OPEN FOR FEEDBACK

Comparing decentralised data publishing initiatives

Open Data Institute

Contents

About

This report has been researched and produced by the Open Data Institute, and first published as a draft in October 2020. Its lead author is Leigh Dodds, with contributions from Josh D'Addario, Jack Hardinges, Elea Himmelsbach, James Maddison, Emily Sinclair and Jeni Tennison.

We would like to thank those who have supported our research and provided initial feedback on the report and the comparison of initiatives, including representatives from OpenActive, Open Contracting, OpenOwnership and 360Giving.

If you want to share feedback by email or would like to get in touch, contact the "Data Infrastructure for Common Challenges" project team at research@theodi.org.

To share feedback in the comments, highlight the relevant piece of text and click the 'Add a comment' icon on the right-hand side of the page.

How can it be improved? We welcome suggestions from the community in the comments.

Introduction

Improving healthcare, switching to renewable energy and responding to crises and disasters are all difficult problems that need to be solved through collaborative approaches. No single organisation has the resources or skills to solve the problems. Or the understanding of how to create solutions that work for everyone.

Multi-stakeholder initiatives are increasingly being set up to do things like advance research or tackle sustainability goals. For example, the UK's Industrial Strategy is geared around four 'Grand Challenges'[1], underpinned by missions that will be tackled through cross-sector collaborations. Businesses are working together to create more resilient supply chains. Local communities are using 3D printers and openly licensed designs to manufacture personal protective equipment to fight the Covid-19 pandemic.

Data can help us to understand and address these challenges. It can help to improve decision making, drive innovation and measure progress. Accessing, using and sharing data are frequent activities within many of these new initiatives.

However, our data infrastructure is often poorly designed or managed. This limits our ability to maximise value from data and opens the door to harmful impacts from its use. Programmes to strengthen and build data infrastructure have become a necessary activity in several of these wider collaborations.

We refer to these programmes as 'data access initiatives'. We describe them as programmes which:

- have a clear challenge, in the form of a specific social, environmental or economic problem that is the focus for the collaboration
- involve multiple stakeholders actively working together to solve the problem
- include a strong focus on collecting, using and sharing data as part of their work.

Over the last few years the ODI has been exploring a range of data access models, including data institutions, data trusts, data observatories and data collaborations[2]. This paper is part of a project exploring how data access initiatives are building data infrastructure to support their work[3].

We have identified one common design pattern for increasing access to data through the adoption of open standards for data. For the purposes of this report we are calling these 'decentralised data publishing initiatives'.

This report provides a short summary of this design pattern and compares how it is being applied across 14 different initiatives. Our intention is to share insights

[1] BEIS (2019) 'The Grand Challenges',
[2] ODI (2019) 'Mapping the wide world of data sharing',
[3] ODI (2020_ 'Data infrastructure for common challenges'

into when and where it may be appropriate to use this approach to increasing access to data, how to make it successful, and the types of initiatives that might benefit from it.

What is a decentralised publishing initiative?

Data access initiatives are programmes that:

- **involve multiple stakeholders, collaborating to address a common challenge**
- **are increasing access to or use of data**.

To define 'decentralised publishing initiatives' we have chosen to build on this definition by adding additional elements.

A 'decentralised data publishing initiative' is a data access initiative in which:

- **data is published in a decentralised way**: the data providers make the data available via their own infrastructure
- the **data providers are publishing data about the same kinds of things,** for example spending data.
- the data is **shared** or **open data**
- a **single, common standard is used by all organisations** such that the data is published in the same kind of way
- the **initiative provides guidance, tools and technology** to support data publication and use, for example a central register of datasets to support discovery

Examples of decentralised publishing initiatives

From our initial research we have identified several initiatives that display these characteristics. The examples include:

1. **OpenActive** – UK activity providers such as gyms publish live data feeds for opportunities to be physically active, such as spin classes, under an open licence, using a common data model and application programming interface (API) standard
2. **Open Contracting** – governments around the world publish openly licensed datasets describing public procurement tenders and contracts
3. **Open Banking** – major banks in the UK publish open data about banking products as well as shared data about bank transactions using a common data model and set of API standards
4. **LIVES** – municipalities in New York and San Francisco publish data about restaurant inspections
5. **UK Bus Open Data Service** – bus operators in the UK publish data feeds about bus timetables

For more examples and a comparison, see the section: 'How do some existing initiatives compare?'

Examples of initiatives that do not share these characteristics

In reviewing a wide range of data access initiatives we identified many that share some of the characteristics of a decentralised data publishing initiative, but vary from the core pattern in different ways.

First, in some initiatives, while the data may conform to a common standard, it is being published or shared via a single central portal, repository or platform.

For example the Europeana project, which helps cultural heritage organisations share data about their collections, asks those organisations to submit data using a common standard[4]. The data is then made available via Europeana.

This shift from a decentralised to a centralised approach to publication and use creates a very different data ecosystem. It may happen by design or through one aggregator or intermediary coming to dominate the ecosystem. Or it may arise because publishers lack capability: a centralised infrastructure is being used to overcome issues with them providing their own data infrastructure to consistently and reliably share open data in a trusted way.

Second, common standards might be in wide use across a sector, industry and community, but are being used to publish data covering a diverse range of different subjects. This is often the case when the underlying standards define general purpose data exchange formats or API designs, rather than common schemas or data models.

For example, the OpenGeospatial Consortium works with a variety of stakeholders to create and support adoption of geospatial data standards. However the organisations adopting those standards are publishing a very wide range of different types of geospatial data, for a variety of different purposes.

Not all programmes to develop and support adoption of an open standard are data access initiatives or decentralised data publishing initiatives.

Third, there are many examples of organisations publishing similar datasets, where there is no common standard being used to publish that data.

In these cases there may not be a data access initiative that is coordinating this activity. Publishers do not see themselves as participating in a common initiative with a shared goal or vision. This may point to a need for a new initiative.

For example, many retailers provide information about their store locations or products, but there is no common standard or approach for making that data available even though it is relatively consistent across different websites.

In other cases there may be a shared initiative, but it is not working to build or strengthen the data infrastructure necessary to increase quality, consistency and support use of the data. This might highlight a need for the initiative to broaden its activities.

[4] Europeana (n.d.), 'Process',

For example, the CovidSecureCheck[5] project is building a register of Covid-19 risk assessments to gather evidence and support analyses of approaches being taken. However it doesn't recommend that organisations use a specific standard format or approach for publishing those assessments.

Like any attempt to apply a new definition to the real world there are many examples that are more or less closely aligned with our definition of decentralised data publishing initiatives.

Our goal in this paper is to outline the general pattern so we can compare how it is being used and implemented in practice. The specific design decisions behind individual initiatives are often more instructive to consider than refining a formal classification.

Discussing edge cases and the different priorities and choices that underlie them is a useful way to explore how the design of data infrastructure reflects the ecosystem in which it exists, and the purposes for which it is being created.

[5] TUC (2020) 'CovidSecureCheck',

Where is this pattern applicable?

Building a data infrastructure around decentralised publishing of data seems to be applicable when one of more of the following preconditions apply:

- the **data is naturally distributed**, for example it is being collected and managed by a large number of different organisations
- there are **multiple similar organisations with a shared purpose**, such as a public task or similar business model, which are collecting or producing the same type of data
- there is **utility in using both original data sources and aggregated data**. In some cases consumers may benefit both from directly accessing data from a single provider, with extra insights or use cases being supported by use of aggregate data across multiple publishers

A decentralised approach may have a range of benefits, including:

- **removing central costs for data collection and management**, which will be higher where there are a large number of publishers who may need to contribute, or where the volume of data to be collected is significant
- **increasing timeliness of access to data**, for example if the data is regularly updated or published and creating an intermediary would slow down the publishing of that data
- making data available from source, rather than indirectly via an intermediary may **reduce risks**, **increase trust** or **clarify the provenance of data**.

Our other research projects on sustainable data access[6] and building trust[7] explore these topics in more detail.

[6] ODI (2020) 'R&D Sustainable Data Access',
[7] ODI (2020) 'R&D Building trust through certification and audit"',

How are decentralised initiatives building data infrastructure?

Data infrastructure[8] consists of:

- data assets, such as datasets, identifiers, and registers
- standards and technologies used to curate and provide access to data assets
- guidance and policies that inform the use and management of data assets and the data infrastructure itself
- organisations that govern the data infrastructure
- the communities involved in contributing to or maintaining it, and those who are impacted by decisions that are made using it.

There are many different activities that support the development, maintenance and use of data infrastructure. For decentralised data publishing initiatives, these activities include:

Standards

- Scoping, development, publication and ongoing governance of an open standard for data, used by all those publishing data covered by the initiative.
- Providing feedback to publishers and data users about whether published datasets conform to the standard. This might take the form of individual validation reports or conformance indicators in a registry.

Data assets

- Maintaining a registry of published data. This might be a list of organisations publishing data, or links to individual datasets. The registry might be maintained by the initiative or contributors.
- Harvesting and collecting published datasets and making them available as a single aggregated dataset either for download or use via APIs, in addition to them being available directly from source.

Technology

- Developing and releasing tools that can be used by data publishers to support them in publishing data that conforms to the standard. Might include spreadsheet templates, API frameworks, anonymisation tools, or other technology.
- Developing and releasing tools that can be used to validate data against the open standard, to check conformance with the specification.

[8] Open Data for Development (2019) '[Data infrastructure]'

Guidance and policies

- Publishing guidance for data publishers about how to publish necessary data. Might include technical guidance on standards, as well as non-technical information.
- Providing support for policymakers, regulators and others in developing legislation, procurement guidance, regulation or other policies that enforce, inform or support the publication of data according to the standard.

Communities

- Carrying out user research and engagement to understand the needs of data publishers, users and other stakeholders. Might involve formal user research methods, interviews or informal collection of needs.
- Providing a help desk or similar support function to allow data publishers and users to request help and guidance on publishing or using data.

A comparison of existing initiatives

Using the activities identified through our analysis (see Methodology), we have compared 14 different initiatives. They vary based on the following:

- The **licensing** of data. We have seen examples of both data sharing and open publication of data.
- Their **geographic scope**. We have seen examples of regional, national and international initiatives
- Whether data sharing and publication of data is **mandatory or voluntary.**
- The **sector or domain** in which the initiative is focused – we have seen examples in transport, health, finance, public policy and physical activity.
- The **type of challenge being tackled**. Some initiatives focus on tackling transparency, others aim to drive innovation in a sector

Comparing decentralised data publishing initiatives

View the result of our comparison here.

The spreadsheet includes:

- a brief summary of each initiative
- a list of activities associated with creating or maintaining data infrastructure
- an indication of whether the individual initiative is carrying out those activities.

How do the initiatives compare?

The comparison identifies a number of similarities and differences between initiatives which are helpful to highlight.

Addressing gaps in data infrastructure

By definition all of these initiatives involve driving adoption of a standard. However in the majority of cases these initiatives are scoping, developing and governing new standards, rather than driving adoption of existing standards.

This suggests that these initiatives are typically addressing a gap in existing data infrastructure.

Validation and ranking

While the majority of initiatives appear to provide feedback to publishers on the quality of published data, the actual approaches vary. Some initiatives provide direct, private feedback to publishers while others produce public reports to assess conformance.

Some initiatives or their communities go a step further and produce an index or ranking of publishers based on the published data. This seems to be more common around initiatives whose purpose is related to transparency and compliance.

Provision of aggregated datasets

While all of the initiatives are maintaining a public registry of datasets to support discovery of the underlying data, only half of the initiatives are producing an aggregated dataset and/or an API.

For some initiatives an aggregated dataset would may not be possible because the underlying data is shared and not open data.

The lack of an aggregated dataset in other circumstances may be due to a range of factors:

- There is more utility in accessing the source data, rather than analysing data in bulk.
- The number of source datasets may be small enough that investing in an aggregation is not useful – data consumers can easily access what they need.
- Participation in the initiative remains low, or it is still at an early stage and developing an aggregation is part of a future roadmap.
- Data consumers have a variety of different needs so a ‘one size fits all’ aggregation may be difficult to create.
- The initiative does not have the technical or financial resources to maintain the aggregate dataset.

Few initiatives are providing tools and technology to support data consumers in aggregating or harvesting data. For those initiatives that are not providing an aggregate dataset, providing additional tools for data consumers may be a gap to be addressed.

Data hosting services

A small number of initiatives offer a data hosting service to provide an alternative to publishers publishing data themselves. For example the UK Bus Open Data service offers to host data for bus operators who have fewer than 40 routes.

This is likely to be a pragmatic approach aimed at increasing participation where some publishers may not have the resources to publish data directly. As noted previously, if all publishers use this infrastructure, it is no longer a *decentralised* data publishing initiative.

Skew towards open data

While some of the initiatives we reviewed are using the decentralised data publishing approach to support sharing of data, there is a clear skew towards open publication of data.

The decentralised data publishing approach is likely to be simpler to implement where data can be published openly.

Much of the groundwork of developing standards, technology and guidance is similar across initiatives regardless of licensing. But where data is shared with restrictions there is additional work required to define, for example, common approaches to data governance and shared API standards that support secure access to data within the terms of the restricted licence.

Next steps

This report summarises our current understanding of how decentralised data publishing initiatives are creating data infrastructure to tackle a range of challenges. We invite feedback to help broaden the range of initiatives being compared and to help develop further insights into the benefits and challenges of applying this approach.

As part of this research project we will be developing and recommending tools and guidance that may help those who are leading or planning initiatives in improving their approach to creating and maintaining data infrastructure.

If you would like to be involved in this project, then please contact the 'Data Infrastructure for Common Challenges' project team at research@theodi.org.

Methodology

To develop this comparison we went through the following process:

1. We created a ‘long list’ of data access initiatives, drawing on those we have worked with, supported and led. This was supplemented with new initiatives identified through wider desk research.
2. Drawing on interviews and desk research we developed logic models[9] for a small number of initiatives. This helped us clarify the activities of those initiatives.
3. We classified the activities to identify those that were related to the development, maintenance or adoption of data infrastructure.
4. Having identified that a subset of initiatives shared some common characteristics, we created a definition of the ‘decentralised publishing initiatives’ pattern and a list of related activities.
5. Using this definition and list, we compared a range of decentralised publishing, drawing from our original long list with additions that were crowd-sourced from social media.
6. We requested feedback from some stakeholders directly involved in the relevant initiatives to help improve the accuracy of our comparison. Any remaining gaps or misunderstandings are our own.

It is important to note that our focus has been on data infrastructure. Successful initiatives will be creating or adopting other types of (digital) infrastructure – for example, collaboration or video conferencing tools, or other software and services that are helpful in tackling their challenge. Exploring those aspects are out of scope for our project.

Data access initiatives will also be involved in a broader range of activities that help to manage, sustain and grow the initiative, but which are not directly related to building or maintaining data infrastructure. Examples include fundraising, recruitment, communications, etc. Documenting these activities is also out of scope for our project but they are obviously important in delivering impactful programmes.

We have not assessed all of the details of how the compared initiatives are carrying out the work of creating data infrastructure. Each initiative may be approaching these activities in different ways and with different levels of investment. Our goal is to highlight similarities and differences to prompt discussion and review.

Where there are other gaps, we invite feedback to help us improve our analysis.

[9] PHE (2018) ‘Introduction to logic models’

Limitations

We are aware that there are several limitations in our survey and initial comparison.

- The set of initiatives that we identified and shortlisted for review are biased towards those known to us through our network and broader engagement. We invite suggestions on further initiatives to include: we would particularly like to identify further initiatives from the Global South, as well as a broader range of domains.
- The initiatives we have reviewed are still active and have a public presence. There may be local initiatives that we have not identified. There may also be 'survivor bias' in the analysis, as those we can easily review are still active and successful enough to be easily identified
- The activities we have identified are those that we have been able to find through desk and user research, as well as some direct feedback from those involved in the initiatives. There may be other activities that we have overlooked or which should be more prominent in our analysis.
- Like all other programmes of work, data access initiatives may cause harm or fail to promote diversity, equity and justice. The activities and outputs they create may not always demonstrate good practice around ethics, equity or inclusion. The activities identified in our comparison are a starting point for discussion and improvement, rather than a definitive list of good practice. We invite feedback on the identified activities and suggestions on how to conduct them in ways that minimise harm and tackle inequities.