**A maturity spectrum for data institutions**

In times of crisis, data sharing is high on policy makers’ agendas. The European Union (EU) recently released a proposal for a data governance act (DGA) “to foster the availability of data for use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the EU.” Additionally, the United Kingdom (UK) Data Standards Authority recommended the use of new open standards to improve data sharing across the government, while the Information Commissioner’s Office published a revised version of its statutory data sharing code of practice to offer practical advice to businesses and organizations for ways to carry out responsible data sharing and launched a data sharing information hub. In the United States, federal agencies have started to execute a federal data strategy, a 10-year plan to leverage the value of information to support the delivery of missions and services across agencies, and a repository of federal enterprise data resources has been created. Canada is also revamping its privacy framework: the draft Consumer Privacy Protection Act aims to facilitate the sharing of deidentified personal information to carry out socially beneficial purposes.

In this article, we analyze data institutions as a means to enable data sharing. To be sure, not every data sharing instance requires setting up an ad hoc data institution. This article adopts a broad definition of data institutions, which covers what have been called data trusts, data foundations, data collaboratives, and data coops and which is aligned with the working definition proposed by the Open Data Institute: “organizations whose purpose involves stewarding data on behalf of others, often toward public, educational or charitable aims.”1

The focus on data institutions stems from the will to reconcile two objectives that are often considered antithetical: accelerating access to data and ensuring responsible data reuse. Data institutions are also conceived as a means to reconcile individual rights, particularly privacy and data protection rights, and collective interests in furthering the public good and solving challenges to benefit the wider community. What is more, data institutions are expected to help “democratize” data access, with a view to appeal to a wide range of users, including researchers and small- and medium-sized enterprises.

The introduction of the notion of “data altruism,” as found within the proposed DGA, can be seen as an attempt to overcome these tensions. Data altruism is defined as “the consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their nonpersonal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services.” Data altruism organizations are conceived as ad hoc legal entities that pursue objectives of public interest and operate on a not-for-profit basis as per Article 16 of the DGA. (Whether this solution is consistent with the EU data protection framework is unclear, as prior guidance on consent and transparency issued by the European Data Protection Board seems to suggest that purpose limitation and specification remain key requirements, even for processing activities that are in the public interest, despite their adaptation in the field of scientific research.)

Beyond the EU, at least two jurisdictions are worth exploring. The UK is particularly interesting in that, although the legislature has not stepped in yet, a few solutions have already been prototyped to identify best practices for sharing heavily regulated data, such as health and social care information. This trend has only been strengthened by the Covid-19 pandemic. The Secure Anonymized Information Linkage (SAIL) Databank is probably the best example to mention (for another example, see the International COVID-19 Data Alliance at <https://icoda-research.org/>) to which the SAIL Databank is a partner. The United States is also interesting, as the absence of a comprehensive data protection framework at the federal level would seem to suggest that such institutions are not needed in this context. The recent emergence of the Center for New Data (C4ND) suggests the opposite. (Canada and the provinces of Ontario and Quebec are also worth evoking. Since Sidewalk Labs’ proposal to build a civic data trust in Ontario, data trusts have received a lot of attention and sparked a debate, which has informed recent legislative initiatives.) This short contribution aims to propose a data institution maturity spectrum to assess the development of current and future initiatives in this space. It unpacks five challenges, or data governance goals, by giving examples of emerging data institutions in two jurisdictions.

**The Data Institution Maturity Spectrum**

The data institution maturity spectrum is built on five variables, as illustrated in Figure 1. This is because, generally speaking, data institutions help solve five challenges, which we conceive as the following data governance goals:



1. standardizing data access through shared data discovery functionalities and common curation standards
2. reducing reidentification risks
3. overseeing data reuse
4. demonstrating safe return
5. empowering citizens.

For each of these variables, it is possible to rank the sophistication of the approach implemented by each data institution, with a view to reach the data governance goal at stake. Let us start with the first data governance goal: standardizing data access. The more sophisticated the data institution is, the more services it will offer to data users. Data access services range from discovery to curation and visualization. The sophistication of the approach pursued by a data institution to reach the goal of reducing reidentification risks will depend on the range and variety of organizational and technical measures implemented by this bona fide, trusted third party. A data institution is usually in charge of minimizing the data, i.e., tailoring the amount of information to a predetermined and legitimate purpose and reducing reidentification risks. When robust data transformation techniques are coupled with organizational measures, a data institution ranks higher within this dimension.

As regards the goal overseeing data reuse, the sophistication of the approach will depend on the robustness of the monitoring and auditing function performed by a data institution. Demonstrating safe return, particularly at the local level, is considered key to giving data providers an incentive to share and citizens and local communities the evidence needed to fully endorse the approach. Being able to demonstrate system transformations places data institutions in the best position to achieve this goal. Finally, empowering citizens remains a central piece to legitimize the data sharing and reuse process, particularly when data pertain to individuals. Securing citizen representation at the decision-making level demonstrates a high level of maturity.

**The Five Data Governance Goals in Context**

**Standardizing Data Access Through Shared Discovery Functionalities and Common Curation Requirements**

Improving data discovery is often the number-one priority that drives immediate consensus and brings stakeholders together. Without coordination among providers, data discovery can generate substantial organizational overhead, as a variety of processes, including compliance, are necessarily duplicated. One key process that requires close collaboration is data integration, which consists of producing a unified description and thereby a cohesive set of metadata to enable seamless access to information. This makes it possible for users to reuse information. Rich contextual information about data often makes a difference when enabling users to access third-party data. Data curation thus goes beyond collecting and storing information in a shared facility. In fact, centralized storage is not necessarily a requirement. Data curation implies managing data sets to enable their reuse by a variety of stakeholders in such a way that data sets become easily findable, understandable, and accessible.

The SAIL Databank is a national data safe haven of anonymized health and administrative data sets about the population of Wales. Established by the Health Informatics Group at Swansea University Medical School, in 2007, and funded by the Welsh government’s Health and Care Research Wales, it is a relatively old data institution, as it has been operating for more than 10 years. Having started with health information, researchers have been progressively adding administrative data sets to their repository and catalog. Notably, they are now able to work with new data types, such as free text, images, and genetic information.

Founded in 2020 by Silicon Valley tech leaders, public health experts, and social entrepreneurs in the United States, the C4ND aims to “accelerate and democratize knowledge production from complex or gated data sets, in order to drive change on pressing policy issues such as pandemic response, inclusive redistricting, and equitable voting access.” One of the most interesting data discovery functionalities of the C4ND is that it goes beyond the creation of a data catalog by curating a comprehensive inventory of assets coupled with the capability to retrieve them. The C4ND built interactive dashboards to enable stakeholders who are not tech savvy, such as social scientists and journalists, to visualize sensitive data sets, perform fact checking, and, ultimately, interpret data for nonpartisan research and investigative journalism in a privacy-preserving manner. Examples include state-level dashboards for the organization’s Observing Democracy program and national dashboards for its COVID Alliance, which has been leveraged for public health reporting by the *Washington Post* and CNN.

**Reducing Reidentification Risks**

Reducing reidentification risks while preserving utility for analysis often requires adopting an approach that has been coined *functional anonymization* by experts in statistical methods, such as the UK Anonymization Network. At this stage, the main concern is to preserve data confidentiality. In practice, this means that data controls (i.e., those that directly impact the data view accessible to the user) and context controls (i.e., those that impact the broader environment, such as access control, contractual obligations, and organizational processes) must be combined. Controlling the data environment is an effective way to reduce the overall reidentification risk, as acknowledged by the medical data literature. However, the precise combination of data and context controls depends on specific use cases, and creating a controlled environment with real-time monitoring and auditing capabilities is often seen as a must-have. Data institutions are well placed to offer such environments in that they are able to offer both a technical infrastructure that is secure and trustworthy and data governance services, which include monitoring and auditing.

By way of an example, the SAIL Databank’s remote access system provides time-limited access to data sets and is subject to researcher verification, a data access agreement, and physical and procedural controls. Key security measures include enhanced user authentication, auditing all Structured Query Language commands, and configuring controls to ensure that data cannot be removed and transferred unless authorized. In addition, a variety of masking techniques is available, such as aggregation and suppression. The number of variables provided is tailored to a research purpose, and the project-specific encryption of key variables aims to prevent cross linkage where data users are involved in multiple projects. The C4ND removes directly identifying attributes; tokenizes device IDs; implements *k*-anonymization for visualization dashboards; applies attribute-based access controls, including project-based access control; and monitors and audits queries and user-level access to specific data sets and data elements. It has also commissioned penetration testing and external security audits.

**Overseeing Data Reuse**

To properly mitigate risks to individuals, it is a best practice to oversee the sharing and reuse of data, particularly when the information is very sensitive, such as with health care. This is because privacy is much more than confidentiality, and goals such as transparency, accountability, and fairness require iterative assessments. The same is actually true with confidentiality, as data products, such as models, can leak personal information. A substantial body of research has demonstrated that accessing a model’s structure and observing its behavior is enough to infer personal information. Data protection by design is best thought of as data protection by process. Of course, other fundamental rights and liberties could be at stake beyond the rights to privacy and data protection.

Data institutions are particularly well placed to oversee data reuse. A green paper produced by the UK Health Data Research Alliance (HDRA) advocates “providing access to data via Trusted (or Trustworthy) Research Environments (TREs) which protect—by design—the privacy of individuals whose health data they hold, while facilitating large scale data analysis using High Performance Computing that increases understanding of disease and improvements in health and care.”2 This is based on the five-safes framework, which is relied on by the UK Office for National Statistics: safe people, safe projects, safe settings, safe data, and safe outputs. The HDRA insists that safe settings should be built on the existence of the following set of minimum requirements:

* a system to hold data securely such that individual-level information cannot be exported.
* systems to enable secure remote access by accredited researchers to carry out analysis, with the ability to keep track of researcher activity and ensure that accounts cannot be shared
* a research environment containing a set of tools to enable data to be analyzed, with a barrier between the safe setting and the outside world to control data and software imports and exports
* processes and systems for the export of summary data and to support data and software imports; it will be necessary to implement systems able to scan data files, such as for viruses hidden in software packages and for identifiable data that should not be imported into a TRE.2

Safe computing is considered an extension of safe settings. Since computing infrastructure is being outsourced increasingly often, which can lead to partnerships with commercial organizations, it is crucial that third-party providers are not in a position to access personal information.

**Demonstrating Safe Return**

Demonstrating safe return means verifying the economic and societal impact of data reuse. “Economic and societal impacts” include all the ways in which data reuse can benefit individuals and communities. In many cases, securing them requires maintaining a close relationship with data providers, which often have a direct interest in leveraging users to solve a series of challenges aimed at transforming practices and systems. Securing the impacts also requires making sure the wider community of citizens is on board and able to appreciate the value of the impact that will be generated. Mature data institutions should thus set a defined purpose from the start and develop a set of metrics to capture and demonstrate impacts. Intellectual property management and mechanisms to ensure fair access to data are intimately linked with this objective and should ensure that collaborative efforts do not lead to exclusive appropriation.

More specifically, in the context of health-related research, the HDRA conceives safe return as a means to augment trusted research environments and make it possible to send results back to the clinical setting that provided the data, which would then be able to match results with personal identifiers for direct care purposes, clinical trials, and other research projects. Meanwhile, the C4ND’s objective is to generate instrumental and conceptual impacts. *Instrumental impact* refers to influence on policy making, while *conceptual impact* leads to the advancement of scientific methods and knowledge.

**Empowering Citizens**

Empowering citizens in a meaningful way is not straightforward. Data institutions appear well placed to take on this challenge and address it in two ways: by facilitating the exercise of individual rights and enhancing citizen involvement in decision making. In particular, by deploying decentralized data sharing infrastructures that are able to support dynamic deidentification (i.e., the production of a bespoke deidentified view of data for each reuse project), it is possible to put individuals in a position to exercise key privacy and data protection rights, such as the right to withdraw consent and to object to subsequent processing activities. (For an overview of the social data foundation model and an assessment of different deployment scenarios see “A Blueprint for a Social Data Foundation: Accelerating Trustworthy and Collaborative Data Sharing for Health and Social Care Transformation,”3 and “Building Trust Through Data Foundations: A Call for a Data Governance Model to Support Trustworthy Data Sharing.”4

In addition, by placing a data institution at the center of the sharing and reuse ecosystem and making it the decision maker for which reuse project to approve, not only is it possible to strengthen citizen information by building a one-stop shop but it is also possible to secure the involvement of citizen representatives to develop engagement programs. Interestingly, the SAIL Databank has opted for an advisory model and the creation of a consumer panel consisting of members of the public who represent a variety of interests. Their task is to advise on data protection issues and directly engage with data users, such as researchers.

With this said, data institutions have a long journey ahead of them to be able to claim that they can guarantee both transparency and accountability and that they enable effective civic collaboration. As the literature on the open data government movement shows, data sharing and reuse do not necessarily generate enhanced collaboration between providers and citizens or increased accountability for the policies and programs that data influences. It has been observed, however, that the rise of new “infomediary business models, positioned between open data providers and users,” although not perfectly, has managed to strengthen citizen representation.5

To conclude, the data institution movement should be welcome. It is much more nuanced than the release-and-forget model that is often associated with the open data movement. While a series of iterations will be needed to refine governance structures, the maturity of data institutions can be assessed and compared using a 5D spectrum, as illustrated in this article.