all).*

The sandbox methodology is used for research and commercial purposes across various industries. As defined above, a data sandbox provides a safe space to assess and experiment with new ideas involving data, including emerging technologies, data applications, and governance techniques. While most data sandboxes are virtual, some physical laboratories exist,[4] and typically allow data access for a short and fixed period of time.[5] Sandboxes share two common elements: their *temporality*—sandboxes are quick sprints to assess or create data products—and their *collaboration*—they engage different actors to co-design models and solutions.[6]

In addition to data sandboxes, three other types of sandboxes are often cited in the emergent literature (and often confused or combined with each other):

- **Regulatory or policy sandboxes**, which provide waivers from existing regulation so organizations can run trials of new products and policymakers can develop iterative regulations on how data and technology is handled.[7]

[4] "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions: A New European Innovation Agenda." *European Commission*, July 5, 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022DC0332

[5] Allen, Hilary J. "Sandbox boundaries." *Articles in Law Reviews & Other Academic Journals*, vol. 694, (2020). https://digitalcommons.wcl.american.edu/facsch_lawrev/694

[6] Datasphere Initiative. "Sandboxes for data: creating spaces for agile solutions across borders." *Datasphere Initiative*, 2022. https://www.thedatasphere.org/wp-content/uploads/2022/05/Sandboxes-for-data-2022-Datasphere-Initiative.pdf

[7] Arntzen, Siri, Zach Wilcox, Neil Lee, Catherine Hadfield, and Jen Rae. "Testing Innovation in the Real World: Real-world testbeds." *Nesta*, October 2019. https://media.nesta.org.uk/documents/Testing_innovation_in_the_real_world.pdf; Datasphere Initiative. "Sandboxes for data: creating spaces for agile solutions across borders." *Datasphere Initiative*, 2022. https://www.thedatasphere.org/wp-content/uploads/2022/05/Sandboxes-for-data-2022-Datasphere-Initiative.pdf

- **Innovation sandboxes,** which provide a virtual space for developers to assess and test the features of an innovation from the perspective of prospective users.[8]
- **Software development sandboxes,** which allow for bespoke, secure, and realistic product demonstrations to trial new digital tools, test potentially dangerous applications, and assess user product experiences.[9]

1.2. Characteristics of Data Sandboxes

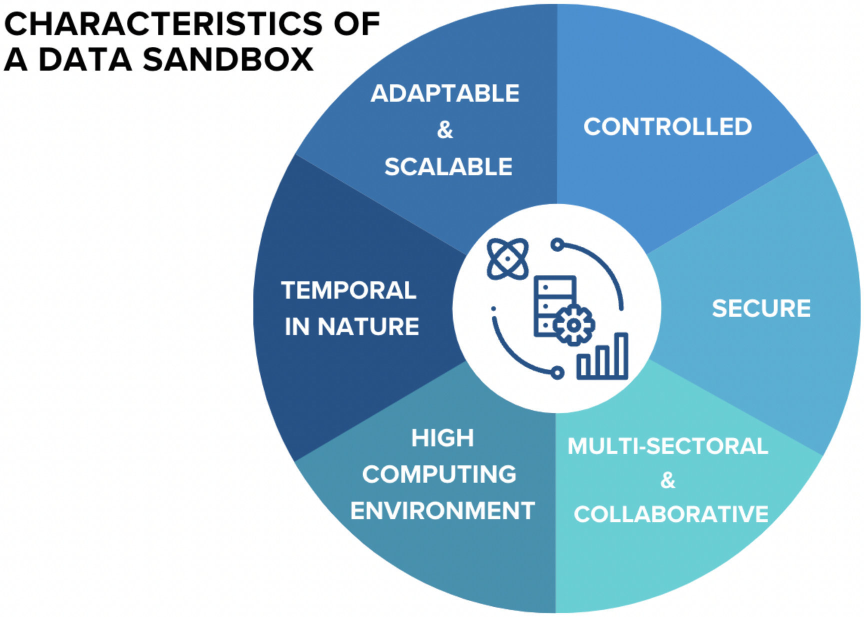


Figure 1. The six characteristics of a data sandbox. Image by The GovLab.

Data sandboxes possess a unique set of features that make them an ideal solution for addressing the challenges associated with open data exploration and re-use. These characteristics include being:

1. **Controlled:** Data sandboxes provide a restricted setting where sensitive data can be securely accessed, analyzed, and manipulated, ensuring compliance with relevant regulations.[10] Further, users are vetted prior to access, thereby

[8] Ribiere, Vincent M., and Francis D. Tuggle. "Fostering innovation with KM 2.0." *Vine* 40, no. 1 (2010): 90-101. https://doi.org/10.1108/03055721011024955

[9] "Sandbox Software." *SourceForge*, n.d. https://sourceforge.net/software/sandbox/

[10] Ibid.

opening the data up to new sectors and actors while maintaining a level of oversight on who uses it. These access control measures can include:

- Requiring data users to submit an application detailing their project and rationale to join the sandbox;[11]
- Vetting and verifying researcher credentials for legitimate use of data; and
- Hosting mandatory data research and literacy training classes to ensure proper data use.

2. **Secure:** The data available in the sandbox does not infringe on data subjects' consent approvals or rights to privacy. The way the data is stored, presented, and handled in the sandbox is protected with data and digital security measures to prevent unauthorized use or manipulation. For example, the data can (but does not necessarily need to) feature some of the following elements:[12]
    - **Aggregated:** individual data is often gathered and presented as a summary; raw data points are not necessarily provided.
    - **Anonymized or De-identified:** the data has no personally identifiable information (PII) and cannot be linked back to the data subjects.
    - **Access controlled:** data that is only accessible to certain individuals for specific use cases.
    - **Synthetic:** artificially generated data that is created by machines rather than collected from human data subjects.

3. **Multi-Sectoral and collaborative:** These platforms enable stakeholders from various sectors, including governments **(G)**, businesses **(B)**, and academia and civil society **(C)**, to work together on shared goals and objectives, fostering innovation and synergies. [13] This interdisciplinary flow of data and insights brings together players in data innovation and regulation, including national statistical and innovation offices and established and startup companies. These stakeholders may include:[14]

[11] "The Digital Sandbox Accelerator." *EIT*, February 2020. https://eit.europa.eu/our-activities/opportunities/digital-sandbox-accelerator

[12] Young, Andrew, Stefaan Verhulst, Nadiya Safonova, and Andrew J. Zahuranec. "The Data Assembly: Responsible Data Re-Use Framework." *The GovLab*, November 2020. https://thedataassembly.org/files/nyc-data-assembly-report.pdf

[13] Ribiere, Vincent M., and Francis D. Tuggle. "Fostering innovation with KM 2.0." *Vine* 40, no. 1 (2010): 90-101. https://doi.org/10.1108/03055721011024955

[14] Kalkar, Uma, and Natalia González Alarcón. "Facilitating Data Flows through Data Collaboratives: A Practical Guide to Designing Valuable, Accessible, and Responsible Data Collaboratives." *The GovLab*, forthcoming.

- **Data holders:** organizations or individuals that possess and control access to data relevant to the sandbox. A data holder can be a government agency, a private sector firm, or a community member.
- **Data users:** organizations or individuals that access and (re-)use the data that has been made available in the sandbox. Data users may include governments, businesses, non-profit organizations, or researchers.
- **Data stewards:** organizational leaders or teams with data and data science skills who are responsible for re-using their organization's data to create public value by identifying opportunities for innovative collaboration and responding proactively to external requests for access to data, insights, or expertise.[15] Data stewards are critical players when it comes to coordinating collaborations, often acting as liaisons between data holders and data users.
- **Data engineers:** individuals with the technical data science skills needed to leverage analysis tools, such as Tableau or PowerBI, or create bespoke data dashboards to visualize data.
- **Data decision-makers:** individuals who oversee the creation of governance systems, including regulatory and legal means, around data (re-)use. They coordinate with other stakeholders to update and advance data governance.

In smaller organizations, it is common for these roles to be consolidated due to resource constraints and the necessity for increased efficiency.

[15] Verhulst, Stefaan, Andrew J. Zahuranec, Andrew Young, and Michelle Winowatan. "Wanted: Data Stewards: (Re-)Defining the Roles and Responsibilities of Data Stewards for an Age of Data Collaboration. *The GovLab*, March 2020. https://thegovlab.org/static/files/publications/wanted-data-stewards.pdf

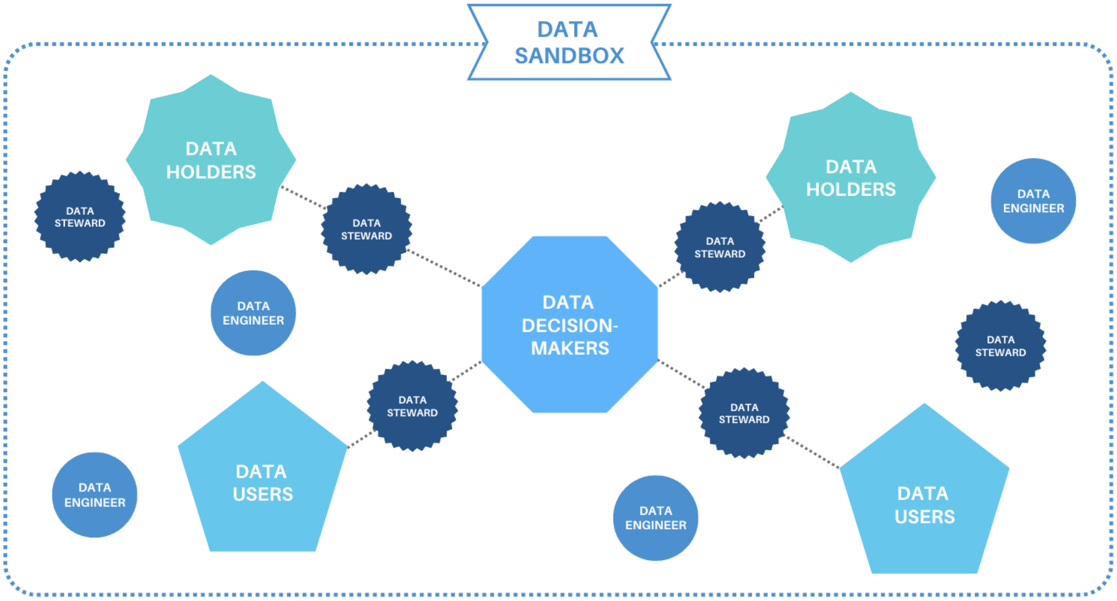


Figure 2. The various stakeholders of a data sandbox. Image by The GovLab.

4. **High computing environment:** Data sandboxes may "provide the computing required for data scientists to tackle typically complex analytical workloads."[16] This can include technical elements, including:[17]
   - Large, sophisticated parallel central processing units;
   - High-end memory;
   - Massive volume of storage area and quick processing times; and
   - Separate areas for handling data and experimenting with data.

5. **Temporal in nature:** Sandboxes typically operate within a limited timeframe, with a focus on addressing specific challenges or exploring particular data sets.[18] This temporal aspect allows for iterative learning, adaptation, and improvement as new insights and discoveries emerge.

6. **Adaptable and scalable:** The flexibility of data sandboxes enables them to be tailored to various data types, use cases, sectors, and contexts, ensuring their relevance and usefulness across diverse scenarios.

[16] Rouse, Margaret. "What Does a Data Sandbox Mean?" *Techopedia*, May 12, 2020. https://www.techopedia.com/definition/28966/data-sandbox-big-data

[17] Ibid.

[18] Allen, Hilary J. "Sandbox boundaries." *Articles in Law Reviews & Other Academic Journals*, vol. 694, (2020). https://digitalcommons.wcl.american.edu/facsch_lawrev/694

1.3. Purposes of a Data Sandbox

A data sandbox can serve various purposes toward opening data and establishing data collaboration. These can include, but are not limited to the following:

1. **Pre-engagement assessment:** In a data sandbox, data holders and data users can evaluate the relevance and value of data before formalizing data-sharing agreements. This ensures the data meets the necessary requirements before it is fully shared, improving the efficiency of and staying true to an initiative's purpose for both parties.

2. **Data mesh enablement:** A data sandbox can map the creation of a "data mesh," a decentralized architecture that organizes data into domains to enable accessible and targeted use of data.[19] Creating this structure allows for seamless data linking and integration.

3. **Rapid prototyping:** A data sandbox can facilitate the development of data products by collaborating with relevant stakeholders for faster and more efficient results.

4. **Familiarization:** A data sandbox can assist organizations that are new to data sharing practices in becoming comfortable and adept at opening up their data resources. This provides a taste of what opening and sharing data entails without requiring organizations to fully commit to such practices.

5. **Quality and privacy assurance:** A data sandbox can enable external validation of the data being shared to make sure it meets the required quality standards and adheres to privacy regulations before it is shared further.

6. **Experimentation and ideation:** A data sandbox can encourage innovative uses of data by exploring potential applications and re-purposing opportunities, including the co-development of proprietary datasets for example.

7. **White labeling and minimization:** A data sandbox can support the process of anonymizing training data and determining the proportional scope of data sharing.

[19] "What is a Data Mesh?" *IBM*, n.d. https://www.ibm.com/topics/data-mesh; Rousseau, Bart. Interview re: Digital Flanders' Digital Sandbox. In-person, April 19, 2023.

8. **Maturing data insights:** Being able to rapidly prototype and experiment with different models and algorithms, and explore the quality and potential of different data sets, encourages innovation and the maturing of new insights.

2. EXAMPLES AND TYPOLOGY OF DATA SANDBOXES

Data sandboxes are being successfully implemented across Europe in diverse settings and contexts. Below, we explore six data sandbox cases taking place at a city, country, and supranational level from across the UK and the European Union that cater to data sharing between governments, businesses, and academic institutions. A comparative review of the examples is provided in **Appendix B**.

2.1. Curation of Data Sandbox Examples

2.1.1. London's Digital Sandbox Pilots

London's Digital Sandbox pilot is the product of a collaboration between the City of London Corporation and the Financial Conduct Authority. [20] The two government organizations initially partnered in 2020 to create a sandbox focused on meeting open data needs in light of the COVID-19 pandemic.[21] Since then, the program's area of focus has expanded to developing and testing emerging technologies, sustainability, and greener solutions for the financial sector in support of the national government's Net Zero aspirations.

Two pilot programs have been run under the initiative. The first sandbox ran from October 2020 to February 2021 and focused on improving financial resiliency and fraud protection. The second sandbox ran between November 2021 and March 2022 and looked at how technology could promote ESG-friendly business practices.[22] For each sandbox, three use cases were identified by hosting workshops with stakeholders to identify the priority issues that could benefit from data innovation.[23]

[20] "Digital Sandbox Pilot." *City of London*, n.d. https://www.cityoflondon.gov.uk/supporting-businesses/business-support-and-advice/digital-sandbox-pilot

[21] "Digital Sandbox." *Financial Conduct Authority*, April 14, 2023. https://www.fca.org.uk/firms/innovation/digital-sandbox

[22] Ibid.

[23] "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

Access to the sandbox is controlled. Firms applied to take part in the sandbox, specifically detailing how the use of sandbox data and digital technologies (such as Privacy Enhancing Technologies (PETs), big data, machine learning, etc.) could lead to genuine innovation.[24] Some of the organizations operating in the sandbox include Vested Impact[25] (using ESG data to measure impact), Greenomy[26] (mapping ESG standards to drive adoption), and QPQ[27] (enabling the digital economy through deep tech solutions).

The data assets provided in the sandbox ranged from consumer transactions and financial data, to SME lending data, business production data, satellite data, and global indicators. This data was a mixture of real and synthetic data; for instance, highly sensitive data, such as financial and business data, was invented. Data was provided at an aggregate and individual level.

Both sandboxes featured data from government and business data holders.[28] Data users were able to collaborate, test and showcase new solutions using real and synthetic datasets via a shared platform for collaboration.[29] In addition to the datasets, the sandbox featured: application programming interfaces (APIs) designed to support the development and validation of solutions; an integrated development environment (IDE) hosted on Jupyter Notebooks for data users to collaboratively write code in; a collaboration platform for sandbox players to discuss their work with expert stakeholders who can provide sector-specific knowledge, specialist skills, and mentorship; and an 'observation deck' to demo products.[30]

Results from both sandboxes show that the pilot program did at least partially meet expectations of enhancing innovation, collaboration, testing, and rapid prototyping of products.[31] Participants noted that the sandboxes furthered collaboration, but further

[24] "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

[25] *Vested Impact*. https://www.vestedimpact.co.uk/

[26] *Greenomy*. https://greenomy.io/

[27] *QPQ Technology*. https://qpq.io/technology/

[28] "Digital Sandbox." *Financial Conduct Authority*, April 14, 2023. https://www.fca.org.uk/firms/innovation/digital-sandbox

[29] Ibid.

[30] "Digital Sandbox." *Financial Conduct Authority*, April 14, 2023. https://www.fca.org.uk/firms/innovation/digital-sandbox; "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

[31] "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

access to relevant datasets and data tools and more investment in validating proof-of-concept and business models was needed.[32] Further, both cohorts of participants noted that it remains too early to see the full impact of these pilots on product and policy action. The Kalifa Review of UK FinTech, an independent policy report on the state of the sector, advocated for the permanent implementation of the Digital Sandbox given its positive impact on the sector as a whole and its potential to drive future innovation.[33] Now, the City of London Corporation and the Financial Conduct Authority are exploring establishing a permanent operating model based on the pilot.[34]

2.1.2. Statistics Denmark's Data Science Lab

In 2022, Statistics Denmark launched their new Data Science Lab to offer researchers access and experience working with "more sensitive and new weird data."[35] Housed under the government agency, this sandbox seeks to streamline access to sensitive administrative data while fostering innovation and open science ideals.

Participation in the sandbox is determined by a strict application process. Researchers are required to submit a proposal for consideration, outlining their research process and goals, as well as the population of data they would like to access. The application also requests the details of all potential data users for the project. To be considered for the sandbox, applicants must be affiliated with an authorized academic, public, or private institution.[36] Uniquely, Statistics Denmark charges researchers for access to the sandbox, quoting a price based on the scope of the project. This not only creates a barrier to entry to encourage meaningful participation but also creates an additional revenue stream and incentive for Statistics Denmark to open their data up for (re-)use.[37]

[32] "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

[33] "The Kalifa Review of UK FinTech." *United Kingdom Government*, February 26, 2021. https://www.gov.uk/government/publications/the-kalifa-review-of-uk-fintech

[34] Ibid.

[35] Tranberg, Pernille. "Statistic Denmark's New Sandbox: The Data Science Lab." *Data for Good Science*, June 14, 2022. https://www.dataforgood.science/statistic-denmarks-new-sandbox-the-data-science-lab/

[36] "Denmark's Data Portal." *Statistics Denmark*, n.d. https://www.dst.dk/en/TilSalg/Forskningsservice/danmarks-datavindue

[37] Ibid.

Prior to the sandbox, researchers were only able to use non-sensitive data through Statistics Denmark's Research Service arm.[38] Now, the Data Science Lab sandbox provides access to further non-sensitive data and previously locked sensitive data. It also allows researchers to experiment with different data analysis tools innovatively, such as with Geographic Information System (GIS) or High-Performance Computing (HPC) methods. The data is housed under a secure firewall and access to it is given via a remote workspace.[39] As well, the sandbox is held to Statistics Denmark's policies for where and how data can be used. For instance, foreign research and analysis of Danish data is not authorized, however, under certain conditions, foreign users can access data.[40]

The Data Science Lab sandbox looks to produce academic research and datasets that can advise on new statistical and data-driven analysis methods. To this end, it hosts all of Novo Nordisk Foundation's Challenge projects and EU-funded projects, providing researchers with high-value, high-demand initiatives to work on. The Data Science Lab seeks to share findings and insights generated within the sandbox for public research.[41]

2.1.3. Lithuania's Traveltech Data Sandbox

Lithuania is fast on its way to becoming a country of sandboxes.[42] The small Baltic country appears to be leading the adoption of the tool, playing host to sandboxes for solutions in the property technology (PropTech),[43] financial technology (FinTech),[44] government technology (GovTech)[45] and tourism technology (TravelTech)[46] sectors, to name just a few. Thanks to its adaptable public sector, a thriving technology industry,

[38] Tranberg, Pernille. "Statistic Denmark's New Sandbox: The Data Science Lab." *Data for Good Science*, June 14, 2022. https://www.dataforgood.science/statistic-denmarks-new-sandbox-the-data-science-lab/

[39] Mortenson, Laust Hvas. 2023. "RE: Request for some information for The GovLab's Sandboxes White Paper Project," May 3, 2023.

[40] "Access to data." *Statistics Denmark*, n.d. https://www.dst.dk/en/TilSalg/Forskningsservice/Dataadgang

[41] Ibid.

[42] "Country as a sandbox: where startups and public institutions come to play." *Government of Lithuania*, October 29, 2021. https://lithuania.lt/governance-in-lithuania/country-as-a-sandbox-where-startups-and-public-institutions-come-to-play/

[43] Turp-Balazs, Craig. "World's First PropTech Sandbox Opens in Lithuania." *Emerging Europe*, March 15, 2018. https://emerging-europe.com/news/worlds-first-proptech-sandbox-opens-lithuania/

[44] "Regulatory Sandbox." *Bank of Lithuania*, March 23, 2020. https://www.lb.lt/en/regulatory-sandbox

[45] "Lithuania Has Launched a Funded Govtech Sandbox Program." *GovTech Lab Lithuania*, n.d. https://govtechlab.lt/lithuania-has-launched-a-funded-govtech-sandbox-programme/

[46] "Lithuania TravelTech Sandbox." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/news/lithuania-traveltech-sandbox

and a robust academic ecosystem, along with a culture of innovation and entrepreneurship, Lithuania is able to create regulatory, policy, and data sandboxes.

Lithuania's TravelTech Sandbox is an example of a sandbox for open data.[47] It is the result of a collaboration between Lithuania Travel, the national tourism development agency, and three government-run economic development programs: Invest Lithuania, the Innovation Agency of Lithuania, and Startup Lithuania.[48]

This sandbox provides free and open access to tourism data, including data on mobility, traffic, expenses, and popular sites.[49] This real, aggregated, and anonymous data is organized as a data dashboard with annually aggregated tourism statistics that can help improve business and labor force development in Lithuania.[50] The sandbox also provides stakeholders with the financial mechanisms, access to regulatory institutions, and strong infrastructure support to help ease the productization of their sandbox solutions.[51] At present, major TravelTech companies like the ride-share app Bolt, the reservation website Booking.com, and the transportation rental site Kayak are among the many players using Lithuania's TravelTech Sandbox to improve their data-driven services.

2.1.4. The E.I.T. Health Digital Sandbox

EIT Health[52], an offshoot of the European Institute of Innovation and Technology, established its Digital Sandbox from 2020 to 2021 with the goal of empowering European start-ups in their access and use of valuable medical data. The sandbox facilitated access to data from European biobanks, registers, and sample holders, which the start-ups were able to access for a maximum of 12 months.[53] In addition to matching the start-ups with relevant datasets, the EIT Health Digital Sandbox also

[47] "Lithuania TravelTech Sandbox." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/news/lithuania-traveltech-sandbox

[48] "Tourism Lithuania." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/

[49] "Lithuania TravelTech Sandbox." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/news/lithuania-traveltech-sandbox

[50] "Tourism Lithuania." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/

[51] "Lithuania TravelTech Sandbox." *Lithuania Travel*, n.d. https://www.lithuania.travel/en/news/lithuania-traveltech-sandbox

[52] "Who We Are" *EIT Health*, n.d. https://eithealth.eu/who-we-are/

[53] "The Digital Sandbox Accelerator." *EIT*, February 2020. https://eit.europa.eu/our-activities/opportunities/digital-sandbox-accelerator

offered the startups a share of funding from a pool of €180,000 and helped them identify and partner with key researchers and experts in the field.[54]

Access to data in the Digital Sandbox was strictly controlled and limited to data sharing between the data holders and the start-ups. Although EIT Health fostered these partnerships between the actors in the sandbox, they were not directly involved in sharing the data and did not have access to any of the data exchanged between actors.[55] Rather, the Digital Sandbox served as a secure forum where trusted stakeholders could come together to form data sharing agreements.[56] In this type of a data sandbox, there were fewer requirements for data governance measures and there was a greater focus on vetting stakeholders to ensure that relevant and high quality data assets are matched with the right competencies and expertise to achieve actionable insights.

Following a competitive application process, EIT Health selected six companies, two each from the Biotech, Medtech, and Digital Health sectors, with unique and specific data needs to access real-world data. The companies were: Amadix (Spain), BeFC (France), Acobiom (France), GlyCardial Diagnostics (Spain), Healthinn: Rehand (Spain), and Radiobotics (Denmark).[57] Acobiom, for example, was working on developing biomarkers to diagnose and treat pancreatic cancer for which they required access to clinical and real-world data from different biobanks to test their results.[58]

2.1.5. The Indigo Initiative's Sandbox For Open Data

The Government Outcomes (GO) Lab, based at Oxford University's Blavatnik School of Government, studies collaborations between government organizations and organizations in the private and social sectors to drive stronger social outcomes.[59] As part of this work, the GO Lab is home to the International Network for Data on Impact and Government Outcomes (INDIGO), an initiative to promote data sharing across sectors.[60]

[54] "Digital Sandbox helps six start-ups access key data." *EIT Health*, September 7, 2020. https://eithealth.eu/news-article/digital-sandbox-helps-6-start-ups-access-key-data/

[55] Meißner, Ruth. 2023. "RE: Request for some information for The GovLab's Sandboxes White Paper Project," May 9, 2023.

[56] Ibid.

[57] "Digital Sandbox helps six start-ups access key data." *EIT Health*, September 7, 2020. https://eithealth.eu/news-article/digital-sandbox-helps-6-start-ups-access-key-data/

[58] Ibid.

[59] *GO Lab*. https://golab.bsg.ox.ac.uk/about/

[60] "INDIGO: Better data for better social outcomes." *GO Lab*, n.d. https://golab.bsg.ox.ac.uk/knowledge-bank/indigo/

The INDIGO initiative leverages sandboxes to promote the adoption of open data practices in the social outcomes sector.[61] They are able to do this by offering large-scale data holders the opportunity to trial sharing their data in a secure environment. Rather than engaging with multiple stakeholders in a shared space, the INDIGO initiative creates bilateral sandboxes with individual clients. Later on, to connect these bilateral data sharing initiatives, INDIGO's data stewards work to build a common governance structure for the sandboxes. The data stewards monitor overall access control, insight sharing, and responsible and sustainable data use across the sandboxes.

Data access is controlled and held in an isolated, password protected sandbox. Only those who need to work with the data directly are granted access to the data. As well, INDIGO labels data as 'only for sandbox' and which sandbox it belongs to.[62]

INDIGO's data stewards work with data holders to own and govern the responsible sharing of data within the safe and trusted sandbox environment.[63] The INDIGO team runs analyses with the data, creates visualizations, and presents a mock-up of how the data could be best packaged for open consumption. These processes help increase the data holders' comfort with openly sharing their data and further help them to understand the value proposition of re-using their data in driving stronger social outcomes and public good.

Once the data holder is convinced of the value and safety of open data practices, the INDIGO team moves the datasets out of the sandbox and onto their public data platform. According to INDIGO, the most common users of the data in their open repository are public procurement officials, policymakers, and impact and evaluation measurement experts. Looking ahead, the INDIGO team plans to continue focusing on open data, exploring the potential of sandboxes for data stewardship, and working to establish a social license for data re-use to build trust and buy-in in the broader community.

[61] Outes Velarde, Juliana. Interview re: Government Outcomes Lab Sandbox for Open Data. Zoom, March 13, 2023.

[62] Outes Velarde, Juliana. 2023. "RE: Request for some information for The GovLab's Sandboxes White Paper Project," May 18, 2023.

[63] Ibid.

2.1.6. Danish National Health Data Science Sandbox For Training And Research

The Danish National Health Data Science Sandbox for Training and Research project provides a safe environment for budding researchers to practice their data analysis and testing skills on biomedical and clinical data "without having to worry" about regulations.[64] A major hurdle facing health data research is that the use of sensitive, real data faces strict access and usage guidelines, making it difficult and high-stakes for students learning how to handle data.[65] This sandbox aims to be a teaching and training platform for data-driven health sciences research.

To this end, this sandbox provides public, anonymous real data and non-person-sensitive synthetic data. It also offers teaching materials on how to work with data, which are openly available on GitHub.[66] Sandbox data is available free of charge for researchers and students at Danish academic institutions. The data is hosted on Computerome and UCloud platforms and offers students a high-performance computing environment to play with the data.[67]

2.1.7. The Europeana Foundation's Metis Sandbox

The Europeana Foundation's mission, as part of the larger Europeana Initiative, is to "empower the [European] cultural heritage sector in its digital transformation."[68] Emerging out of a partnership between the Europeana Foundation and the Deutsche Digitale Bibliothek (DDB), the Metis Sandbox is the work of 18 partner organizations who collaboratively steward a shared data space to hold European cultural heritage data.[69]

Data input and access are open to anyone. To ensure the data is usable and meets a set of established quality standards, inputs need to include the object name, country, language, and file type are required. Using dataset and record IDs, data users can search for existing data.[70] This allows the Sandbox to include diverse data sources and types, such as data on art, music, languages, history, and more.

[64] "National Health Data Science Sandbox for Training and Research." *University of Southern Denmark*, n.d. https://escience.sdu.dk/index.php/news/national-health-data-science-sandbox-for-training-and-research/

[65] Ibid.

[66] "Health Data Science Sandbox." *GitHub*, n.d. https://github.com/hds-sandbox

[67] "National Health Data Science Sandbox for Training and Research." *University of Southern Denmark*, n.d. https://escience.sdu.dk/index.php/news/national-health-data-science-sandbox-for-training-and-research/

[68] *Europeana*. https://pro.europeana.eu/

[69] Ibid.

[70] "Metis Sandbox." https://metis-sandbox.europeana.eu/

Further, the Metis Sandbox enables data holders to test their datasets for their interoperability within the existing data portal infrastructure. It also evaluates the quality of the data, ensuring that the data records are of high quality and meet strict data quality and metadata requirements that govern the shared data pool. This reduces back and forth between data holders and the Europeana team to aggregate and prepare the data for publishing. Finally, the sandbox enables data holders to preview their datasets on the 'Europeana.eu' database to test potential use cases and synergies between their data and existing records on the platform.[71]

These three functions of testing, quality control, and previewing data facilitate more efficient and effective data collaboration and (re-)use between stakeholders. To date, stakeholders have used the Metis Sandbox to share over 56 million data records from 45 different countries. [72] The Europeana Foundation will continue to use the Metis Sandbox to further build out the Europeana database. They will also consider adding increased functionality in the sandbox, such as mapping software and additional infrastructure to encourage experiments with the datasets.[73]

2.2. Diversity of Implementation

Taking stock of the common features present across the data sandbox examples, we identify the following differences in their implementation:

- **Type of data flows:** The sectors involved in providing and playing with data, as well as the types of data on offer.
  Across the data sandbox examples, we found that G2G, G2B, and B2G data sandboxes facilitate more unlocking of administrative data, sensitive PII data, and government and business data to researchers so they can improve knowledge generation, spur innovation and accelerate product development. Meanwhile, C2B and B2C data sandboxes primarily originate from the research community to encourage sharing of data from other institutions, businesses, and government for academic purposes and seek to improve data quality and data analysis skills alongside data (re-)use for innovation.

[71] Berta, Cosmina, and Valentine Charles. "The METIS Sandbox - or finding joy in working with data." *Europeana*, n.d. https://pro.europeana.eu/post/the-metis-sandbox-or-finding-joy-in-working-with-data

[72] "Metis Data Statistics Dashboard." *Europeana*, n.d. https://metis-statistics.europeana.eu/en/data/contentTier

[73] Scholz, Henning, and Jochen Vermeulen. "How aggregators work with the Metis Sandbox to deliver high-quality data." *Europeana*, n.d. https://pro.europeana.eu/post/how-aggregators-work-with-the-metis-sandbox-to-deliver-high-quality-data

- **Data access control level:** The process and selectiveness of entering the data sandbox.
  Some sandboxes require an application and selection process, specific institutional membership, or are freely available for public use.

- **Data security measures:** The type of data provided in the sandbox and how it is presented.
  We found that data sandboxes host various types of data, including administrative data, public data, or sensitive and personally-identifiable data. Some data holders may provide anonymized, de-identified, aggregated data while others may provide identifiable and individual data points–the role of the data sandbox therefore also becomes to provide a platform for the secure storage and sharing of sensitive data. For instance, PETs may be used to provide synthetic data for the sandbox.

- **Computing and experimentation elements:** The technical features available to sandbox participants to play and experiment with data.
  These can include APIs and collaborative codebooks to develop data re-use, platforms for engaging with mentors and experts, and digital spaces to test out demo products.

- **Temporality:** The permitted time frame of data access and re-use in the sandbox.

3. VALUE PROPOSITIONS BEHIND DATA SANDBOXES

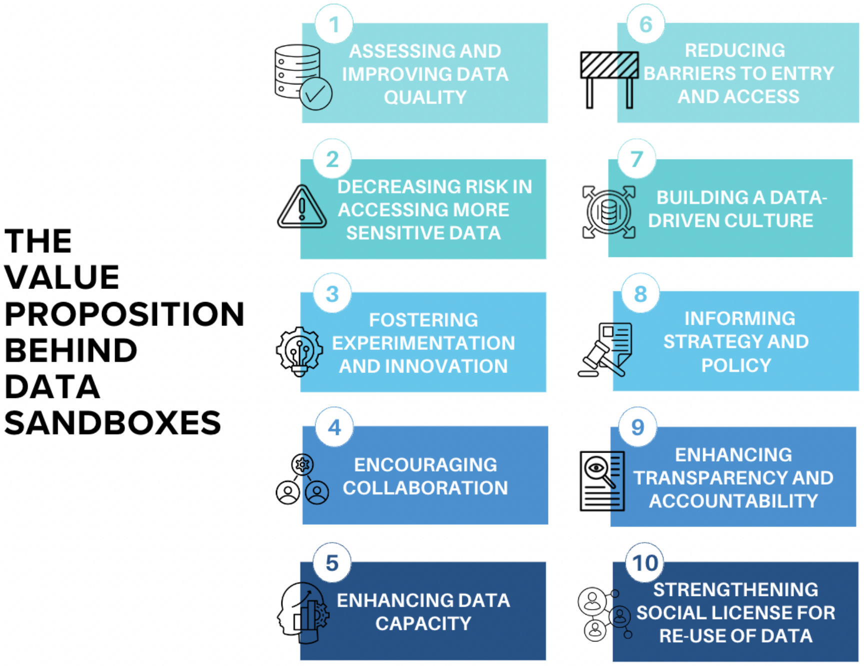

THE VALUE PROPOSITION BEHIND DATA SANDBOXES
1
ASSESSING AND IMPROVING DATA QUALITY
2
DECREASING RISK IN ACCESSING MORE SENSITIVE DATA
3
FOSTERING EXPERIMENTATION AND INNOVATION
4
ENCOURAGING COLLABORATION
5
ENHANCING DATA CAPACITY
6
REDUCING BARRIERS TO ENTRY AND ACCESS
7
BUILDING A DATA-DRIVEN CULTURE
8
INFORMING STRATEGY AND POLICY
9
ENHANCING TRANSPARENCY AND ACCOUNTABILITY
10
STRENGTHENING SOCIAL LICENSE FOR RE-USE OF DATA

Figure 3. The 10 value propositions of data sandboxes for participants. Image from The GovLab.

Data sandboxes enable data holders to share their data for greater (re-)use by data users and become more familiar with open data practices. They offer a secure environment for open data, where some of the benefits of open data, such as greater insights, can be achieved while minimizing the data privacy and security risks associated with open data. Below we detail 10 core value propositions for governments, businesses, and researchers:

1. **Assessing and improving data quality**
   Users can evaluate data quality within sandboxes, identifying inconsistencies, inaccuracies, or gaps.[74] This feedback loop enables data providers to make necessary improvements, ensuring that (open) data remains reliable, relevant, and useful.

[74] Wang, Richard Y., and Diane M. Strong. "Beyond accuracy: What data quality means to data consumers." *Journal of Management Information Systems* 12, no. 4 (1996): 5-33. http://www.jstor.org/stable/40398176

2. **Decreasing risk in accessing more sensitive data**
   Data sandboxes facilitate secure and controlled access to sensitive data, de-risking access and the re-use of data.[75] Users can work with anonymized or aggregated data to protect individual identities while still gaining valuable insights.

3. **Fostering experimentation and innovation**
   By offering access to diverse datasets and tools, sandboxes enable experimentation and innovation, possibly leading to novel insights that address challenges or capitalize on opportunities.[76] This environment nurtures the development of new insights, services, and solutions across various sectors.

4. **Encouraging collaboration**
   Data sandboxes bring together individuals with different backgrounds and expertise, fostering a collaborative atmosphere. This exchange of ideas, knowledge, and skills accelerates problem-solving and promotes cross-disciplinary innovation.[77]

5. **Enhancing data capacity**
   Data sandboxes serve as hands-on learning environments, helping users develop critical data skills such as analysis, visualization, and interpretation.[78] While this capacity-building generally begins on the individual level, as it becomes more widely adopted it will play a crucial role in fostering a skilled workforce that can leverage data effectively.

6. **Reducing barriers to entry and increasing confidence in providing access**
   By offering secure access to data and tools, sandboxes enable data holders, especially those with limited experience or resources, to participate more confidently in the data ecosystem.[79] This inclusive approach expands the pool of potential data users and contributors.

[75] "Data Sandbox: giving innovators visibility of existing data." Rail Technical Strategy, n.d. https://railtechnicalstrategy.co.uk/case-study/data-sandbox-giving-innovators-visibility-of-existing-data/

[76] Arntzen, Siri, Zach Wilcox, Neil Lee, Catherine Hadfield, and Jen Rae. "Testing Innovation in the Real World: Real-world testbeds." *Nesta*, October 2019. https://media.nesta.org.uk/documents/Testing_innovation_in_the_real_world.pdf

[77] Zheng, Jiaxin, and Thomas Schaffter. "Introducing NLPSandbox.io." *SAGE BioNetworks*, June 8, 2021. https://sagebionetworks.org/in-the-news/introducing-nlpsandbox-io/

[78] "Supporting innovation in ESG data and disclosures: the digital sandbox sustainability pilot." *Financial Conduct Authority*, June 2022. https://www.fca.org.uk/publication/corporate/digital-sandbox-sustainability-pilot-report.pdf

[79] OECD. "Responding to societal challenges with data: Access, sharing, stewardship and control." *OECD Digital Economy Papers*, No. 342, OECD Publishing, Paris, 2022. https://doi.org/10.1787/2182ce9f-en

7. **Building a data-driven culture**
   Data sandboxes serve as practical demonstrations of the benefits of data-driven decision-making.[80] By facilitating openness and collaboration around data, sandboxes contribute to a shift towards a more data-driven culture that harnesses data for the greater good.

8. **Informing strategy and policy**
   Data sandboxes can inform the development of data-related strategies and policies. By analyzing experience and feedback from those involved in a data sandbox environment, decision-makers can identify best practices and prioritize areas for improvement, testing new approaches to ensure that (open) data initiatives remain effective and sustainable.[81]

9. **Enhancing transparency and accountability**
   When designed well, data sandboxes promote transparency and accountability in data governance. This increased scrutiny can lead to more trusted data services.[82]

10. **Strengthening social license for re-use of open data**
    Data sandboxes can strengthen the social license for data re-use, resulting in a more trusted and diverse data ecosystem.[83]

4. IMPLEMENTING A DATA SANDBOX

4.1. Operational Methodology

We seek to provide a 'methodology' for sandboxes to explore data openness and (re-)use. To this end, we provide details on the operational and technical aspects to consider to responsibly oversee the data sandbox.

[80] Yiu, Chris. "The Big Data Opportunity: Making government faster, smarter and more personal." *Policy Exchange*, 2012. http://ict-industry-reports.com.au/wp-content/uploads/sites/4/2013/05/2012-The-Big-Data-opportunity-in-Government-UK-Policy-Exchange-July-2012.pdf

[81] Attrey, Angela, Molly Lesher, and Christopher Lomax. "The role of sandboxes in promoting flexibility and innovation in the digital age." *OECD Going Digital Toolkit Note*, No. 2. https://goingdigital.oecd.org/data/notes/No2_ToolkitNote_Sandboxes.pdf

[82] OECD. "Responding to societal challenges with data: Access, sharing, stewardship and control." *OECD Digital Economy Papers*, No. 342, OECD Publishing, Paris, 2022. https://doi.org/10.1787/2182ce9f-en

[83] Attrey, Angela, Molly Lesher, and Christopher Lomax. "The role of sandboxes in promoting flexibility and innovation in the digital age." *OECD Going Digital Toolkit Note*, No. 2. https://goingdigital.oecd.org/data/notes/No2_ToolkitNote_Sandboxes.pdf

The methodology spans the four stages of a sandbox project: planning, implementation, monitoring and evaluation, and sunsetting the project.

- During the **planning phase**, stakeholders develop their goals, timelines, milestones, and resource allocation strategies. This is ideally initiated by a project sponsor or problem owner, who has a legitimate stake in generating actionable insights leveraging the sandbox.

- The **implementation phase** usually involves rapid iterative cycles, where the plan can be adjusted as needed to best meet the project's objectives.

- At the **monitoring and evaluation phase**, stakeholders continuously monitor the performance of the data sandbox project using review and retrospective systems. The results of these reviews inform future interactions and improvements.

- Finally, **sunsetting the project** can involve either graduating the project to a product or developing a further iteration of the project via a more long-term data (re-)use model (see **Appendix A** for some examples).

The nature of a sandbox is to provide a space to experiment and play. As such, its operational methodology needs to be flexible and agile. Applying a 'scrum' framework, which adapts and adjusts processes and practices as needed on the go, can help support innovation and out-of-the-box thinking to manage and develop data products within the sandbox effectively. Some of the operational considerations to help achieve this balance include:

1. **Appointing a Data Sandbox Facilitator,** who will lead the sandbox and team, facilitate implementation, and ensure smooth communication among stakeholders. This person should have expertise in both the sandbox domain and data management. We refer to these individuals as 'bilinguals.'

2. **Defining the purpose and problem** by clearly articulating the objectives of the data sandbox project (i.e. is it a space to test the quality of data, prototype new data products, or improve existing data solutions?) and identifying the specific problem(s) the project aims to address. This can also involve creating a list of prioritized milestones.

3. **Identifying additional roles and positions** to ensure all necessary skills and expertise are represented. It can also be helpful to establish the roles and responsibilities of the various team members and stakeholders at this time. This can include system administrators, data stewards, and technical project managers who can speak to the good data practices, funding and sustainability needs, and overall management of the data sandbox.

4. **Establishing data and technological infrastructure requirements,** including the minimum specifications (min-specs) of the data and technological resources needed to support the data sandbox project.

5. **Establishing the role of a data steward** to define the standards of quality for data and continuously monitor, audit, and help enhance data in the sandbox to uphold quality standards. A data steward is also responsible for summarizing and communicating the results of projects in the sandbox with relevant stakeholders.

6. **Developing a methodology for categorizing the sandbox's data assets** in collaboration with industry players to map key datasets. This can be especially important when using PETs such as synthetic data to create bespoke datasets for sensitive use cases.

7. **Providing strong metadata** about the data available in the sandbox to help potential users understand the sandbox's purpose and its alignment with their needs.

8. **Defining data access and use parameters** that clearly outline the eligibility requirements and process of joining the sandbox. This can include creating a system for soliciting, vetting, selecting, and communicating with sandbox users to stay aligned to core values.

9. **Building dispute resolution processes** before they are needed to proactively handle disagreements or legal/regulatory issues that may arise in the sandbox.

4.2. Focus on Governance of a Data Sandbox

The GovLab defines governance as the *processes and practices that enable an organization to fulfill its purpose in accordance with a set of principles*. As such, when designing the governance framework of a data sandbox, it is important to specify and implement the “4Ps” of governance:

- Clarifying the **purpose** of the sandbox for open data,
- Defining the guiding **principles**,
- Structuring the **processes** of decision making, and
- Translating decisions into **practices** that seek to meet the purposes and principles of the sandbox.[84]

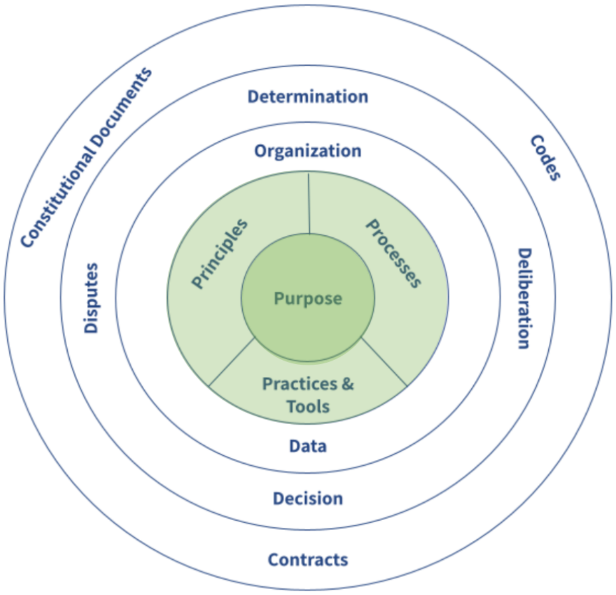


Figure 4. The 4Ps of Governance Framework. Image adapted from Verhulst and Young (2022).

In the below we provide further detail regarding the steps and approaches that can enable an adequate and responsible governance framework for data sandboxes.

4.2.1. Step 1: Specify the Purpose of Data Sandboxes

As mentioned previously, data sandboxes can serve various purposes (see *Purposes of a Data Sandbox*). From a governance perspective, it is important to clearly and narrowly define the purpose of a data sandbox and how it (re-)uses data and advances social good.

[84] Verhulst, Stefaan G., Anthony Zacharzewski, and Christian Hudson. “Governing the Environment-Related Data Space.” *Data and Policy Blog*, October 3, 2022. https://medium.com/data-policy/governing-the-environment-related-data-space-98794ed9accb

4.2.2. Step 2: Define Principles: What does the sandbox stand for?

Principles provide a common set of ideals and values that exist across the data sandbox's structure. Some of the priority principles we identified include being:

1. **Fit-for-purpose:** The design of the data sandbox enables its mission, leading to more impactful and targeted outcomes.

2. **Trusted:** The sandbox makes decisions aligned with its agreed-upon principles and values and balances the interests of all its members, increasing trust as a result.

3. **Secure:** The sandbox protects the data and data derivatives it holds with top-notch privacy and security measures and data (re-)use audits.

4. **Legitimate:** The sandbox is accepted as a data (re-)use hub by the relevant stakeholders.

5. **Standardized:** The sandbox has established processes for data ingest, usage, management, and sharing.

6. **Accountable:** The sandbox is held accountable for its actions; independent review procedures are in place to monitor conduct.

7. **Effective:** The sandbox is given sufficient support and resources to implement governance principles.

8. **Timely:** The sandbox provides access to its data and interfaces via rapid approval and for a short, defined period of time.

9. **Collaborative:** The sandbox's work complements existing data, standards, and policy work and involves partnerships with governments, the private sector, data communities, and other organizations.

10. **Supportive:** The sandbox has low barriers to participation and learning to involve new and emerging players to take part in the initiative.

4.2.3. Step 3: Design Processes and Practices: How will the sandbox achieve its purpose and adhere to its principles?

To meet its purpose, the sandbox will need to develop a set of processes and practices for governing and overseeing the sandbox that adheres to its principles. These include the following:

Process	Practice
Creation of an oversight/ management system: Implement a high-level governance structure that oversees the core objectives of the sandbox in a way that is fit-for-purpose, and cost-effective.	This management system could include bodies such as a: ● **Oversight Board:** decision-makers that advise on overseeing decisions regarding access to sandbox data, the temporality of the sandbox, and if/when the time comes, the transition from a sandbox to another data (re-)use model. The Oversight Board assesses these points by looking at: ○ *What* types of data access should be provided; ○ *Who* can access the data; ○ *When and Where* the sandbox will take place and its operational parameters; ○ *How* to prevent undue corporate, academic and policy influences in decision-making and *how* to prevent mis-use of the insights generated. ● **Advisory Board:** Advisors that seek to enable financial and operational sustainability, fiduciary oversight, and technological advancement of the sandbox project. They maintain relationships with funders, data holders, data engineers, etc. They help set the strategic direction and oversee the operational and financial performance of the sandbox (i.e. they can help decide whether to run one-off or multiple sandboxes around a specific type of data or issue area). ● **Dispute Resolutions Committee (see also "Management of dispute resolution"):** This committee consists of individuals trained in organizational conflict resolution who can be called up to address any disputes between the sandbox facilitators and participants if the need should occur.

Establishment of conditions for access: Depending on the data, stakeholders, and purpose for the sandbox, it will require different access controls.	**Set data access protocols.** Consider whether the data at hand permits sandbox participants to download data and work with it on their own interfaces, or if it is too sensitive to be extracted. Work with participants to set permissions and explore how PETs including differential privacy, federated learning, and secure multi-party computation can improve data user comfort while balancing against data leakage/privacy. (Financial Conduct Authority, 2022)
	Provide strong metadata about the data available in the sandbox to help potential users understand the sandbox's purpose and its alignment with their needs. (Metis Sandbox, 2023)
	Bring in data stewards who oversee and manage the data access process to the sandbox. Data stewards will be responsible for reviewing and processing data access requests in line with the purpose and principles of the sandbox, and will ensure the data is used responsibly by participants. (Verhulst et. al., 2020)
Review of sandbox participant applications: There is a need to define the protocol for sandbox eligibility and access, as well as verify researchers and research projects accurately and rapidly.	**Assess motivation to participate in the sandbox and ensure they are aligned with the sandbox's purpose.** This will involve evaluating the potential benefits and risks associated with each data access request, as well as assessing the qualifications and competency of the researchers who submit the requests. Involve data and industry experts when vetting the applications. (Rousseau, 2023)

	Create common standards to manage participant vetting, data access, and the auditing process effectively and efficiently. This can include establishing an application template detailing what the project is, what data can help, where can the sandbox help, what the intended outcome of the sandbox is, and who is involved in the project. (Rousseau, 2023)
Management of dispute resolution: At times, conflicts and disputes may emerge between sandbox participants. Have established processes and necessary people in place beforehand to address these situations.	**Prevent conflicts of interest** between sandbox participants by requiring transparency from participants. This can include implementing disclosure requirements, creating protocols for handling conflict of interests between members, and establishing clear guidelines for the purpose of the sandbox and parameters of data access and (re-)use. (University of Southern Denmark, 2023)
	Establish a Dispute Resolution Committee. This committee will adjudicate disputes in the sandbox with the help of an internal ombudsperson or neutral third-party who can review and resolve disputes.
Active engagement with sandbox participants: Cultivating strong, long-lasting connections between data holders and data users across various sectors helps foster a culture conducive to data access and (re-)use.	**Facilitate a 'teaching and doing' environment** in the sandbox between data holders, users, governments, businesses, etc. Use data stewards to help improve data analysis, (re-)use, awareness, and literacy skills among participants. This can include creating a common communication channel and directory so people can connect with each other.

	Collaborate with industry associations, research consortia, and other organizations to establish best practices for data sharing and access within a sandbox structure and promote an open data culture and community among key stakeholders. (Rousseau, 2023)
Conduction of routine impact assessments: Consistently monitoring the direction, structure, and features of the sandbox helps ensure that the sandbox is performing to meet its established purposes and needs.	**Consider the needs of the sandbox users.** For instance, early-stage products could benefit more from data-related features (APIs, data dashboards, etc.) to test and validate products while mature stage products may benefit more from product preview decks and collaborative networks to beta-test their product and its business model. Create a decision tree to assess and tailor sandbox functionalities to user needs to avoid costly creation of unnecessary tools. (Financial Conduct Authority, 2022)
	Design a funding and sustainability model for the sandbox. Consider the maturity of the sandbox users (e.g. are they established companies that would be open to paying for data access, or are they start-ups that would require the sandbox to inject funding to scale up?) Design these policies while keeping in mind the duration of the sandbox and whether the initiative is a one-off or part of a series of sandboxes. (Rousseau, 2023)

Table 1. Suggested processes and practices that operationalize the purpose of the data sandbox.

4.2.4. Enabling Conditions

In order to operationalize the governance framework, the sandbox's infrastructure needs to satisfy **three** enabling conditions.[85] Decision-makers should keep the following in mind when implementing a data sandbox:

1. **Sufficient resources to implement the governance framework:** To be trusted and efficient, the sandbox must have the resources and human capital to uphold governance decisions—from purpose to principles to processes and practices—and protect data security and privacy.
2. **Strong relationships between sandbox hosts and participants:** A sandbox is only as strong as its participants. Cultivating strong connections with data holders (who provide high-quality data and trust its (re-)use) and data users (purposeful initiatives run by vetted researchers) is a key component of sustaining the sandbox and achieving overarching goals.
3. **Adequate decision-making processes and staff support:** The sandbox should seek to establish collective agreement and compliance whenever and wherever possible. To achieve this, it requires dedicated staff focused on policy, research, engagement, and technological development and processes and practices that foster collective agreement in a timely, coordinated, and transparent fashion.

5. CHALLENGES OF DATA SANDBOXES

Data sandboxes provide value to a diverse set of stakeholders. However, establishing, adopting, and maintaining a sandbox also poses certain risks and challenges. In what follows, we explore key challenges facing sandbox implementation and offer strategies to mitigate them. The literature highlights **five** main data challenges to sandbox creation, namely **attracting relevant data; obtaining high-quality data; keeping data fresh; translating sandbox solutions to the real world; and managing risks across the data lifecycle.**

[85] An enabling condition refers to resources, both operational and human, required to carry out the initiative. These are broader, ecosystem-level requirements without which the initiative would not be able to sustainably be implemented.

5.1. Attracting Relevant Data

For a sandbox for open data to generate value, its data "must be considered within the context of the task at hand."[86] Yet getting data users and data holders to agree on what they intend the sandbox to explore may be difficult due to different backgrounds and motivating agendas. Navigating asymmetries, especially with regard to data and agency, may prove difficult at this initial stage of a sandbox for open data. Data holders may not see how their data is relevant for the sandbox; data decision-makers may be wary of relinquishing regulations around certain types of data.

Overcoming these obstacles requires a clear and mutually agreed-upon sandbox subject scope. This can be crafted through participatory topic mapping, which brings stakeholders together to organize the problem space and identify the priority focus areas for which data can be used to experiment on new solutions.[87] Through this exercise, sandbox participants can reflect on the "conceptual and technical pre-occupations" of a specific discipline to understand what topics are important, how they interact and connect with one another, and where data can play a role in answering longstanding questions.[88]

5.2. Obtaining High-Quality Data

Once the focus area and objectives for the data are defined, how the data is sourced, processed, and presented must be considered. For instance, sandbox stakeholders must consider if the data at hand is precise and granular enough for their purposes, if the data is accurate, reliable, and believable, and if the data is representative and complete to mimic real-world applications. These considerations must also be weighed against popular concepts of data minimization, which means that only the essential data needed is collected/kept, and data proportionality, or "the practice of collecting and storing only personal information that is directly relevant (i.e., proportional) to a given task or purpose (i.e. purpose specification)."[89]

[86] Wang, Richard Y., and Diane M. Strong. "Beyond accuracy: What data quality means to data consumers." *Journal of Management Information Systems* 12, no. 4 (1996): 5-33. http://www.jstor.org/stable/40398176

[87] Verhulst, Stefaan, Constanza M. Vidal Bustamante, Liliana Carvajal-Velez, Fiona Cece, Jennifer Harris Requejo, Alexandra Shaw, Michelle Winowatan, Andrew Young, and Andrew J. Zahuranec. "Toward a demand-driven, collaborative data agenda for adolescent mental health." *Journal of Adolescent Health* 72, no. 1 (2023): S20-S26. https://doi.org/10.1016/j.jadohealth.2022.05.027

[88] Sutherland, William J., Robert P. Freckleton, H. Charles J. Godfray, Steven R. Beissinger, Tim Benton, Duncan D. Cameron, Yohay Carmel et al. "Identification of 100 fundamental ecological questions." *Journal of Ecology* 101, no. 1 (2013): 58-67. https://doi.org/10.1111/1365-2745.12025

[89] Verhulst, Stefaan. "Questions as a Device for Data Responsibility: Making data science responsible by formulating questions in a polycentric way." *Forthcoming*.

A well-defined and actionable governance structure on how data and data management systems are to be structured can help mitigate these challenges to uphold data quality standards and maintain the security of data and data subjects. Outlining the data format requirements, sourcing information, update status, and metadata reporting requirements (such as tagging data by type, update frequency, fields, data holder, etc.) to build a 'data codebook' creates strong content management practices. These practices help to better categorize and match relevant datasets with those seeking data, bringing order and accountability to the sandbox.[90]

5.3. Keeping Data Fresh

Having access to relevant and current datasets is crucial in ensuring that the sandbox tests align as closely as possible to their real-world use cases. Yet in addition to bringing data holders and their data to the sandbox, updating and maintaining relevant data in the sandbox can present a challenge (if not linked automatically). [91]

Indeed, data 'freshness' is critical for information system success. Bouzeghoub and Peralta (2004) note that data freshness has two factors. First, they consider the *currency* dimension, which refers to the time between data extraction and data delivery, the number of updates to data since its extraction, and the number of data categories that are up-to-date.[92] Second, they consider the timeliness of data or the time since its last update.[93] Sandbox users must agree on their data freshness expectations and consider the time and resources needed to update, maintain, and audit data relevancy against these quality standards. Specifically, they need to consider the type of data at hand (does it need to be frequently or periodically updated or is it stable?); the type of data sourcing system (is the sandbox retrieving data in real-time, refreshing cached data, or holding large troves of data that are periodically refreshed?); and the synchronization policies to facilitate coordination between the data management system and the data sources (does the system synchronously or asynchronously pull and push data between data sources and the system?).

[90] Hayler, Andy. "Common data lake challenges and how to overcome them." *TechTarget*, April 17, 2020. https://www.techtarget.com/searchdatamanagement/feature/Common-data-lake-challenges-and-how-to-overcome-them

[91] Hutchison, Matthew. "5 Salesforce Sandbox Seeding Challenges and How to Solve Them." *OwnBackup*, June 24, 2020. https://www.ownbackup.com/blog/5-salesforce-sandbox-seeding-challenges/

[92] Bouzeghoub, Mokrane. "A framework for analysis of data freshness." In *Proceedings of the 2004 international workshop on Information quality in information systems*, pp. 59-67. 2004. http://apmd.prism.uvsq.fr/Publications/Articles/A%20Framework%20for%20Analysis%20of%20Data%20Freshness_Peralta,%20Bouzeghoub.pdf

[93] Ibid.

Monitoring and updating available data, and aligning new data to the sandbox's format standards, require time and resource investment. To mitigate this issue, sandboxes for open data must consult with their users to understand the types of data, data sourcing, and data synchronicity needed to build a fit-for-purpose freshness model.

5.4. Translating Sandbox Solutions to the Real World

A fourth drawback of a sandbox for open data is the applicability and viability of sandbox projects in practical settings. Indeed, sandboxes are somewhat of a double-edged sword: while they allow for open experimentation, they can sometimes leave users blind to their external regulatory and emerging technologies environments.

Berndtsson, Ericsson, and Svahn (2020) note some of the difficulties that occur when attempting to scale up data-driven pilots. Poor team management, lack of skills, and unrealistic expectations of what the initiative aims to achieve, coupled with low-quality data, lack of support from senior management, and an inability to turn insight into action can hinder a sandbox project from entering the real market.[94] Before taking a promising project outside of the sandbox, stakeholders must ensure that the project has the required skills, resources, and business/managerial buy-in to support and sustain a larger launch.

As well, for initiatives that fall under existing regulatory and legal frameworks, failing to properly account for real-world stochasticity in the sandbox can result in numerous barriers when bringing data applications to the real market.[95] For initiatives that give rise to new governance measures, although the sandbox involves policymakers and legislators, the lag between governance implementation and data product launch may still be inefficient. To overcome these challenges, the sandbox environment needs to be adaptive and practical. It needs to closely mimic the environment of the use case and consider the hurdles facing new governance measures to support innovative products in order to avoid oversimplifying real contexts and develop products and systems that will work just as well in the real world as they do within the confines of a sandbox environment.[96]

[94] Berndtsson, Mikael, AnnMarie Ericsson, and Thomas Svahn. "Scaling Up Data-Driven Pilot Projects." *AI Magazine* 41, no. 3 (2020): 94-102. https://doi.org/10.1609/aimag.v41i3.5307

[95] Gurram, Nagender, and Yudhish Batra. "How to Tackle the Problem of an Unmanaged Data Sandbox." *CapitalOne*, August 29, 2022. https://www.capitalone.com/software/blog/solving-unmanaged-data-sandbox/

[96] Datasphere Initiative. "Sandboxes for data: creating spaces for agile solutions across borders." *Datasphere Initiative*, 2022. https://www.thedatasphere.org/wp-content/uploads/2022/05/Sandboxes-for-data-2022-Datasphere-Initiative.pdf

5.5. Managing Risks across the Data Lifecycle

There are numerous risks present at each stage of a data project.[97] These risks extend to data sandboxes as well. Below, we outline some of the most prevalent risks across each stage of a data project:

1. **Planning:** During the planning stage of a data sandbox, stakeholders face the risk of a lack of trust. This in turn can lead to misaligned expectations and values when it comes to (re-)using the data in question. At this point of the project, it is also important to outline accountability systems to mitigate the risk of poor governance systems.

2. **Collecting:** When collecting the various data resources for the sandbox, actors must consider the risk not only of biased or bad data but also the ethical risks around data exploitation and poor handling of data. Considerations around data subjects' rights and consent mechanisms also pose both a risk and an opportunity for sandbox actors.

3. **Processing:** The risks at the processing stage are more technical in nature, and center around data storage and use. The main risks at this point of the process are unsafe data storage practices, loss of data, and data misuse.

4. **Sharing:** In a data sandbox, there are two aspects of sharing. The first is in terms of sharing data and resources within the sandbox. Here, the risks of stakeholders are similar to those during the collecting stage, focused mainly on ethical and responsible data sharing practices. The second aspect is sharing data and insights publicly. In this case, stakeholders must consider reputational and privacy risks that could emerge from publicly sharing their data or insights.

5. **Analyzing:** When conducting analyses in a data sandbox, it is important to consider the risks of biases or blindspots among stakeholders and within the datasets. The presence of biases and blindspots can further lead to the risk of reproducing inequalities and of monolithic evaluations.

6. **Using:** At this final stage, where stakeholders seek to translate insights into action, the risk of unintended consequences may emerge alongside the risk of missed uses of the data. Towards the end of a sandbox project, long-term data

[97] "Data Responsibility Journey." *The GovLab*. https://dataresponsibilityjourney.org/steps

retention strategies may also pose a risk to stakeholders if they are not carefully implemented.

6. EMERGING TRENDS IN THE DATA SANDBOX ECOSYSTEM

As we look ahead, we consider the question of ‘What could be next for emerging data sandboxes?’. In what follows, we highlight critical trends in the ecosystem in terms of not only technical advancements but also forthcoming changes in regulation and legislation in the European Union.

6.1. Technical Advances

When it comes to advances in technology, we identify **four** key areas of progress: D**ata Generation; Data Preparation, Storage, and Computation; Analytics; and Access Management.** Understanding and following these trends in data generation, preparation, and analytics can help drive the creation of more secure data sandboxes.

6.1.1. Advances In Data Generation

Recent developments in the Internet of Things (IoT) and synthetic data creation methods have transformed the ways in which data is generated and used across contexts.

IoT refers to the global infrastructure of sensors, software, and other technologies that connect and exchange data over the Internet.[98] Currently, this network consists of more than 7 billion devices, a number that is expected to triple by 2025.[99] IoT devices can capture real-time, reliable data on a wide range of phenomena, from air quality and individual heart rates to urban traffic patterns and daily coffee orders. As the network of IoT devices continues to grow and expand, the volume of high-quality data generated quickly and efficiently will continue to grow. In the context of data sandboxes, this data can offer better intelligence leading to stronger insights for stakeholders. Combined with advances in data analytics, larger volumes of data can also improve operational efficiency both within and outside of the sandbox environment.[100]

[98] “Internet of Things Global Standards Initiative.” International Telecommunication Union, July 2015. https://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx

[99] “Internet of Things.” *Oracle*, n.d. https://www.oracle.com/in/internet-of-things/

[100] Neto, Jose Antonio Ribeiro. “Data Generated by IoT.” *Medium*, May 17, 2020. https://medium.com/xnewdata/data-generated-by-iot-2873c9f9d0c5

In addition to developments in the IoT environment, there have also been rapid advancements in the creation and use of synthetic data. Synthetic data refers to datasets that are created via algorithms either by sampling real-world data or through simulated interactions between models and processes with the goal of creating new data without taking data directly from the real world.[101] As a PET, synthetic data can mitigate concerns around using personal data and sensitive data because it does not tie back to real individuals yet still mirrors their real-world contexts.[102] This technology can help drive the adoption and implementation of data sandboxes as stakeholders can share synthetic data gleaned from their real data, thereby minimizing their risk of data sharing while allowing partners in the sandbox to conduct accurate and robust analyses and achieve relevant insights. As the technology around synthetic data continues to improve, the process of creating and using data will only become more efficient and effective.

6.1.2. Advances In Data Preparation, Storage, And Computation

Alongside advances in data generation, there has been a similar transformation in data preparation, storage, and computation methods. As previously mentioned, PETs are especially relevant for data sandboxes because they play a key role in strengthening the security of the sandbox and building trust.

Privacy Enhancing Technologies (PETs), also sometimes referred to as "Partnership Enhancing Technologies" or "Trust Technologies," are defined as "a suite of tools that can help maximize the use of data by reducing risks inherent to data use."[103] PETs offer new ways for data holders to provide access to closed or sensitive datasets, to outsource computations, to implement governance frameworks, and to decentralize data services.[104] They also enable multiple stakeholders across organizations to combine datasets and run shared analyses.[105] As with other technologies and systems, there is no one-size-fits-all approach to implementing PETs, and it is important to understand what approaches will best fit the needs of the stakeholders

[101] "Synthetic Data." *Gartner Glossary*, n.d. https://www.gartner.com/en/information-technology/glossary/synthetic-data

[102] "How To Generate Synthetic Data? Populating a Sandbox Development Environment." *Awesense*, November 10, 2022. https://www.awesense.com/how-to-generate-synthetic-data/

[103] "Privacy Enhancing Technologies." *The Royal Society*, January 23, 2023. https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/

[104] "Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis." *The Royal Society*, March 2019. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Protecting-privacy-in-practice.pdf?la=en-GB&hash=48A28CDF4FB012663652BE671CFFED08

[105] Ibid.

based on scale, data types, and existing data infrastructure[106] For data preparation and handling in sandboxes, two types of PETs are important to consider: confidential computing and secure, multi-party computation.

Confidential computing uses cloud technology to "isolate sensitive data in a protected area, for instance, in a specific CPU, during data processing."[107] This secure environment means that only authorized stakeholders can access the data being processed, as well as the models or techniques being used to analyze the data.[108] This setup offers data holders a private, protected, and secure channel through which sensitive data and computing workloads can be shared.[109]

In addition to confidential computing, secure multi-party computation (SMPC) offers data holders another way to contribute to a data sandbox. In this model, rather than sharing data with stakeholders to form one shared dataset, the SMPC distributes a computation across multiple parties.[110] This method of taking the algorithm to the data (rather than the more traditional method of bringing the data to the algorithm) allows stakeholders to gain insights from distributed data without exposing their data assets to the risks of a shared platform.[111] While this model doesn't require stakeholders to share their data directly, an SMPC model can help stakeholders build trust and observe firsthand the way the value of their datasets grows when they are combined with other datasets.

6.1.3. Advances In Analytics

The development and evolution of generative artificial intelligence (AI) systems have dominated global headlines since early 2023, with promises of dramatically transforming the ways in which we work, learn, and engage with the world. It also has the potential to transform the way we interact with and leverage open data.

Generative AI "describes algorithms that can be used to create [or generate] new content, including audio, code, images, text, simulations, and videos."[112] Recent

[106] Ibid.

[107] "What is confidential computing?." *IBM*, n.d. https://www.ibm.com/topics/confidential-computing

[108] Ibid.

[109] Ibid.

[110] "What is Secure Multiparty Computation?." *Inpher*, n.d. https://inpher.io/technology/what-is-secure-multiparty-computation/

[111] Ibid.

[112] "What is generative AI?" *McKinsey & Company*, January 19, 2023. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

examples of generative AI tools include Chat GPT,[113] Bard,[114] and Bing AI.[115] These technologies largely depend on open data portals and access to large datasets to improve the quality of their outputs by reducing biases and expanding capabilities.

For data sandboxes, generative AI, combined with Natural Language Processing (NLP) technology, could change the way we access and use datasets.[116] Given generative AI's ability to interpret and respond to a line of questioning, this technology can make datasets more easily searchable, helping to break down silos in the open data ecosystem and match the supply of data with relevant demand. For stakeholders in a data sandbox, this will likely improve interactions between data holders and data users by making it easier and more efficient to access relevant data sources for re-use. When combined with other advancements in the data ecosystem, generative AI can also improve computational outcomes.

6.1.4. Advances In Access Management

Data access management is a critical element of a data sandbox's infrastructure in ensuring a secure environment. Data access management refers to a "set of processes and technologies used to control access to applications or data."[117] This not only helps to create a secure environment by preventing unauthorized access or use, but it is also key in preventing data breaches and compliance issues.[118] It can include creating defined access privileges, using password managers, maintaining activity monitors, and implementing multi-factor authentication systems. Technological advancements in the data security ecosystem are enhancing the security of data sandboxes by minimizing the risk of data sharing within the controlled environment.

In order to implement effective data access management, it is helpful to maintain certain best practices, such as: establishing an inventory of data assets and stakeholders; understanding data provenance structures; and running regular activity and user audits.[119] As part of a sandbox's governance framework, a security group representing all of the stakeholders involved can help to monitor data access and

[113] *OpenAI*. https://openai.com/blog/chatgpt

[114] *Bard*. https://bard.google.com/

[115] *Bing AI*. https://www.bing.com/?/ai

[116] Nagy, Sarah. "Generative AI and the Natural Language Interface for Data." *Data Council*, n.d. https://www.datacouncil.ai/talks/generative-ai-and-the-natural-language-interface-for-data

[117] "What is Data Access Management?." *ONEWEB Business Application Platform*, n.d. https://www.oneweb.tech/api/661WhatisDataAccessManagement.html

[118] Ibid.

[119] "Data Access Management Basics & Implementation Strategy." *OvalEdge*, September 29, 2021. https://www.ovaledge.com/blog/data-access-management-basics-implementation-strategy

ensure that actors are upholding the governance principles and regulatory requirements of the sandbox.[120] Additionally, there has also been a recent push for increased data literacy focused especially on secure access and responsible (re-)use. This can take the form of certification requirements for access, for example, that check the degree of data literacy of users.[121] These checks and balances also make sure that all users have a shared understanding of governance practices and regulatory requirements of the data sandbox.

6.2. Shifting Policy Regulations

As policy regulations in the European Union around the responsible and innovative use of data and technology gain traction and increased calls for oversight, data sandboxes could act as a secure way to facilitate and experiment with new forms of data sharing in an evolving regulatory space.

6.2.1. The European Union Digital Markets Act

In May 2023, the European Union enacted the Digital Markets Act (DMA). The DMA aims to establish fair markets in the digital sector, focusing especially on access to and (re-)use of data to promote greater business and innovation for the digital economy.[122] The Act also plans to foster greater market contestability, allowing new players to enter the market.[123]

In order to implement the Act, policy and data experts in the region advocate for the implementation of institutional data sandboxes as a key policy tool.[124] In addition to providing greater access to data for (re-)use for the public good, a data sandbox could help improve the quality of data being shared and foster a stronger culture around data sharing and collaboration. Finally by partnering with trusted institutions, these sandboxes would not only receive greater legitimacy but would also likely be more sustainable in the long-run.[125]

[120] Ibid.

[121] Ibid.

[122] Krämer, Jan. "Data Access Provisions in the DMA." *Centre on Regulation in Europe (CERRE) Issue Paper*, 2022. https://cerre.eu/wp-content/uploads/2022/11/DMA_DataAccessProvisions-2.pdf.

[123] "DMA: rules for digital gatekeepers to ensure open markets start to apply." *European Commission Directorate-General for Competition and Directorate-General for Communications Networks, Content and Technology*, May 2, 2023. https://digital-markets-act.ec.europa.eu/dma-rules-digital-gatekeepers-ensure-open-markets-start-apply-2023-05-02_en

[124] Krämer, Jan. "Data Access Provisions in the DMA." *Centre on Regulation in Europe (CERRE) Issue Paper*, 2022. https://cerre.eu/wp-content/uploads/2022/11/DMA_DataAccessProvisions-2.pdf.

[125] Ibid.

6.2.2. The European Union Data Governance Act

The European Union's Data Governance Act was ratified in 2022 and goes into effect formally in September 2023. While similar to the Data Act, the Data Governance Act "seeks to increase trust in data sharing, strengthen mechanisms to increase data availability and overcome technical obstacles to the reuse of data."[126]

As part of this work, the European Commission advocates for the creation of shared European data spaces for diverse industries across sectors.[127] They aim to foster trusted data sharing through 4 measures: sharing sensitive public sector data in closed environments for re-use; implementing a network of trusted data intermediaries; creating systems for businesses and citizens to share their data for use to promote public good; and facilitating cross-sectoral data sharing across borders.[128]

Once the Data Governance Act goes into effect, data sandboxes can play a key role in fostering trusted data sharing across industries and sectors. As demonstrated by the previous examples, data sandboxes provide a customizable system of creating secure and trusted environments for data sharing to pilot the sharing of sensitive data and the use of data intermediaries before they become public-facing systems. Additionally, their ability to replicate real-world conditions and contexts can enable stakeholders to run robust risk assessments on their data re-use cases to identify unanticipated challenges and results before they're opened up for public use.

6.2.3. The European Union Data Act

One of the central tenets of the proposed Data Act is to make data more accessible for responsible use to drive economic growth and technological innovation.[129] In order to do so, the act aims to dismantle existing legal, economic, and technological barriers to data re-use, including mistrust between public and private actors, challenges with data quality and interoperability, and the high costs of building sustainable and effective data infrastructures.

[126] "European Data Governance Act." *European Commission*, n.d. https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[127] "European Data Governance Act." *European Commission*, n.d. https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[128] Ibid.

[129] "European Data Strategy." *European Commission*, n.d. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

A key focus involves promoting safe and effective data collaborations between public and private sector entities.[130] They plan to, for example, include systems to prevent the emergence of power asymmetries in data sharing agreements; to enable public sector bodies to access private sector data in emergency contexts; and to create legal safeguards around data transfers.[131]

Thus, data sandboxes could offer a promising tool for the implementation of the Data Act and its proposed amendments. By creating a safe environment for stakeholders to share data and run analyses, a data sandbox can enable actors to collaborate while protecting their intellectual property rights and trade secrets from infringement. Furthermore, governance frameworks and dispute resolution strategies within a data sandbox can help prevent power asymmetries from emerging between stakeholders, and ensure that fair and responsible data practices are upheld. In this way, data sandboxes can play a key role in transforming the Data Act from policy to practice.

7. CONCLUSION

This white paper aimed to introduce policymakers to the concept and application of data sandboxes as a method for opening and re-using data while protecting sensitive information.

Earlier, in addition to defining what data sandboxes are, we considered two key questions surrounding data sandboxes: **Could they empower decision-makers to unlock the power of open data while maintaining the necessary safeguards for data privacy and security? And could they help decision-makers overcome barriers to data access and promote purposeful, informed data (re-)use?**

To answer these questions, we first defined the six characteristics of data sandboxes, noting that they are controlled, secure, collaborative, and high computing environments that are temporal and adaptable. Following a review of different types of existing data sandboxes, we provided the 10 value propositions that the model offers to decision-makers, specifically with regard to improving data quality and data handling capacities, building trust and collaboration between sectors, and increasing confidence in data (re-)use initiatives across the stakeholder ecosystem. Moreover, under the appropriate governance structures, data sandboxes can indeed provide an environment that balances data innovation and re-use and data privacy and security.

[130] "Data act: member states agree common position on fair access to and use of data." *Council of the EU*, March 24, 2023. https://www.consilium.europa.eu/en/press/press-releases/2023/03/24/data-act-member-states-agree-common-position-on-fair-access-to-and-use-of-data/

[131] Ibid.

Using the “4Ps” of governance methodology, we recommended principles, processes, and practices that decision-makers seeking to set up a data sandbox could follow.

Ultimately, we believe that data sandboxes are a powerful and effective tool for testing data (re-)use initiatives in a low-risk, controlled, and safe manner. The specific parameters of data access and (re-)use of data sandboxes allow data holders and data users to learn how to handle data sharing in a synergistic manner. Furthermore, data sandboxes allow for experimentation with both data sharing and re-use practices and data privacy and security methods, as evidenced by the emergence of PETs, sensor data, and data preparation and management systems. These technical, operational, and governance advances help promote purposeful data (re-)use for decision-makers. Data sandboxes allow actors to test initiatives at scale, trialing the parameters for data sharing and getting an understanding of the costs associated with data-driven initiatives. These features help promote an open data culture and provide easy-to-follow and manage entryways for actors seeking to join the data (re-)use movement.

APPENDICES

Appendix A: Glossary of Different Data (Re-)Use Models

There are other data (re-)use and data collaborative models that are similar to data sandboxes. Due to their similar goal of collaborating with data for (re-)use, oftentimes these models are confused with each other.

Yet it is important to reiterate that data sandboxes are defined not only by their collaboration but also by their **temporal nature**. That said, a data sandbox could transition towards a more long-term collaborative model. Below, we define some of the other prevalent models in the data (re-)use ecosystem to raise awareness of differences across terminology.

Data Clean Rooms:[132] Data clean rooms act as a secure environment for multiple stakeholders to share “clean” or de-identified data with each other. Data clean rooms are quick to set up and are often temporary arrangements to help actors rapidly garner and act on data-driven insights.

[132] Durand, Julian. “How ‘Data Clean Rooms’ Evolved From Marketing Software To Critical Infrastructure.” *Forbes*, April 26, 2023. https://www.forbes.com/sites/forbestechcouncil/2023/04/26/how-data-clean-rooms-evolved-from-marketing-software-to-critical-infrastructure/?sh=2f207b84619c

Data Enclaves:[133] A data enclave can be used to transfer confidential data between stakeholders across a secure network. While it allows for access to the data, it does not support shared analysis functions. Data enclaves can exist as both physical and virtual spaces.

Data Labs:[134] A data lab is a type of closed data science system that enables data holders to run experiments and analyses on their own data. They are often used to uncover new business opportunities and innovative processes.

Data Lakes:[135] A data lake is an internal repository where data holders can store their structured and unstructured data assets. The data can then be used for future analyses after it is processed.

Data Pools:[136] A data pool is a repository where multiple stakeholders can store, access, and share data in a structured manner or standard format. Data pools act mainly as a data storage platform rather than a safe environment to conduct data analyses.

Data Spaces:[137] A data space is a virtual storage space that contains data alone. A data space is not shared, nor does it allow users to execute analyses or other functions on the data.

Data Trusts:[138] Data trusts act as a tool to empower individuals or communities to decide what data they choose to share and how that data is used. Data trusts are important mechanisms when it comes to data (re-)use for public good, however, they are not designed to be a space for experimentation with data.

Appendix B: Audit of European Data Sandbox Ecosystem

Drawing on the findings of our seven data sandbox case studies, we crafted the following audit of the European data sandbox ecosystem. Our ecosystem audit identifies the types of data flows, levels of control, data security measures, computing

[133] "Data Enclave." *Network of the National Library of Medicine,* n.d. https://www.nnlm.gov/guides/data-thesaurus/data-enclave

[134] "What is a Data Lab?." *Talend*, n.d. https://www.talend.com/uk/resources/what-is-data-lab/

[135] "What is a Data Lake?." *Amazon Web Services*, n.d. https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

[136] "Data Pool." *IBM*, March 16, 2023. https://www.ibm.com/docs/en/imdm/11.6?topic=gds-data-pool

[137] "Creating and using data spaces." *IBM*, March 22, 2021. https://www.ibm.com/docs/en/zos/2.2.0?topic=guide-creating-using-data-spaces

[138] *Data Trusts Initiative*. https://datatrusts.uk/

capacities, and timeframes of the different sandboxes to provide a summary of the diverse strategies and approaches that exist in this growing ecosystem. By understanding the diverse approaches at play, stakeholders interested in designing and implementing their own data sandboxes can identify the most appropriate strategies that best suit their use case.

Example	Data Flows	Control Level	Data Security	Sandbox Computing and Experimentation Elements	Temporality
London's Digital Sandbox Pilots	G2B; B2G **G** (City of London Corporation; Financial Conduct Authority) **B** (small to medium private enterprises)	Controlled (application for access)	Anonymized/ de-identified data Synthetic and real data Aggregated and individual data points	APIs Jupyter Notebook Integrated development environment (i.e. a 'Google Docs for code') Collaboration platform (messaging/ intersectional connecting) Observation deck for technical testing	12 months; program based
Statistics Denmark's Data Science Lab	G2G; G2B; G2C **G** (Statistics Denmark) **B** (Novo Nordisk Foundation; other SMEs) **C** (academic researchers)	Controlled (application for access; paywall to access data)	Real data		Ongoing

Lithuania's TravelTech Data Sandbox	G2B; G2C **G** (Lithuania Travel; Invest Lithuania; the Innovation Agency of Lithuania; Startup Lithuania) **B** (Bolt, Booking.com, Kayak)	Open access	Real Aggregated Anonymous	Data dashboard	Ongoing
The EIT Health Digital Sandbox	G2B **G** (EIT) **B** (Med/Health Tech firms)	Controlled (application for access)		Data sharing agreements Funding for start-ups Collaboration platform between start-ups, data holders, and mentors	12 months
The INDIGO Initiative's Sandbox for Open Data	C2G; C2B; C2C **G** (governments) **B** (SMEs) **C** (GO Lab; academic researchers)	Controlled with a move towards open access		Open data platform	Ongoing
Danish National Health Data Science Sandbox for Training and Research	C2C **C** (academic institutions)	Open to Danish academic institutions	Real Anonymous Synthetic	Teaching Materials (available in R and Jupyter Notebooks) Data hosted on Computerome and UCloud	Ongoing

The Europeana Foundation's Metis Sandbox	C2G; C2B; C2C **G** (Europeana) **C** (archival/ artistic institutions)	Open	Real Aggregated Anonymous	Data quality standards Europeana.eu data portal	Ongoing

Appendix C: Further Reading

The literature around data sandboxes is growing as the adoption and implementation of data sandboxes become more prevalent around the world. Below, we share some additional readings decision-makers can explore to learn more about the existing discourse on data sandboxes and further bolster their understanding of the concept. The readings are listed in alphabetical order.

- Arntzen, Siri, Zach Wilcox, Neil Lee, Catherine Hadfield, and Jen Rae. "Testing Innovation in the Real World: Real-world testbeds." *Nesta*, October 2019. https://media.nesta.org.uk/documents/Testing_innovation_in_the_real_world.pdf

- Attrey, Angela, Molly Lesher, and Christopher Lomax. "The role of sandboxes in promoting flexibility and innovation in the digital age." *OECD Going Digital Toolkit Note*, No. 2, 2020. https://goingdigital.oecd.org/data/notes/No2_ToolkitNote_Sandboxes.pdf

- Datasphere Initiative. "Sandboxes for data: creating spaces for agile solutions across borders." *Datasphere Initiative*, 2022. https://www.thedatasphere.org/wp-content/uploads/2022/05/Sandboxes-for-data-2022-Datasphere-Initiative.pdf

- Kwok, Wai Min. "Sandboxing and experimenting digital technologies for sustainable development." *UN DESA Policy Brief*, No. 123, 2021. https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-123-sandboxing-and-experimenting-digital-technologies-for-sustainable-development/.

- OECD. “Responding to societal challenges with data: Access, sharing, stewardship and control.” *OECD Digital Economy Papers*, No. 342, OECD Publishing, Paris, 2022. https://doi.org/10.1787/2182ce9f-en

- Outes Velarde, Juliana. “Bigger, better and more data for better social outcomes: what's next for the INDIGO data and learning collaborative?” *Oxford Government Outcomes Blog*, 2023. https://golab.bsg.ox.ac.uk/community/blogs/bigger-better-and-more-data-for-better-social-outcomes-whats-next-for-the-indigo-data-and-learning-collaborative/.

SMIT

Vlaanderen
verbeelding werkt

VUB VRIJE UNIVERSITEIT BRUSSEL

imec

Open.Data.Policy.Lab

opendatapolicylab.org