Open Data Portal Management Framework

From set-up to sustainability

MBA Thesis

By

Anne Scanlon

Professional MBA Digital Transformation and Data Science

12-Jul-2021

Supervisor

Dr. Sabrina Kirrane

Acknowledgements

I would first like to thank my thesis supervisor Dr. Sabrina Kirrane of the Institute for Information Systems & New Media, Vienna University of Economics and Business. Without Sabrina's advice, feedback and support, this thesis would not be the work that it is today. She consistently allowed this paper to be my own work but steered me in the right direction whenever needed.

I would also like to acknowledge the support offered by my employment to allow for the time to research and develop this thesis.

Finally, I must express my very profound gratitude to my spouse, family and friends for providing me with unfailing support and continuous encouragement throughout my study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

Anne Scanlon

Abstract

Data is the foundational layer to so many opportunities today, powering digital transformation and innovation. Opening access to data under clear usage terms can foster innovation and improved efficiency, promote transparency and accountability, and stimulate engagement. Whilst the opening of access to government data has been a trend for over a decade, more recent developments have seen multilateral organisations recognizing the value in opening their data in a drive to fulfil mandates, promote transparency of their programmes and bring efficiencies.

However, Multilaterals face particular challenges when it comes to opening up their data. A lack of consistent data management practices coupled with limited resources and high personnel rotation can prove a barrier to sustaining an Open Data initiative. The hypothesis of this thesis is that having a guiding framework in place, informed by Open Data and FAIR data principles, will ensure a consistent repeatable approach to managing and publishing Open Data, which, in turn, ensures sustainability.

This thesis seeks to create an actionable framework for setting up and sustaining an Open Data Portal. Following a design science research approach, we iteratively develop the framework and test by applying elements in a use case implementation, representative of any multilateral organisation. In answer to the research questions, we derive an aggregation of principles into assessment categories, a list of requirements or desired features to implement in an Open Data portal and a list of components to action as part of a set-up and sustainability plan.

Applying these three elements of the framework, an Open Data Portal is set up and a minimum number of datasets are migrated. Any reflections on the framework are documented for potential incorporation into the next cycle. The framework could be enhanced further by adding a stronger focus on data quality and taking metadata and linked data research into consideration.

Table of Contents

List of Figures

List of Tables

1 Introduction

Data makes the world go round, or at least data helps the world go round (Fraizer, 2021). Forbes considers data the fuel that is powering digital transformation[1], which coupled with advances in digitization and technological innovations, is transforming the socioeconomic world (Strohmaier, et al., 2019). According to a McKinsey report[2] on the topic, leveraging data is impacting industries from transport to healthcare, from top-line, customer-facing activities to bottom-line internal processes such as supply chain optimization, predictive maintenance, fraud prevention. In the healthcare industry, integration of biomedical and healthcare data, has the potential to revolutionize medical therapies and personalized medicine (Dash, et al., 2019). Data has also transformed the field of management, enabling data driven, evidence-based decision making in businesses (McAfee & Brynjolfsson, 2012). It is no surprise that data has been referred to as "*feedstock for the world economy*"[3]. The private sector is not alone in leveraging data as an asset. Governments and public sector organisations are recognizing the value of their data (Zeleti, et al., 2016), in creating opportunities for transparency, innovation, public participation and efficiency (Christodoulou, et al., 2018).

The practice of opening data, that is, making data available and accessible so that anyone can access, use or share it, has grown tremendously over the past two decades (Hook, 2021). Opening of government data streams was encouraged by former United States President Obama in his memorandum on "Transparency and Open Government" (McDermott, 2010). European legislation has been in place since 2003 to facilitate the availability and re-use of public sector data in the form of a Public Sector Information (PSI) directive 2003/98/EC[4] which was further amended in 2013 to emphasis opening of data. Since then, Open Data, as it has been named, has been widely adopted by governments to proactively disclose public data instead of waiting for requests from citizens.

[1] Data: The Fuel Powering AI & Digital Transformation (forbes.com)
[2] Achieving business impact with data | McKinsey
[3] Data is giving rise to a new economy | The Economist
[4] Revision of the PSI Directive | Shaping Europe's digital future (europa.eu)

Open Data serves many purposes. These can be summarized in three objectives (Charalabidis Y., 2018):

1. To promote "*transparency and accountability*" around how governments or organisations function.
2. To promote "*innovation and improved efficiency*"; as making data readily available and reusable can encourage innovation, adding value or generating new products.
3. To stimulate *"engagement and participation"*, broadening the communication channels and allowing feedback on activities.

The provision of Open Data also promises many social, political and economic benefits (Kucera & Chlapek, 2014). With expected benefits including stimulating innovation and promoting economic growth (Wieczorkowski, 2019), it is understandable that Open Data is of interest to not just governments (Attard, et al., 2016) and the public sector (Below, 2015) but also to the private sector and multilateral organisations. The latter are international organisations formed between three or more nations to work on issues that relate to all the countries in the organisation. They include the United Nations (UN) entities and development banks such as the World Bank Group and the Asian Development Bank (ADB).

According to Hammer (2019), multilaterals can provide the medium- and long-term support needed for social and economic development impacts from Open Data. An increasing number of multilaterals have joined the International Aid Transparency Initiative[5] to promote transparency and accountability. They are opening up their data from projects and activities by developing internal and external information-sharing platforms.

However, along with the benefits of opening data, there are also challenges (Martin, 2014). The main challenge of Open Data is that the data by itself has little value; with value generation coming from its use and re-use (Janssen, et al., 2012). It is therefore important to facilitate a mechanism for data owners to make their data available in a manner that promotes usage opportunities for data consumers (Kučera, et al., 2015). Information sharing platforms or Open Data portals are software systems that are typically employed to connect

[5] About IATI | International Aid Transparency Initiative - iatistandard.org

the data publisher with the data consumer, and ideally should allow maximum access and re-use of data, in order to benefit from the data being released and opened.

1.1 Problem Formulation and Motivation

The work described in this thesis is guided by a motivating use case scenario relating to a fictitious organisation, OrgX, that is representative of any multilateral organization. OrgX is funded by its Member States to run programmes that deliver the mandate of the organisation, across a variety of topics, such as energy, health, safety and security, food and agriculture and climate change. For example, a human health programme could be to apply scientific techniques to prevent, diagnose and treat diseases. The programmes are achieved by implementing a series of projects which deliver outputs and outcomes such as publications, training materials, websites, web applications and datasets. Over the years, the data collected on these programmes and outputs has grown steadily.

OrgX is also a funder of research and capacity building projects. For research projects, OrgX accepts proposals from research institutes and bodies and provides funds to implement the approved research projects. The capacity building projects are a mechanism to transfer expertise and technology to address key development priorities in Member States. Data is generated from these types of projects, both financial administration data, and subject-specific outputs from the project's research and development. There is a global initiative to improve the transparency of resources spent on development projects[6] and a growing pressure from the Members States to release data on the programmes as Open Data.

OrgX, is faced with many challenges and opportunities when it comes to managing the data from its various programmes and projects.

Lack of a consistent approach

In the absence of a central data portal, programme managers at OrgX have created multiple stand-alone websites and applications for publishing their programmatic deliverables and sharing data with Member states and the public. There is no consistency or standards in how the data is published and presented, which formats are used or how the data is described. The websites typically present the data in non-interactive fashion, with neither data

[6] International Aid Transparency Initiative - iatistandard.org

download features nor access to the raw data. The data is not explicitly shared under a licence agreement that is clear to those looking to innovate. Reuse is governed by a generic "Terms of reference" more suitable to web content rather than data. Little user input is available on how and where these data sources are being used and for what purpose.

Lack of data management practices

There is no overarching data strategy or plan in place on how to manage data during the programme life cycle, from creation to archival. For research projects generating new data, there are no data management guidelines. For example, as a funder of research projects, OrgX does not request a data management plan (DMP) as part of the project proposal submission process, yet this is considered best practice by other funding organisations such as the European Commission.

Limited resources

OrgX has limited financial and human resources. There are limited financial resources for spending on data management and the required supporting information systems and these elements are frequently not budgeted for as part of the programme budget. In fact, there is pressure to reduce overall costs for development and maintenance of IT systems.

As part of OrgX policies, staff rotation is frequent, resulting in a lack of continuity in terms of the subject matter and data publishing expertise. Therefore, focussing on sustainability is extremely important when it comes to establishing a new data management practice or portal.

The motivation of this thesis is to provide organisations like OrgX with a path or guiding framework to follow to address the challenges outlined above.

1.2 Objectives of Master Thesis

We propose an actionable framework to facilitate the creation and maintenance of an Open Data portal which will address the challenges outlined above by promoting:

- A consistent approach to data sharing and data management,
- Improved data management services to data providers for publishing data, and to Member States and the public for consuming the data,

- Cost efficiencies by streamlining onto one platform, thus reducing maintenance costs,
- Sustainability of the initiative by taking an iterative approach.

Therefore, the main objective of this thesis is to design and develop an approach in the form of an Open Data Portal Management framework for establishing and sustaining an Open Data portal and data, based on a literature review and assessment of the features of successful Open Data portals. To validate aspects of the framework, we implement one cycle of the proposed plan and evaluate the results.

1.3 Research Questions

Given our main objective, the overarching research question posed is, what principles, guidelines and plans should we include in an Open Data Portal Management framework to ensure a sustainable Open Data portal that can address the data management challenges of organizations like OrgX?

To answer that, we further define more specific research questions:

What are the guiding principles for opening or sharing data?

How have these principles been applied in practice and to what extent in existing Open Data portals?

What are the key elements of plans for data management and sustainability?

1.4 Research Methodology

To address our research questions, we sought out an iterative research approach that complements Information system development, and selected the design science research methodology (Vaishnavi & Kuechler, 2004). Design science research is research specifically undertaken to support the strategic design and development of an artefact, which can be products and/or programs. Hevner et al. (2004) described the approach as characterized by its intention of improving a current state, where knowledge and understanding of a problem domain and its solution are achieved in the exploration, building and application of the designed artefact. It is a suitable strategy for our research, due to the iterative and exploratory nature where knowledge may be gradually and collaboratively developed through the design process. This problem-solving process is depicted in *Figure* 1 and is described in more detail below.

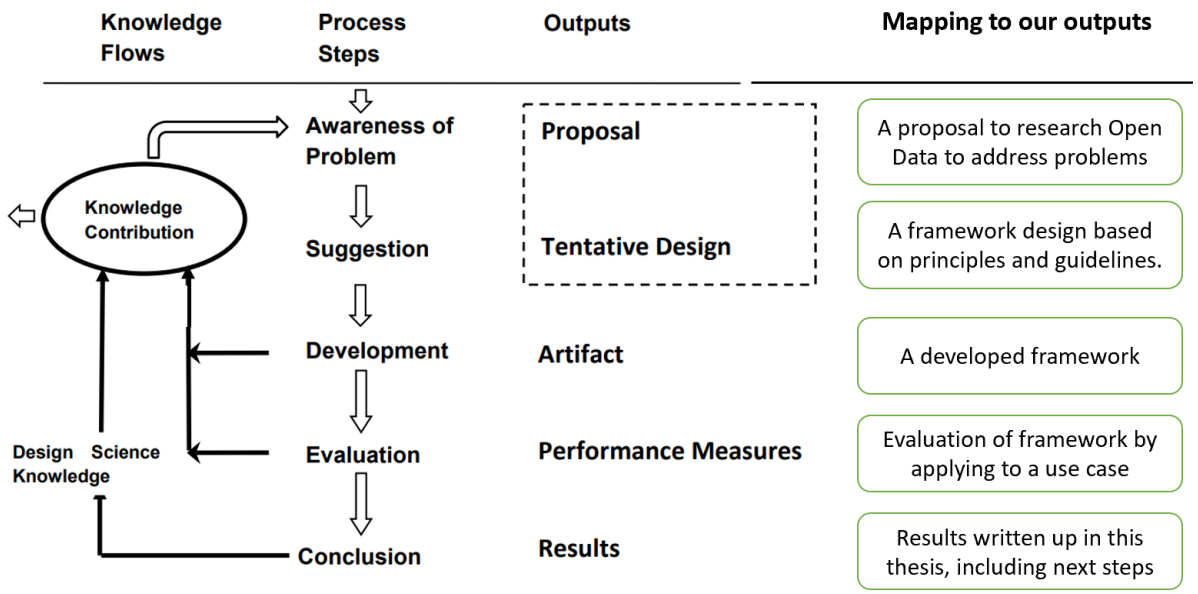

Figure 1: The design science research general methodology, adapted from (Vaishnavi & Kuechler, 2008)

Awareness of problem

The process starts with an awareness of the problem, in our case, an awareness of the problem of managing and opening up data in multilateral type organisations. The output from this process step is typically a proposal for a new research effort. For us, this translates into a proposal to research Open Data and Open Data portals for applying to the problem.

Suggestion

The suggestion phase follows immediately behind the proposal and is closely connected with it, hence the reason for the dotted line around the outputs from this and the previous process steps. This creative step outputs a tentative design in which new functionality is envisioned based on either existing or new and existing elements. In our case, we are suggesting an Open Data Portal management framework and the output from this step is a design of a framework based on principles, guidelines, and best practices.

Development

The Tentative Design is further developed and implemented in this phase, outputting an artefact. The novelty is primarily in the design, not the construction of the artefact. Thru a

process of integrative literature review and an assessment on how principles are applied in organisations like the use case, the framework is developed further.

Evaluation

In the next phase the artefact is evaluated. In our case, the framework is evaluated by applying it to the use case of OrgX and establishing an Open Data portal proof of concept. Feedback is collected through focus groups with key stakeholders. The results from the evaluation phase and any additional information gained in the construction and running of the artefact are brought together and fed back to another round of Suggestion. This build-and-evaluate loop is typically iterated a number of times before the final design artefact is generated.

Conclusion

This step marks the end of the research effort, when it is acknowledged that whilst more iterations and refinement are possible, the artefact is deemed good enough. Results and any research contributions or knowledge gained are written up.

2 Open Data

This chapter presents an overview of Open Data principles. It begins by providing a history of Open Data and the various initiatives and movements through to current day. It summarizes Open Data principles and policies from different sources and then compares the various principle schemes, identifying distinctions and commonalities. The chapter concludes with a consolidation and aggregation of all the principles into potential assessment categories.

2.1 Terminology

Before getting into the history and the principles of Open Data, we begin by defining Open Data and some related concepts and terms used within this thesis. *Figure 2* depicts the relationship between Open Data, government data and linked data.

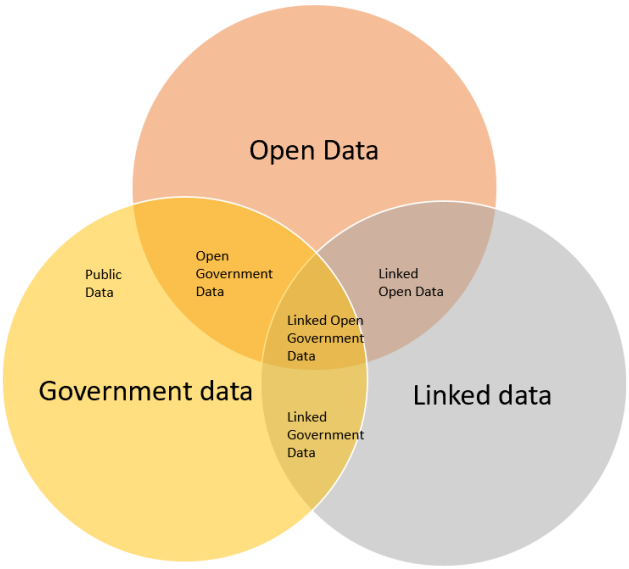

Figure 2: Relationship between Open Data, Government Data and Linked Data

Open Data

The International Open Data Charter[7] defines open data as "*digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere"* (International Open Data Charter, n.d.).

[7] The International Open Data Charter

The Open Knowledge Foundation (OKF)[8] also published a definition of Open Data, entitled the Open Definition[9]. It defines openness in relation to content and data as "*Open data and content can be freely used, modified, and shared by anyone for any purpose*" (OKF, 2015).

Government Data

Government data is data produced or commissioned by government or government-controlled entities.

Open Government Data (OGD) is a subset of government data and Open Data. It is data that is produced by government bodies and is open, according to the Open Definition criteria; as in, it can be freely used, reused, and shared by anyone[10].

Public data can be defined as information in the public domain. It is data that is made freely available to the general public but it not necessarily accessible, structured or open. It could for example be offline data that are only accessible via Freedom of Information requests. It is not subject to valid privacy, security or privilege limitations.

Linked Data

The term Linked Data, coined by Tim Berners-Lee, refers to connecting structured data on the web following four rules. These are to:

1. Use URIs (Universal Resource Identifiers) to refer to data
2. Use HTTP URIs so that it is discoverable on the web,
3. Use standards to serve the data such as RDF or SPARQL and
4. Include links to other data so as to build on the web of data. (Berners-Lee, 2009)

Linked Open Data (LOD) is linked data that is released under an Open Licence which allows its reuse (Berners-Lee, 2012).

Linked government data is government data that is not open but is linked over the web.

---

[8] Home | Open Knowledge Foundation (okfn.org)
[9] The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge
[10] Open Government Data

2.2 Background on Open Data

The concept of Open Data is not new. The principles behind openly publishing information and/or data and the philosophy of transparency have been around since the early years of the European Enlightenment. As early as 1665, the Philosophical Transactions of the Royal Society journal, required that contributors had to include evidence (the data) on which the article was based. This sharing of data enabled peer reviews, replication of experiments and reuse of the data (Boulton, 2014).

However, it is in the past 20 years that Open Data has really come to the fore, as we can see from *Figure 3.*

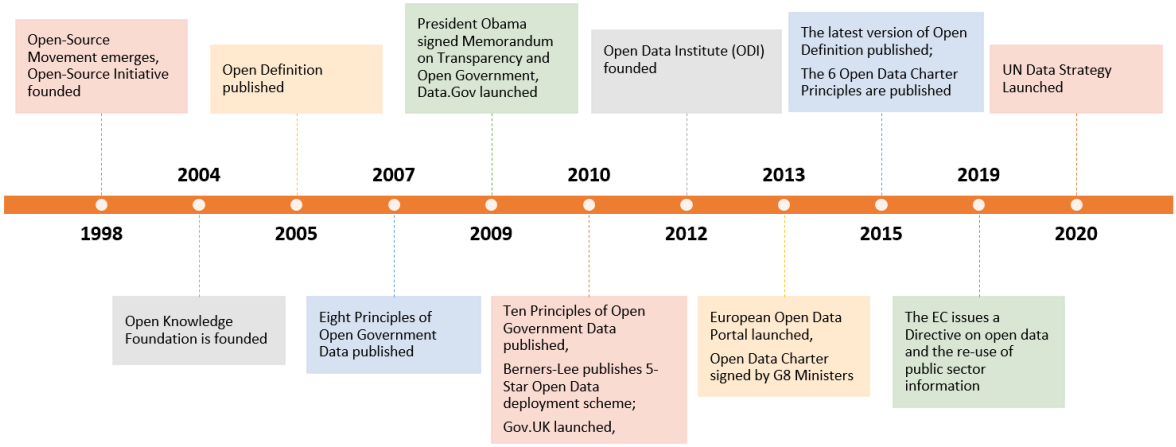

Figure 3: Timeline of Open Data

Towards the end of last century, the term Open Source Software was coined to describe the growing movement promoting the open development process and open access to code. In 1998, the Open Source Initiative (OSI)[11] was formed to execute the vision of Open Source and promote the usage of open source software.

Drawing direct connections with the Open Source and Open Access movement, the Open Knowledge Foundation[12] was founded in 2004 in the UK with the mission to "*create a more open world – a world where all non-personal information is open, free for everyone to use, build on and share; and creators and innovators are fairly recognised and rewarded*" (Open

[11] Open Source Initiative - News | Open Source Initiative
[12] Home | Open Knowledge Foundation (okfn.org)

Knowledge Foundation, n.d.). The OKF published the Open Definition[13] in 2005, explicitly defining Open Data.

Concurrently in the United States (US), initiatives were taking place strengthening the concept of open government data. Some US federal government agencies and civic administrations had begun making their data available online but with inconsistent approaches. In December 2007, thirty open government advocates gathered in Sebastopol, California and wrote a set of eight principles of open government data, with the aim of facilitating the evaluation of openness and accessibility of government data. The initiative was funded by Yahoo, Google and the Sunlight Foundation[14], a non-profit organisation, with the mission to make government and politics more accountable and transparent to all. This foundation went onto expand the list to ten principles in 2010.

Open government data was also being promoted at the highest level of office with US President Barack Obama signing "The Memorandum on Transparency and Open Government" (Obama, 2009), officially paving the way to make open and machine readable data the default setting for government. The US Government data portal, Data.gov was launched that same year with just 47 datasets. At the time of writing, it has over 300,000 datasets.

Berners-Lee, who defined linked data as early as 2006, published the 5-star deployment scheme for Open Data[15] in 2010, to describe the incremental steps to reach linked open data. That same year, he contributed to the launch of the UK's first public data site for open government data.[16] In 2012, he co-founded the Open Data Institute[17] with Nigel Shadbolt. This non-profit organisation's mission is to demonstrate the value of open data and advocate for its use by working with companies and governments (ODI, 2012).

In 2013, the European Open Data portal[18] was launched as an access point to information held by various open government data portals within the EU. It has since grown to contain

[13] Open Definition 2.1 - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge
[14] Sunlight Foundation
[15] 5-star Open Data (5stardata.info)
[16] Find open data - data.gov.uk
[17] The ODI – Open Data Institute
[18] data.europa.eu

links to over one million datasets including country, EU and International data. Also, in 2013, the Group of Eight (G8) industrialized nations' leaders signed the Open Data Charter, following a global consultation of governments and civil societies (G8, 2013). The principles in the charter were further refined and published in 2015 (Open Data Charter, 2015).

In 2019, the European Commission issued a "Directive on open data and the re-use of public sector information", encouraging EU Members States to make as much information available for reuse as possible.

In June 2020, the Secretary General of the United Nations (UN) announced the "UN Data Strategy for Action by Everyone, Everywhere"[19], noting that *"data permeates all aspects of the work of the United Nations, and its power – harnessed responsibly – is critical to the global agendas that we serve"* (UN, 2020). The strategy addresses the broader context of general data management but does include a section on Open Data and sharing of data between organisations. It promotes leveraging data as a strategic asset, fostering UN data standards and interoperability. By 2021, over 33 UN organisations and development banks have already established Open Data portals

2.3 Open Data principles

In this section, we take a closer look at the various existing principles and guidelines that can be adopted by data providers and publishers in order to make data more available, accessible and reusable.

2.3.1 Open Government Data (OGD) Principles

Open government advocates declared that government data would be considered open if it is made public in a way that complies with eight principles (The Open Government Working Group, 2007), listed in *Table 1*. The principles cover a range of aspects from what is published, to how it is published.

In 2010, the original list of 8 principles was updated and expanded upon by the Sunlight Foundation to bring it to 10 principles in total (Sunlight foundation, 2010), as summarized in *Table 2.*

[19] UN Secretary-General's Data Strategy

The Original 8 Principles of OGD	Description
1. Complete	All public data is made available.
2. Primary	Data is as per the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. Timely	Data is made available as quickly as necessary to preserve the value of the data.
4. Accessible	Data is available online to the widest range of users for the widest range of purposes.
5. Machine Processable	Data is reasonably structured to allow automated processing.
6. Non-Discriminatory	Data is available to anyone, with no requirement of registration.
7. Non-proprietary	Data is available in a format over which no entity has exclusive control.
8. Licence free	Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Table 1: Original 8 Principles of OGD, extracted from (The Open Government Working Group, 2007)

Sunlight's 10 Principles of OGD	Summary
1.Completeness	Datasets should be complete and include metadata explaining the raw data and any data derivations.
2.Primacy	Datasets should be primary source data and include a description on how the data was collected.
3.Timeliness	Datasets should be available to the public in a timely fashion while they can still make use of it.
4.Ease of Physical and Electronic Access	Datasets should be as accessible as possible; obtained with ease, whether through physical or electronic means.
5.Machine readability	Data should be stored in widely used file formats that easily lend themselves to machine processing.
6.Non-discrimination	Any person can access the data at any time without having to identify him/herself or provide any justification for doing so.
7. Use of Commonly Owned Standards	Data can be accessed without the need for purchasing a software license.
8. Licencing	Datasets should be available, as much as possible, without restrictions or barriers on use.
9. Permanence	Data made available online should remain findable online over time, with appropriate version tracking and archiving.
10. Usage Costs	Fees for accessing the data should not be imposed.

Table 2: Ten Principles of Open Government Data adapted from (Sunlight foundation, 2010)

The first 8 principles are similar to, or expanded versions of, what was in the original 8 (*Table 1*). The 'Completeness' principle adds the requirement for metadata and information on how the data was derived. The 'Machine Processable' principle was renamed to 'Machine Readability', with emphasis on the need for widely used file formats for machine processing. The 'Non-proprietary' principle has become the 'Use of Commonly Owned Standards' but both have the same intention, that data is available in a format that does not impose any unnecessary restrictions, such as costs, over who can use the data. The 'Licensing' principle allows for more flexibility than the more specific 'License free' principle, although both encourage minimal barriers for access and re-use.

Adding 'Permanence' is to ensure links to online data are managed and maintained. The 'Usage Costs' principle reasons that a government who already has this data at their disposal should not charge for public access as it would act as a barrier for some and limit re-use.

The US-based Open Government Data organisation also annotated the original list (Tauberer, n.d) with 7 additional principles that were not considered originally but take into consideration among others, the Open Definition and Sunlight Foundation principles and guidelines.

7 Annotated Principles of OGD	Description
1. Online & free	Data should be findable on the Internet and available at no charge.
2. Permanent	Data should be made available at a stable Internet location indefinitely and in a stable data format for as long as possible.
3. Trusted	Data should be either digitally signed or have attestation of publication/creation date, authenticity, and integrity.
4. A presumption of openness	Public information should be available as open data, proactively and with no restrictions on its reuse and consumption.
5. Documented	Data should include a description of the data (metadata) to make it useful.
6. Safe to open	Data should be published in safe formats that do not include executable content.
7.Designed with public input	Seek input from the end users of the data to ensure value is delivered.

Table 3: The 7 Additional Principles of Open Government Data, extracted from (Tauberer, n.d)

These additional principles as listed in *Table 3* include variations of principles listed already, such as data should be online and free, permanent and documented. This list introduces new concepts applicable to the consumers of the data, such as, the data should be trusted, safe to open, and designed with public input.

2.3.2 The International Open Data Charter Principles

The Open Data Charter is a collaboration between governments and organisations working to make data open and freely available based on the shared set of 6 principles (Open Data Charter, 2015) summarized in *Table 4.*

Open Data Charter Principles	Summary
1. Open by Default	There should be the presumption that data will be published for all, with justification for when it is not, taking security, data protection and privacy into account.
2. Timely and Comprehensive	Data should be published quickly, while still relevant and comprehensively, as much as possible, in its original, unmodified form.
3. Accessible and usable	Data should be free of charge, under open licence, machine readable, easy to find and in a usable file format.
4. Comparable and Interoperable	Datasets that can "talk to each other" bring more potential value. Commonly agreed data standards are key to making this happen.
5. For Improved Governance and Citizen Engagement	Open data promotes transparency which can, in turn, improve public services and help hold governments to account.
6. For Inclusive Development and Innovation	Open data can spur economic development, and drive entrepreneurs to innovate and create new products.

Table 4: Open Data Charter 6 Principles summarized from (Open Data Charter, 2015)

'Open by Default' is encouraging a proactive approach to how data is managed; rather than waiting for requests to open datasets, the data should be available as open by default unless there are justifications for not doing so. The latter two principles in the charter are more about the purpose or promised benefits of providing Open Data, rather than specific to the characteristics of the data. As of May 2021, this charter has been adopted by 81 national and local governments from around the world and endorsed by 72 organisations[20].

2.3.3 The 5-Star Deployment Scheme

Tim Berners-Lee's 5-Star Open Data deployment scheme[21] is not so much a set of principles, but a plan that maps out progressive steps to follow to make data more open. It can also be used as a rating system for already published data, to determine where the data sits on the openness scale. *Table 5* enumerates the incremental levels of the scale, starting from one star

[20] The International Open Data Charter
[21] 5-star Open Data (5stardata.info)

indicating the most basic requirements for openness, with each additional star building on the prior step.

Star	Description
★	Make data available on the Web, in any format, but under an open license.
★★	Make data available as structured data (e.g., Excel instead of image scan of a table).
★★★	Make it available in a non-proprietary open format (e.g., CSV instead of Excel).
★★★★	Use URIs to denote data so that it can be referenced from other sources.
★★★★★	Link your data to other data to provide context.

*Table 5: The 5 * Open Data levels, adapted from (Berners-Lee, 2012)*

Berners-Lee's plan promotes linked data as the highest star rating. Linked data promises to make data more discoverable, increase the value of the data and provide benefits to both consumer and the publishing organisation but it can be more resource intensive to produce and broken links can arise if not frequently maintained (Berners-Lee, 2012).

2.4 FAIR Data Principles

So far, this chapter has focused heavily on principles related to or driven by open government data, but there has also been a need for specific principles and policies to ensure openness for scientific research data and materials, particularly if the research was publicly funded (Ramachandran, et al., 2021). The G8 Science ministers met on this topic in 2013 and agreed to support a set of principles for open scientific research data. They declared that *"Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards"* (G8 Science Ministers, 2013).

These initial principles evolved into the concept of Findable, Accessible, Interoperable and Reusable data (FAIR). The FAIR guiding principles were first formulated in 2014, by stakeholders sharing an interest in scientific data publication and reuse but were not formally published until 2016 in the Nature Publishing Group journal, Scientific Data (Wilkinson, et al., 2016). FAIR principles can be applied on a broader level than Open Data principles, with a view to making all aspects of the research process available, from data to analytical pipelines.

FAIR Principles	Details
To be Findable	F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To be Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
To be Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To be Reusable	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

Table 6: FAIR Guiding Principles, reproduced from (Wilkinson, et al., 2016)

The FAIR guiding principles (*Table 6*), place an emphasis on discovery and re-use, and can be applied to both metadata and data. They are minimally defined to encourage adoption by data publishers. The principles can be applied on a continuum for increasing degrees of FAIRness. FAIR principles can be applied on a broader level than Open Data principles, to both data and non-data assets, as making all aspects of the research process available assists discovery and reuse by third parties plus ensures transparency, reproducibility, and reusability (Mons, et al., 2017).

2.5 Comparison of Principles

Reviewing the principle and guideline groupings from the previous sub-chapters, we're now beginning to see some repetition emerge in the principles, especially on the topic of openness, timeliness and accessibility. Some are more domain specific (government) whereas others are more domain independent (FAIR). Some are very specific to Open Data (Open Data Charter) and others are more high-level so that they can be applied to both metadata and data, as well as non-data assets (FAIR). FAIR does not impose any constraint on the data being openly available whereas the others all expect open by default (Higman, et al., 2019).

Principle	Open Government Data	Open Data Charter	5 Star Open Data	FAIR
Accessible	Accessible	Accessible and Usable	Available online under open licence *	Accessible
Available online under open licence *	Accessible	Accessible and Usable	Available online under open licence *	Accessible
Comparable	Machine processable/readable	Comparable and Interoperable	Structured Data **	Accessible & Interoperable
Complete/Completeness	Complete/Completeness	Timely and Comprehensive	-	Findable
Comprehensive	Complete/Completeness	Timely and Comprehensive		
Designed with public input	Designed with public input	-	-	-
Documented	Documented	Timely and comprehensive	Structured Data **	Findable, Reusable
Findable	Documented, Permanence/permanent	Accessible and Usable	Structured Data ** and URIs ****	Findable
For Improved Governance and Citizen Engagement	-	For Improved Governance and Citizen Engagement	-	-
For Inclusive Development and Innovation	-	For Inclusive Development and Innovation	-	-
Interoperable	Accessible, Documented, Machine processable	Comparable and Interoperable	Linked Data *****	Interoperable
Licence free	Licence free	Accessible and Usable	Available online under open licence *	Reusable
Linked Data *****	-	Comparable and Interoperable	Linked Data *****	Interoperable
Machine processable/readable	Machine processable/readable	Comparable and Interoperable	Linked Data *****	Interoperable
Non-discriminatory	Non-discriminatory	Accessible and Usable	Non-proprietary open format ***	-
Non-proprietary	Non-proprietary	Accessible and Usable	Non-proprietary open format ***	-
Non-proprietary open format ***	Non-proprietary	Accessible and Usable	Non-proprietary open format ***	-
Online and free	Online and free	Accessible and Usable	Available online under open licence *	Accessible (but not necessarily free)
Open by Default	Presumption of openness	Open by Default	Available online under open licence *	-
Permanence/permanent	Permanence/Permanent	Accessible and Usable	URIs ****	Findable
Presumption of openness	Presumption of openness	Open by Default	Available online under open licence *	Accessible

Primary	Primary	Timely and Comprehensive	-	-
Reusable	-	Accessible and Usable	URIs ****	Reusable
Safe to open	Safe to open	-	-	Accessible
Structured Data **	-	-	Structured Data **	Interoperable, Reusable
Timely	Timely	Timely and Comprehensive	-	-
Trusted	Trusted	-	-	Reusable
URIs ****	Permanence/Permanent	-	URIs ****	Findable
Usable	Accessible	Accessible and Usable	Non-proprietary open format ***	Reusable
Usage Costs	Usage Costs, Online and free	Accessible and Usable (free of charge)	-	-

Table 7: Comparison of Principle Groups

As for the individual principles, some are specific to characteristics of the data (format, machine processable) and others target what data (complete, primacy) and how (non-discriminatory) that data should be published. Some are overlapping, some imply the others e.g. Linked Data implies Interoperable (Hasnain & Rebholz-Schuhmann, 2018). Some are a grouping of others; for example, 'Available online under an open license' already groups 'Accessible' and 'Licensing'.

We take each of the individual principles from the four tables presented in this chapter, separate them into individual items, and list them alphabetically. Then we cross check each item against the principle schemes and if that item is covered by a principle in the scheme, we list the applicable principles, and represent our findings in *Table 7.*

From the listing in *Table 7*, we see that some principles or data characteristics have emerged common to all principle schemes. Others are unique to only one grouping; namely 'Designed with public input' from the annotated list of Open Government Data Principles, and both 'For Improved Governance and Citizen Engagement' and 'For Inclusive Development and Innovation' from the Open Data Charter set of principles.

The open government data principles provide a comprehensive list to guide data publishers on how to open up, mostly pre-existing, government data, but do not explicitly mention linked data, structured data or reusability. The Open Data Charter principles are very concise and general and as a result, the broad principles could be interpreted to cover most of the

other principles in one way or another. The 5 Star Scheme for Linked Open Data provides a more technical guide towards publishing (Linked) Open Data and acts as a quality rating system as to how open the data is. FAIR principles focus on making data accessible under well-defined conditions, but not necessarily open or free. The data should be described enough to be discoverable, with clarity and transparency on how it can be (re-)used. All the principles and guidelines provide valuable input to an Open Data initiative. By combining them all we have a very comprehensive list of what should be included in managing our Open Data and portal.

2.6 Data Principle Aggregation

We now have a comprehensive list of principles from the various principle schemes. We are interested in how these principles can be applied in Open Data portal implementations and data management. But first we need to consolidate the list further by aggregating overlapping and similar items and forming higher level groups of related principles or categories (*Table 8*).

This results in six categories. The *Level of Openness* category is a grouping of the progressive steps of opening the data. The *data quality* category is a grouping of characteristics of the data or dataset that reflect some aspects of the quality. The *(meta)data management* category is the measure of how much the data is documented, findable, accessible, interoperable and (re)usable. The *legally open* category groups all the licence or terms of use related principles together. The *underlying purpose* category is more about the objectives of Open Data initiatives to begin with. The *technically open* category groups all the principles related to technical formats, standards or machine readability of the data.

In this chapter we have reviewed the history of Open Data and explored some of the more commonly known principles for managing and publishing (open) data. Comparing the principles across the different principle schemes, we see some overlaps occurring and some different emphasises. Combining all the principles together provides a comprehensive list of items to consider when embarking on an Open Data initiative. Consolidating further, we see some natural categories emerge and conclude with an actionable list of principle items grouped by category.

Category	Principle	Description
Level of openness	Available online under open licence *	The level 1 step to opening up data. * of 5 Star Open Data.
	Linked Data *****	The top level to opening up data. ***** of 5 Star Open Data.
	Non-discriminatory	No barriers to access; non-proprietary open formats. *** of 5 Star Open Data.
	Online and free	Making data available on the Web. * of 5 Star Open Data.
	Permanence/permanent	Achieved by using URIs - **** 5 Star Open Data.
	Structured Data **	Making data available as structured data. ** of 5 Star Open Data.
	URIs ****	Use (referenceable) unique IDs to refer to the data. **** of 5 Star
Data Quality	Complete/Completeness	Datasets should be as complete as possible, including metadata.
	Comprehensive	Publish as much as possible in its original form.
	Primary	Datasets should be primary source data, not aggregated.
	Safe to open	Quality measure.
	Timely	Data should be published quickly while it is still relevant.
	Trusted	A quality measure attesting to authenticity and integrity.
	Designed with public input	Engaging end user in the process to ensure value delivered.
(Meta)data Management	Accessible	Data and metadata retrievable by an identifier. The A in FAIR.
	Documented	Comprehensive metadata, guided by F, I and R of FAIR
	Findable	Discoverable data described by rich metadata. The F in FAIR.
	Interoperable	Promoted by use of standard (meta)data ontologies. The I from FAIR.
	Reusable	Rich accurate (meta)data with provenance, machine readable.
	Usable	Usable is closely tied to re-usable. The R in FAIR.
Legally Open	Licence free	Relates to how legally open the data is.
	Open by Default	Observable by portal or dataset terms of use. Legally open.
	Presumption of openness	Observable by portal or dataset terms of use. Legally open.
	Usage Costs	Observable by open access/licence
Underlying Purpose	For Improved Governance and Citizen Engagement	Promoting transparency, one of the purposes of Open Data
	For Inclusive Development and Innovation	One of the purposes for opening up government data
Technically Open	Comparable	To be comparable, data needs to be structured using commonly agreed standards.
	Machine processable/readable	Structure data in machine processable file formats.
	Non-proprietary format	Use of open formats. *** of 5 Star Open Data.

Table 8: Aggregation of principles into assessment categories

3 Open Data Portals

This chapter looks at how the Open Data principles are applied in practical implementations of Open Data portals. An Open Data portal is an online platform where Open Data is published, consumed and disseminated. A portal can host the actual data with associated metadata (a description of the data) or in some cases the portal acts as a data catalogue of metadata but the actual datasets are hosted elsewhere. It is interchangeably used with the terms Open Data hub or Open Data platform.

Taking the categories from the last chapter, we assess are they measurable and define metrics to measure them on the portals. Then we look at a sample of existing open data portals from across multilateral organisations and EU governments and assess them on the set of defined metrics. We also identify features from the assessment, that could add value to OrgX's portal implementation. We conclude with a derived list of key requirements and features to consider when setting up an Open Data portal to ensure openness and fairness.

3.1 Open Data Portal Metrics

From the principles categorised in *Table 8* we consider how their characteristics could be assessed on an Open Data portal, on a category basis. From a portal perspective, we consider openness as the degree to which data is *technically open* and *legally open*. In addition, we want to assess the *level of openness* of the data by measuring it against the 5 Star Linked Data scheme, on a continuum from one to five stars. We are also interested in determining how *(meta)data management* is supported on the portals; to what degree the data is documented, findable, accessible, interoperable and reusable on a scale of one to five.

Not all the principles relate directly to Open Data portal implementations or data characteristics. For example, some principles are more policy or objective oriented and therefore difficult to rate the implementation of these principles from a portal assessment alone. For example, 'For Improved Governance and Citizen Engagement' and 'For Inclusive Development and Innovation' are more about the *underlying purpose* for an Open Data initiative which is difficult to measure from a portal assessment. Both 'Open by Default' and 'Presumption of openness' are also policy related but the application of these principles could be measured by what access applies to the datasets, under the legally open category. The volume on these portals might also be a good indication.

Some of the aspects of the *data quality* dimension are difficult to measure from assessing a portal alone and could be subjective. For example, for the 'Timely' characteristic, we could check the dataset last updated date but without knowing the life cycle of the data concerned, it would be difficult to rate if the data is timely or not. It is also difficult to rate the implementation of the principles related to 'Complete', 'Completeness', 'Comprehensive' and 'Primary', without knowing the breath and source of data available within an organisation and reviewing data management plans. Therefore, we do not assign any metrics to the *underlying purpose* and *data quality* categories, but we take note of any indications of these in the assessment sheet (*Appendix A: Multilateral Portal Assessment*), in the 'Comments' field.

3.1.1 Technically Open Metrics

By technically open we mean that there are minimum to no technical barriers to accessing and reusing the data. Ideally the data is published and available using non-proprietary and machine-readable formats. Use of open source for development of the portal is also a plus as others can get access to and re-use the code, if desired. We propose a rating scale as per *Table 9* which awards more points, the more technically open the portal is.

Description	Rating
Provides support for linked data file formats (JSON-LD, N-Triples, RDF/XML, Turtle) and SPARQL search.	5
Provides APIs for machine processing	4
Provides easy access to raw data such as customizable bulk downloads, easy to use filtering	3
Use of non-proprietary file formats (CSV) and software (open source)	2
Provides access to the data online, but in proprietary formats (PDF, Excel)	1
Data not technically accessible at all	0

Table 9: Technically open rating scale

3.1.2 Legally Open Metrics

Legally open means data is published under terms of use with minimum restrictions, to encourage maximum access and reuse of the data. *Table 10* displays the common types of licences that information and data can be published under. Each type of licence is described, and some examples of each type provided. We suggest a rating system of 5 being awarded for the most open type of licence when data is available under public domain and therefore

all rights are waived, so the consumer of the data can use, modify and share as they wish. This encourages maximum re-use and innovation. At the opposite end of the rating spectrum, 0 is awarded for when no licence is specified as this stops any re-use of the data. In between the rating scale reflects more points for more openness; less points for the more conditions placed on a consumer of the data.

Common Licence Types	Description	Examples	Rating
Public Domain	Available in the public domain with all rights waived.	Public Domain mark, CC-0	5
Attribution	Anyone can use the data but must give credit, provide original licence and details if data modified.	CC-BY, ODC-BY	4
Share-alike	Anyone can use or build upon the data but must distribute using the same licence as the original.	CC-BY-SA, ODC-OdbL	3
Non-commercial	Data cannot be used for commercial purposes	CC-BY-NC,	2
No Derivatives	Data cannot be altered, transformed or built upon	CC-BY-NC-ND, CC-BY-ND	1
No licence specified	No one can use, share, distribute, re-post, add to, transform or change the data	Data published with no terms of use or Copyright	0

Table 10: Types of licence and legally open rating scale

3.1.3 Level of Openness Metrics

The level of openness of data, can already be measured by conveniently slightly adapting the 5 Star Open Data scheme (Berners-Lee, 2012), which already provides an incremental rating system from 1 to 5 for evaluating the level of openness of the data. We remove the open licence mention as that is covered under legally open. The rating scale is displayed in *Table 11*. At the lowest level of the scale, data is made available online but with accessibility challenges to get the data in a format that can be re-used. Moving up the scale, the barriers to access and re-use are reduced by structuring the data (**), using open formats (***) and using permanent links or URIs (****). At the top of the scale is linked data which facilitates maximum discoverability and reuse of data.

Description	Rating
Data is available online, open licence, as structured data, non-proprietary open format, using URIs to denote data and linked to other data to provide context.	5
Data is available online, open licence, as structured data, non-proprietary open format, using URIs to denote data e.g. RDF	4
Data is available online, open licence, as structured data, non-proprietary open format e.g. CSV	3
Data is available online, open licence, as structured data, e.g. Excel	2
Data is available for free, online, under any format, e.g. PDF	1

Table 11: Level of openness rating scale

3.1.4 (Meta)data Management Metrics

This category contains the FAIR elements plus 'Documented' and 'Usable', the latter being very similar to 'Reusable' of FAIR. It is difficult to rate data on portals against all FAIR aspects due to the multi-dimensionality of the FAIR principles. The fact that the data is published on an Open Data portal, already makes it findable and accessible to some extent. The use of URIs and usage licences are measured under our other metrics. What can be assessed is the presence and richness of metadata and catalogues and the use of standard ontologies. This also helps with measuring how well documented the data is, or how comprehensive the metadata is for that data. Measuring how well the data and metadata are managed on the portal indicates how findable, accessible, interoperable, reusable and documented the data is. We suggest a simplified point system to rate the portals for (meta)data management, with portals gaining a point for each characteristic observed as presented in *Table 12*.

(Meta)data Requirement	**Points**
Presence of data catalogue which is indexed and searchable.	1
Presence of rich metadata describing data.	1
Accessible using unique identifier over the internet, including the unique identifier and licence type in the metadata.	1
Includes reference to other (meta)data.	1
Metadata uses community standards, formats and/or ontologies.	1

Table 12: (Meta)data management points rating

Now we apply these metrics to portal implementations from both international organisations and governments and assess how much the portals have implemented the Open Data principles, using our defined categories and metrics.

3.2 Multilateral Organisation Data Portals

As our use case is based on a fictitious multilateral type of organisation, we start our assessment looking at existing portals in this category. In March 2016, the UN launched the United Nations System Data Catalogue[22], a single interface for finding UN data. Its purpose is to promote sharing of Open Data from across the organisations that make up the UN system.

[22] Home | UN System Data Catalog (undatacatalog.org)

No	Abbrev.	Organisation Name	Data Topic	Portal Name	Portal URL
1	FAO	Food and Agriculture Organisation	Food & agriculture statistics	FAOSTAT	http://www.fao.org/faostat/en/#home
2	ILO	International Labour Organization	Labour statistics	ILOSTAT	https://www.ilo.org/global/statistics-and-databases/lang--en/index.htm
3	IMF	International Monetary Fund	International finance data	IMF DATA	https://www.imf.org/en/Data
4	IOM	International Organization for Migration	Migration data	Missing Migrants	https://missingmigrants.iom.int/
5	OCHA	Office for the Coordination of Humanitarian Affairs	Humanitarian data	Humanitarian Data Exchange	https://data.humdata.org/
6	OECD	Organisation for Economic Co-operation and Development	Economic data	OECD Data	https://data.oecd.org
7	UN	United Nations	SDG data	UNdata	http://data.un.org/
8	UN Habitat	UN Habitat	Urban indicators	UN-HABITAT OPEN DATA	https://data.unhabitat.org/
9	UNCTAD	United Nations Conference on Trade and Development	Trade investment and development data	UNCTAD Statistics	https://unctad.org/statistics
10	UNDP	United Nations Development Fund	IATI data	d-Portal	http://d-portal.org/ctrack.html#view=search
11	UNDP	United Nations Development Fund	Project and donor data	UNDP Transparency Portal	https://open.undp.org/
12	UNDP	United Nations Development Fund	Homan developments	Human Development Data Center	http://hdr.undp.org/en/data
13	UNECE	United Nations Economic Commission for Europe	SDGs, MDGs, economy, population, transport	UNECE Statistical DB	https://w3.unece.org/PXWeb/en
14	UNECLAC	United Nations Economic Commission for Latin America and the Caribbean	Regional overview	Databases and Statistical Publications	https://estadisticas.cepal.org/cepalstat/Portada.html
15	UNEP	United Nations Environment Programme	Environment	Environmental Data Explorer	http://geodata.grid.unep.ch/
16	UNESCAP	United Nations Economic and Social Commission for Asia and the Pacific	Asia & Pacific development statistics	SDG Gateway	https://dataexplorer.unescap.org/

17	UNESCO	United Nations Educational, Scientific and Cultural Organization	Education, science, culture and communication	UNESCO Institute for Statistics (UIS)	http://data.uis.unesco.org/Index.aspx
18	UNFPA	United Nations Population Fund	Programme expenses	Transparency Portal	https://www.unfpa.org/data/transparency-portal
19	UNHCR	United Nations High Commissioner for Refugees	Refugee situations	Operational Data Portal	https://data.unhcr.org/en/situations
20	UNICEF	United Nations International Children's Emergency Fund	Children	UNICEF Data	https://data.unicef.org/
21	UNIDO	United Nations Industrial Development Organization	Industry	UNIDO SATISTICS	https://stat.unido.org/
22	UNODC	United Nations Office on Drugs and Crime	Crime, corruption	Data	https://www.unodc.org/unodc/en/data-and-analysis/statistics/data.html
23	UNOPS	United Nations Operations	Programmes and projects data	UNOPS Open Data	https://data.unops.org/
24	UNSD	United Nations Statistics Division	MDG Indicators	Millennium Development Goals Indicators	http://mdgs.un.org/unsd/mdg/default.aspx
25	UPU	Universal Postal Union	Postal	Postal Statistics	https://www.upu.int/en/Universal-Postal-Union/Activities/Research-Publications/Postal-Statistics
26	WB	World Bank	Development	World Bank Open Data	https://data.worldbank.org/
27	WFP	World Food Programme	Food security	VAM	https://dataviz.vam.wfp.org/
28	WHO	World Health Organization	Health	Global Health Observatory	https://www.who.int/data/gho/
29	WIPO	World Intellectual Property Organization	Intellectual property	Intellectual Property statistics	https://www.wipo.int/ipstats/en/

Table 13: List of portals from the UN Data Catalogue

The catalogue does not contain the actual data but rather publishes the metadata about datasets from across the different portals. It provides a listing of 33 Open Data portals operated by various agencies and organisations within the UN system[23].

Immediately we discover that 3 of the links in the catalogue do not resolve or do not take us to where we expect. From Google, we eventually find the correct URL for two of the portals, the other one we remove from our list. This demonstrates how quickly catalogues can age if not actively maintained. We discover that two of the portals require registration and authentication to access the data and have a costing model in place to charge for the data. These cannot be assessed as they do not embrace the 'Open by default' principle so they are removed from the list. Finally, one portal does not present any data but is more a website with web content. It cannot be rated so it is also removed from the list. That leaves 29 portals remaining which can be seen in *Table 13*, with the organisation name, abbreviation, topics covered, portal name and URL.

We assess the portals in this list against the rating system described in the previous section. In some cases, the data are published in an inconsistent standard across the portal; for example, some datasets have metadata or are accessible in non-propriety format, whilst others on the same portal are not. In these cases, our assessment takes the best effort datasets into account and rates the portal's support for these features. The results are summarized and visualized in *Figure 4* and the complete scoresheet can be found in *Appendix A: Multilateral Portal Assessment.*

From the summary, we see immediately very little activity at the higher range of the *level of openness* and *(meta)data management* scales. The portals score better on the *technically open* and *legally open* scales. We drill into the details for each scale in the next sections.

[23] Open Data Portals | UN System Data Catalog (undatacatalog.org)

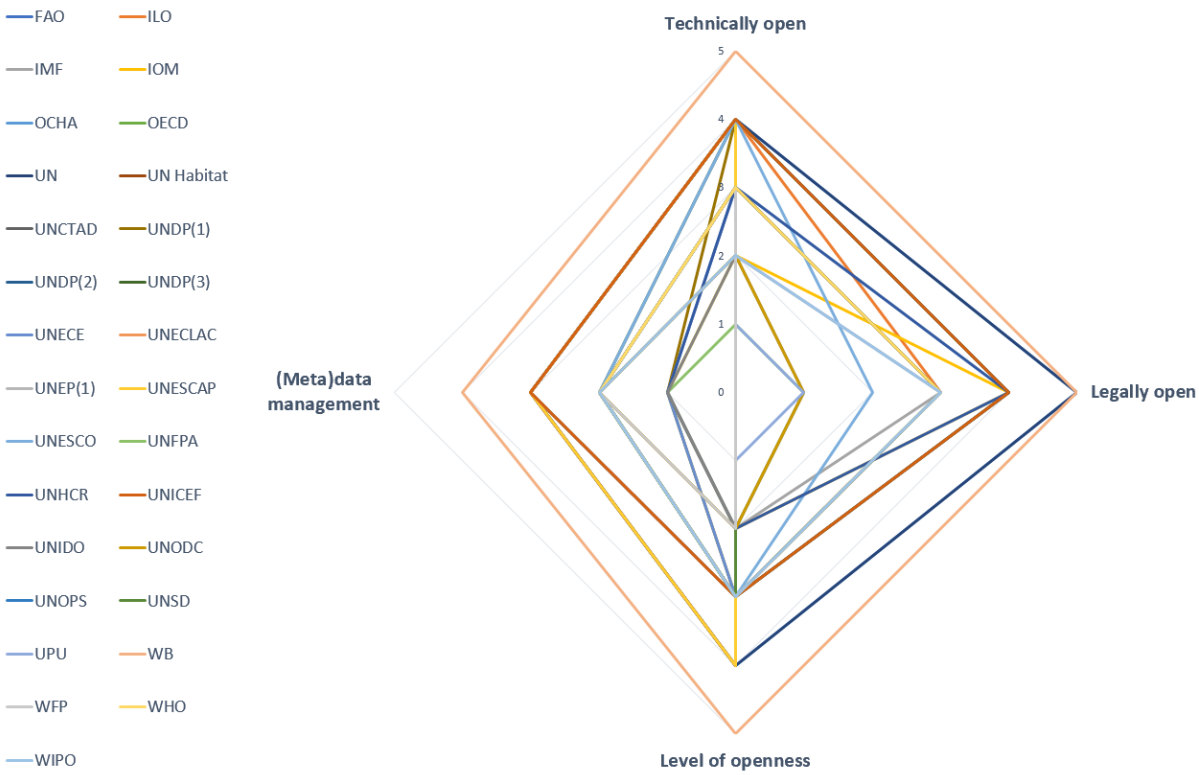

Figure 4: Comparison of multilateral Open Data portals

3.2.1 Technically Open Ratings

Looking at the technology behind the portals (*Figure 5)*, twenty two of the portals are custom built using either open source technologies or the Microsoft stack and ASPX pages. Four portal implementations took a product-based approach and It was not possible to identity which technology was being used for the remaining three of the portals. From those that took a product-based approach, as in leveraged existing software rather than custom building a portal from ground up, only one was identified as using specific open data portal software, CKAN. CKAN is an open source data management system[24] developed by OKF. Another portal used the ArcGIS platform[25], a product that works well with location-based data and displaying data on maps. The other 2 portals use a .Stat Suite[26] framework which is commonly used for working with statistical data.

[24] CKAN - The open source data management system
[25] About ArcGIS | Mapping & Analytics Software and Services (esri.com)
[26] .Stat Suite Open Source Framework - .Stat Suite documentation (sis-cc.gitlab.io)

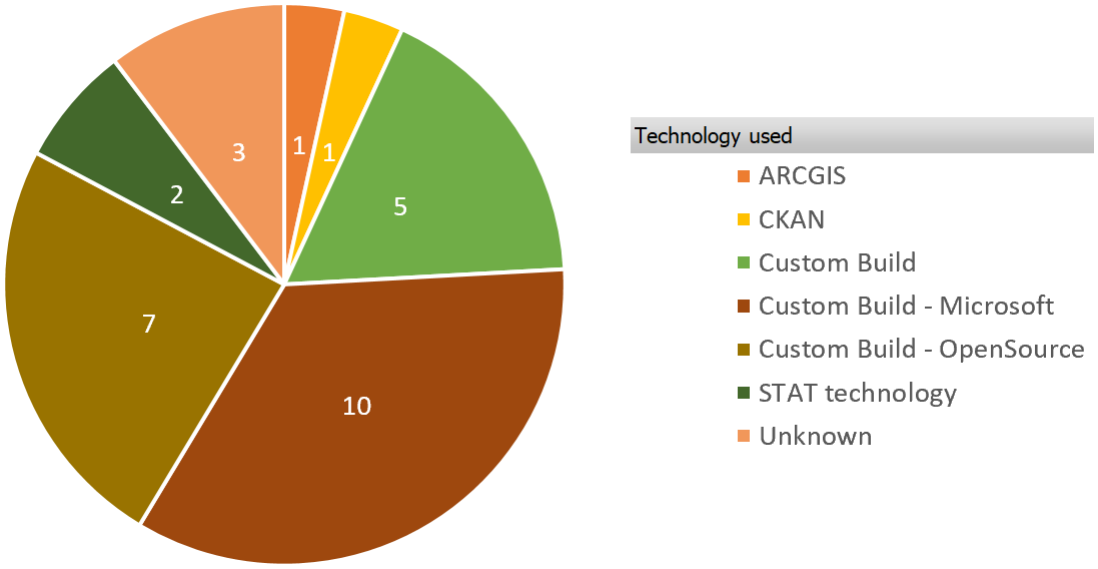

Figure 5: Technology used to implement multilateral portals

The first column in *Figure 6* displays the technically open rating for the twenty-nine portals. All the portals assessed meet the first criteria on the *technically open* rating scale, by providing access to the data online, in at least proprietary formats. Two portals do not score more than this base score of 1, as they do not offer the data in non-proprietary format, using either PDF or Excel instead. On the opposite end of the scale, only one portal clearly supports and provides linked data, scoring maximum 5. In between we find five portals offering just the basic technical functionality, providing data as CSV, scoring 2. Thirteen portals offer an application programming interface (API) for programmatic access to the (meta)data, but two of those require registration for use of the API, so 11 portals score 4 points and the other two are marked down by 1. Those two portals along with 8 other portals score a mark of 3, for enabling users to download the data, either in bulk or allowing some filtering before download.

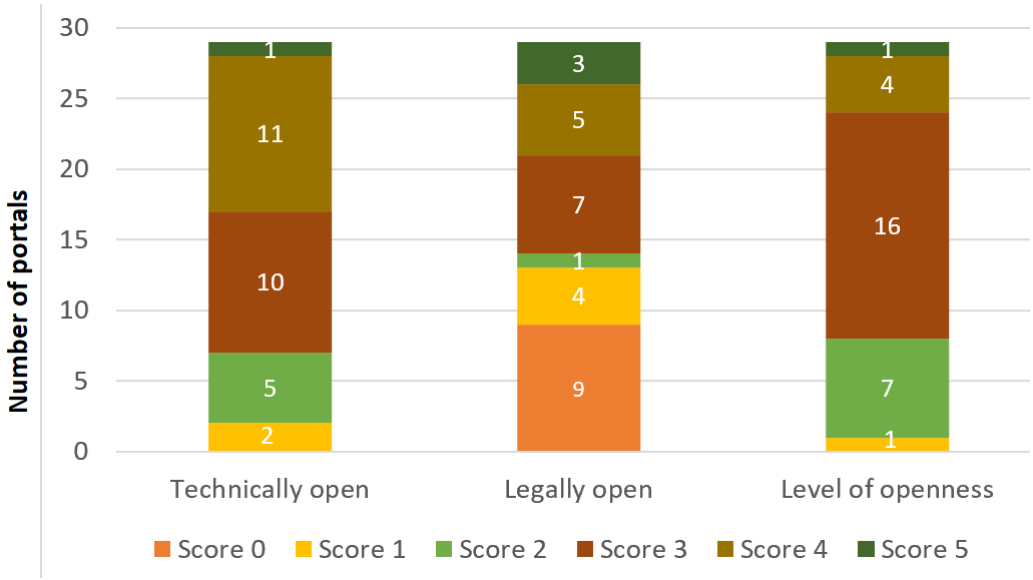

Figure 6: Aggregated assessment ratings for multilateral portals

3.2.2 Legally Open Ratings

The results from the legally open part of the assessment are depicted in the central column of *Figure 6.* Nine portals did not have any terms of use or licencing specified on the portal or the link to the terms of use gave an error (in one case). The remaining twenty portals had either specified terms of use or specific data licences to varying degrees although at times it was difficult to determine what applied. For example, in most of the cases, the use of the data was not mentioned at the dataset level, but was specified in a Terms of Use page, which made reference to how the overall portal content, including data, may be used. These terms are generally worded as text in web content format, and thus not readily machine processable. In the case of four of the portals, there was more than one set of terms published specifying use for the different types of content, some contradicting the others. This is very confusing to know which one applies. Only one portal had licencing labels at the dataset level making it easy for both humans and machines to understand how the data can be re-used.

Three portals make some of their data available in the public domain, waiving all rights, scoring maximum points. Five portals share the data using an attribution licence including the latest version of Creative Commons (CC BY-4.0). The Share-alike licence type used by seven portals is more common than the Attribution type. One portal did not make it clear how the data could be used for commercial purposes, requesting that the consumer get in touch with

the organisation to discuss further. Four portals stated that their data could not be altered, transformed or built upon.

3.2.3 The Level of Openness Ratings

As seen in *Figure 6*, in the *Level of openness* column, over half of the portals (16) manage to make the data available online, in a structured, non-proprietary, open format (CSV), which warrants a score of 3. Eight of the portals do not reach this mark usually due to using proprietary formats. An additional four make use of URI or offer the data in RDF format. Only one portal showed evidence of using linked data, scoring top marks.

3.2.4 (Meta)data Management Ratings

The (Meta)data Management rating scale differs from the other incremental-level scales in that in this case individual points are assigned per criteria, and the combined score is assigned. This is the rating category with the lowest score, with none of the portals scoring a 5 and one portal scoring 0, due to not appearing to have any metadata management features, as seen in *Figure 7.*

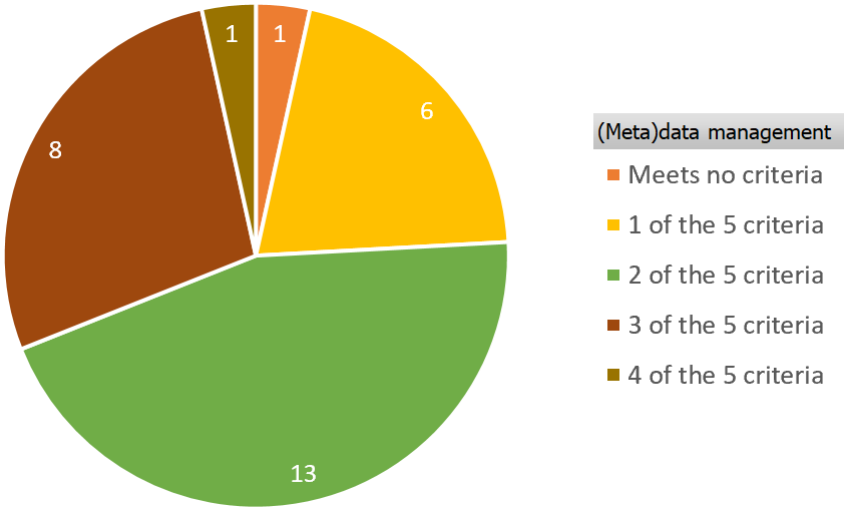

Figure 7: (Meta)data management ratings for multilateral portals

The remainder of the portals score a point for having some sort of catalogue for browsing the data, with metadata on the data. The richness of the meta data varies across the datasets and portals. In most cases the metadata does not include any time stamps to indicate the freshness of the data. There did not appear to be any standard ontologies used for describing the data.

3.2.5 Additional Comments

Browsing thru the portals, we note that exploring the data is easier when the catalogue can be filtered by many dimensions, such as topics and keywords and that the datasets are presented in a consistent fashion or template. In some cases, it was difficult to identify where the raw data could be accessed, especially in portals that presented the data in aggregated form in PDFs or data visualizations.

Nineteen of the portals are multilingual with six of those offering the six UN official languages of Arabic, Chinese, English, French, Russian and Spanish. The other ten portals are English only. Changing the language selection only affects the web content but the metadata and data remain in the original source language.

It was difficult to grasp how much data is available, as the number of datasets or providers was not always apparent. Less than half of the portals indicated either on the home page or the catalogue page how many datasets, data providers and topics. For the others we marked them as unknown.

Just three of the portals mention Open Data; one displays the Open Definition's Open Data label on the site, indicating that the data satisfies the Open Definition, the other two go into more detail to describe what Open Data is and how their data complies.

3.3 Government Data Portals

The USA and UK government data portals were among the first government data portals to be launched back in 2009[27] and 2010[28] respectively. Since then many other governments world-wide, both at the national and local level, have launched portals (Parycek, et al., 2014) and signed up to the Open Government Partnership[29], an organisation committed to transforming how governments serve citizens. Data.europa.eu is the official portal for European data combining the former EU Open Data Portal and the European Data Portal[30], and aggregates data from the national and regional portals across the EU. It has built up a body of knowledge, impact and case studies, best practice guides, reports and assessments

[27] About Data.gov - Data.gov
[28] About - data.gov.uk
[29] About (opengovpartnership.org)
[30] About data.europa.eu | data.europa.eu

on Open Data. It publishes an annual Open Data Maturity report[31] that assesses the level of Open Data maturity in the EU Member states (Knippenberg, 2020).

No.	Country	Ab.	URL	Technically open	Legally open	Level of openness	(Meta)data Mgmt.
1	Austria	AT	https://www.data.gv.at/	4	4	4	4
2	Belgium	BE	https://data.gov.be/en	4	1	3	2
3	Bulgaria	BG	https://data.egov.bg/	3	5	4	3
4	Croatia	HR	https://data.gov.hr/	4	0	3	2
5	Cyprus	CY	https://www.data.gov.cy/	5	4	5	3
6	Czechia	CZ	https://data.gov.cz/english/	5	4	4	2
7	Denmark	DK	https://www.opendata.dk/	3	3	3	2
8	Estonia	EE	https://avaandmed.eesti.ee/	3	4	3	2
9	Finland	FI	https://www.avoindata.fi/	4	5	4	3
10	France	FR	https://www.data.gouv.fr/en/	4	5	4	5
11	Germany	DE	https://www.govdata.de/	5	5	5	4
12	Greece	GR	https://data.gov.gr/	2	0	3	1
13	Hungary	HU	http://www.opendata.hu/	2	1	4	2
14	Ireland	IE	https://data.gov.ie/	5	4	5	5
15	Italy	IT	https://dati.gov.it/	3	4	3	4
16	Latvia	LV	https://data.gov.lv/	3	5	3	2
17	Lithuania	LT	https://data.gov.lt/	4	4	3	3
18	Luxembourg	LU	https://data.public.lu/	3	5	3	3
19	Malta	MT	https://data.gov.mt/	0	0	0	2
20	Netherlands	NL	https://data.overheid.nl/	4	4	4	3
21	Poland	PL	https://dane.gov.pl/pl	4	4	4	3
22	Portugal	PT	http://www.dados.gov.pt/	3	4	3	2
23	Romania	RO	https://data.gov.ro/	4	4	3	3
24	Slovakia	SK	http://data.gov.sk/	4	5	3	4
25	Slovenia	SI	https://podatki.gov.si/	4	4	5	3
26	Spain	SP	https://datos.gob.es/en	5	3	5	5
27	Sweden	SE	https://www.dataportal.se/	4	4	4	4

Table 14: Ratings for the EU27 portals

Table 14 displays the snapshot of the scores for each of the EU 27 Open Government portals. There is a lot of uniformity across these different portals in terms of structure, navigation items, user experience and features offered, which makes them a lot easier to assess, despite a language barrier. Whereas the UN family of portals were all presented in English, each of the government portals are presented in the country's national language. All but four of the portals (AT, DE, DK and IT) offered additional languages including English, but the translation was applied only to central elements and content, not to the metadata or data. Thanks to

[31] Open Data Maturity | data.europa.eu

google translate we can extract the information we need to assess, but the metadata language could cause a barrier to interoperability across regions.

Figure 8 displays the ratings plotted in a grid. From first glance, we see a lot more activity on the outer rims of the grid in comparison to the multilateral portals (*Figure 4)*. These portals have been in place for a longer period and the maturity shows in the more advanced features as discussed in more detail in the next section.

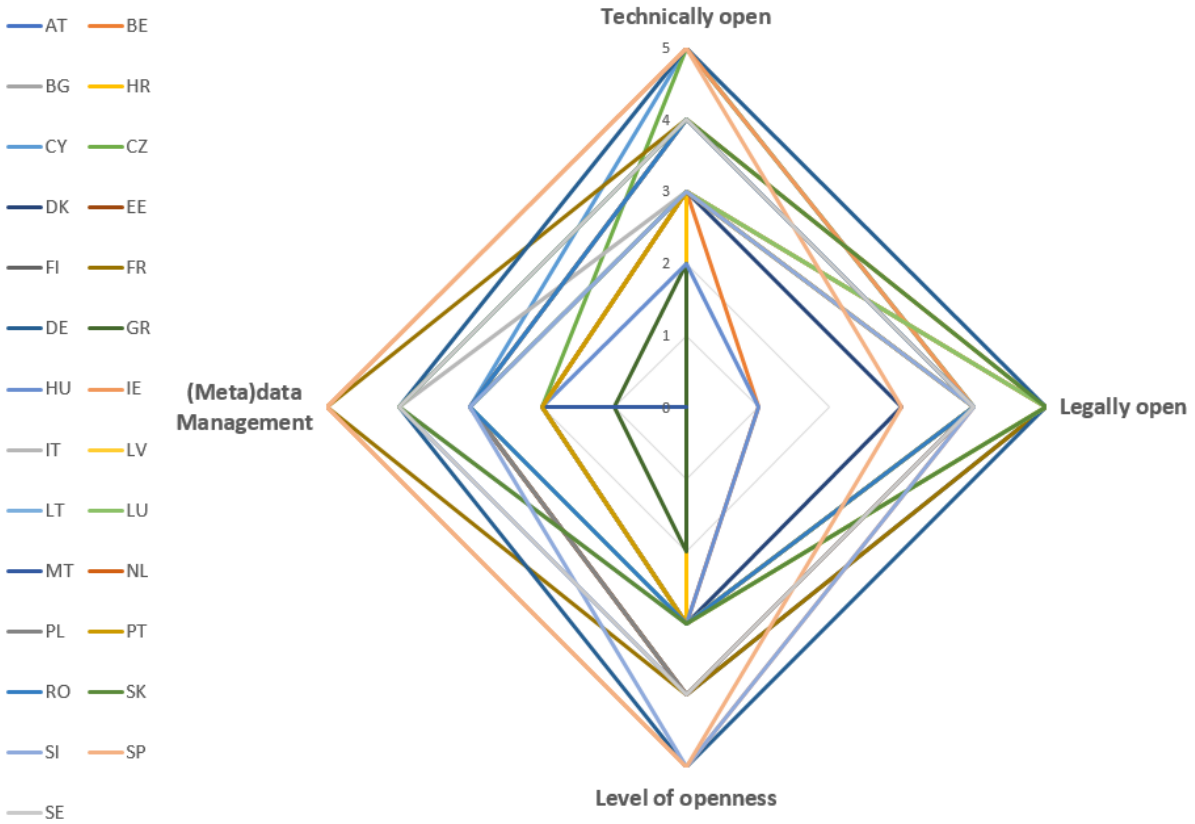

Figure 8: Comparison of EU27 Portals

3.3.1 Technically Open Ratings

As mentioned, there is a lot more uniformity across these portals. On closer inspection we see that nineteen of the twenty-seven portals are built on the CKAN platform *(Figure 9)*, which explains the similarity in the experience across those sites. One portal is built on DKAN[32] which is an open source Open Data platform, similar to CKAN, but written in PHP, designed to

[32] DKAN Open Data Platform | DKAN Open Data Platform (getdkan.org)

integrate into the Drupal content management platform[33]. Another portal is custom built directly on Drupal, also open source. One portal is also custom built but on proprietary Microsoft software. Two portals are custom built using a combination of open source tools and the code base is shared on GitHub.

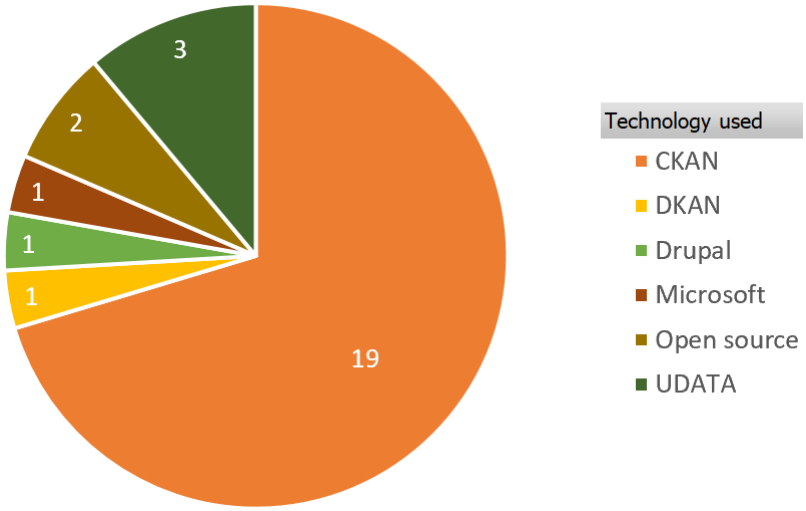

Figure 9: Technical platforms used in EU27 portals

UDATA is a customizable Open Data platform, developed by the French public agency in charge of Open Data, Etalab. It is built as an extension of CKAN and in addition to the FR portal, it is used by two others, LX and PT.

The number of portals per each score on the technically open scale is displayed in the first column in *Figure 10.* One portal scores zero as we are unable to access data, neither from browsing the catalogue nor downloading, but it does appear to be a beta version.

The remainder of the portals present the datasets in a consistent manner, with a link to either preview or bulk download the dataset. These portals support a broad range of formats, over 20 in most cases, including the main non-proprietary format CSV. Two of the portals technically should score 3 points but features were not working which marked them down to 2. Sixteen portals provide API access to the data.

[33] Drupal - Open Source CMS | Drupal.org

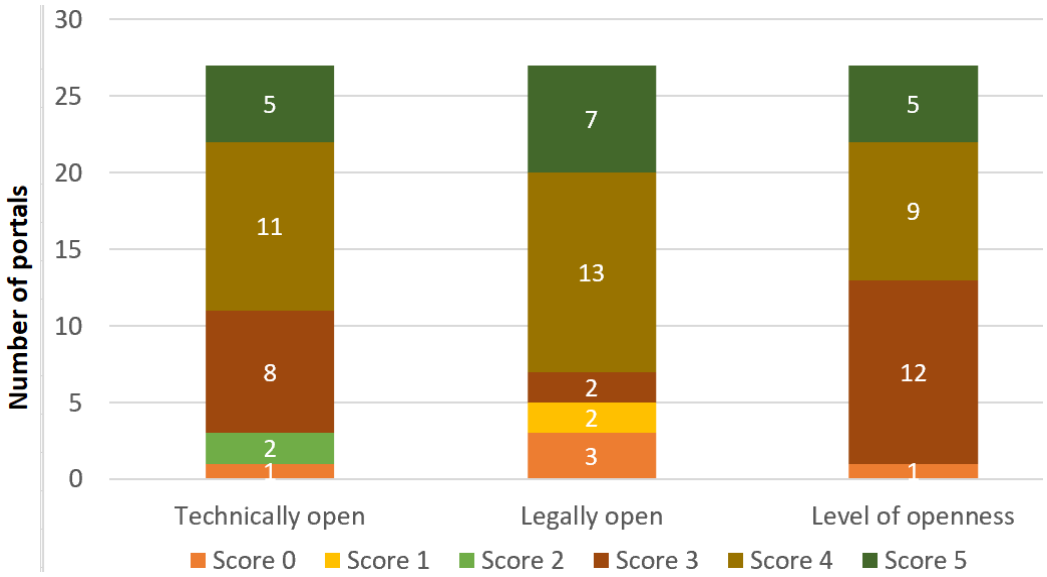

Figure 10: Aggregated assessment ratings for EU27 portals

The others either showed no evidence of API access or required registration or the API responded with an error message. Five of the portals provided some of their datasets in linked data file formats, either RDF or Turtle. Five portals provide a SPARQL search feature, although in some it appears to be in testing mode.

3.3.2 Legally Open Ratings

The majority of the portals consistently label the datasets with a licence type and allow filtering of the catalogue by licence type. Some portals define their own licence types, such as Belgium, Denmark and Romania using Open Data Flanders, opendata.dk and OGL-ROU-V1.0 respectively. It is difficult to interpret these licences as they do not map directly to our predefined metrics and appear to be a combination.

The aggregated scores from the legally open scale are displayed in the middle column of *Figure 10.* Seven portals score top marks in this category for labelling the majority of their datasets as public domain/free to use. Three portals score zero for not mentioning any licence nor terms of use. CC-BY is the most commonly used license type with the majority of portals publishing some of their datasets under this standard licence type. A couple used the CC-Share-alike. The portals at score 1 were the ones with ambiguities over the licence or the licence not being as easy to machine process.

3.3.3 The Level of Openness Ratings

Five of the portals (HR, IE, LT, PL, RO) display a visual star rating of the datasets in the catalogue making it very easy to see immediately how open the data is. However, in the case of Lithuania, it causes confusion as there is also a 5 Star rating for the overall quality of the dataset as perceived by the end user. The other sites do not mention the 5 Star model, but the datasets are mostly 3 and 4-star with unique identifiers and data available as RDF for some of the datasets. Five portals have a very small percentage of 5-star datasets. One portal has datasets with "Linked" in the title, but they are rated as 4-star on the portal. One portal scored zero as the data was not available in any format. The aggregated results are displayed in the right column of *Figure 10.*

3.3.4 (Meta)data Management Ratings

Thanks to how the CKAN product operates, all the CKAN based portals have a catalogue listing the datasets including metadata to various degrees of richness. The catalogue is searchable, and results can be filtered using metadata fields such as topics, organisations, format, licence and keywords. The functionality is there to document rich metadata although the quality varies by dataset. Some portals use the out-of-the-box CKAN metadata offering, others extend it by using DCAT and OWL ontologies. The metadata contains a space for a unique identifier, the last updated date and frequency of updates of the datasets. The datasets are labelled with the associated licence. Even the non-CKAN based portals offer similar features. Only three portals appear to cross reference the datasets. The combined scores are aggregated and displayed in *Figure 11.*

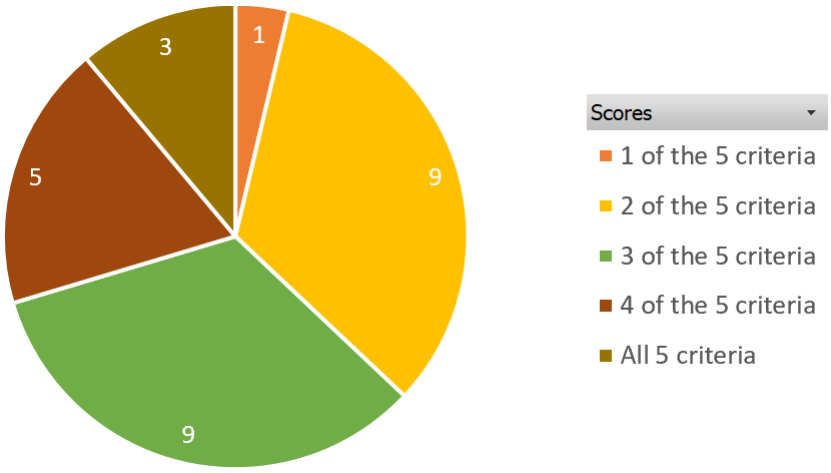

Figure 11: (Meta)data management ratings for EU27 portals

3.3.5 Additional Comments

All the portals display the number of topics, providers and datasets making it easy to get a sense of what's available to consumers. The volume is on a much larger scale than for the multilaterals, but many of these government portals are aggregating datasets from a large number of local and regional sources, contributing to the volume. It is clear from *Figure 12* that there is a broad range in volume from 47 datasets to 135,855.

Three of the portals provided a SPARQL[34] search for querying the RDF data but there are no instructions on how to utilize it.

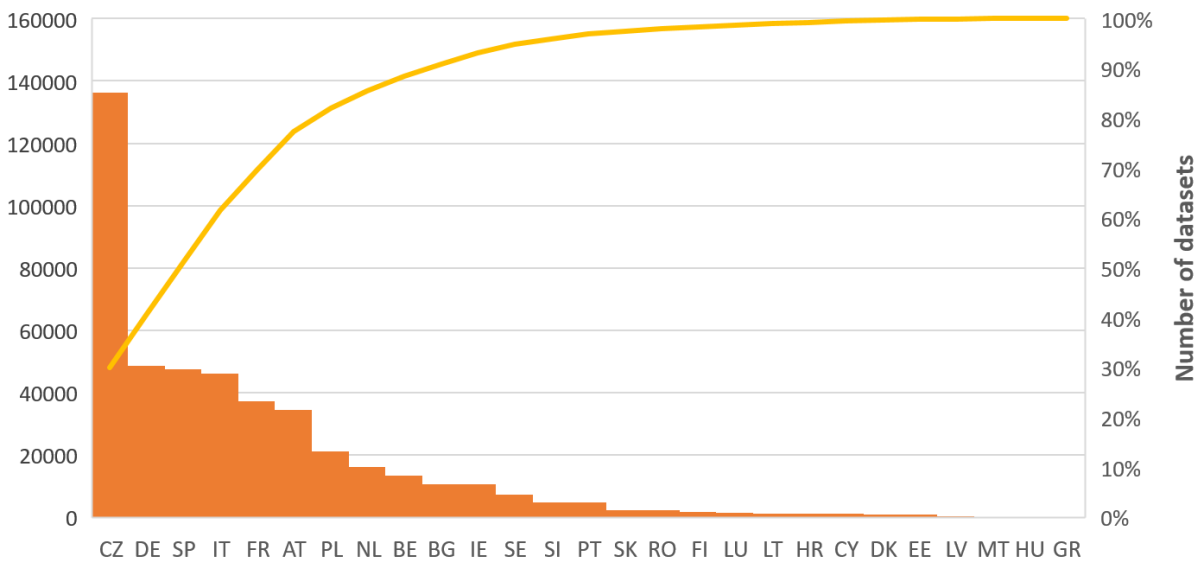

Figure 12: Volume of datasets per EU 27 country

3.4 Desired Portal Features

From the assessment of the portals, we see some key requirements emerging for a portal to support both providers of data and consumers. Firstly, the portal needs to support the aggregated principles and categories identified from chapter 2 (*Table 8)* and metrics from this chapter.

[34] SPARQL Query Language for RDF (w3.org)

Category	Requirements and features of an Open Data portal
Level of openness	Ability to make data available online, with no barriers to access.
	Ability to publish data as structured data.
	Ability to share data in non-proprietary formats; CSV, XML, JSON.
	Ability to publish (meta)data with a unique identifier.
	Ability to link data with other datasets.
(Meta)data management	Support for the findability of (meta)data; including use of rich metadata, keywords, URIs.
	Ability to register or index the (meta)data in a searchable resource.
	Ability to make (meta)data accessible by both humans and machines.
	Support for standards and ontologies to describe the data.
	Ability to reference other datasets.
	Ability to indicate the usage licence and provenance for the data.
	Ability to document the data.
Legally open	Ability to indicate the licence applied to the data.
	Support for machine readability of the licence type.
Technically open	Support for machine readable file formats (CSV, XML, JSON).
	Ability to filter and (bulk) download the data.
	Support for APIs.
	Support for linked data (file formats JSON-LD, N-Triples, RDF/XML, Turtle and SPARQL search).
Data Quality	Ability to indicate the quality of the data.
Underlying purpose	Ability to engage end users, such as support for discussion boards or feedback channels.
User experience	Display of site metrics; how many datasets, topics, publishers.
	Ability to preview the data.
	Ability to apply filters by topic, formats, licences and other high-level tags.
	Ability to indicate the frequency and last modified dates for the datasets.
	Ability to indicate the level of openness from the 5 Star Open Data scheme.
	Display or suggest related datasets.
	Quick access to copy the permanent link to the data.
User engagement	Ability to request data, leave comments on datasets or discussions.
	Ability to rate the dataset or "like" the dataset.
	Ability to follow the dataset and get notified of updates.
	Contact data publisher feature, suggest correction feature.
	Ability to easily share content from the portal to social media platforms.
	Ability to create/view visualizations to gain insights from data.
	Access to impact stories, showcases of the data and reuse count.
	Publishing content that fosters the community, news, hackathons.
User assistance	Documentation for data providers, how to publish data.
	Documentation for consumers, how to use the data.
	Description of tags, field names.
	Access to a Knowledge base, code base.
	Data tools available on the portal.

Table 15: Requirements for an Open Data portal

In addition, we noticed some usability features of the assessed portals that made them easier to use and easier to make sense of the data. These additional user-oriented requirements are added to produce *Table 15.*

These requirements and desired features represent the breath of what could be implemented in an Open Data portal to ensure openness, FAIRness and a good user experience for publishers and consumers of the data.

4 Open Data Set-up and Sustainability Planning

Managing an Open Data portal, to meet both the needs of data publishers and consumers, whilst ensuring openness and FAIRness, requires planning to establish the portal and ongoing operational plans to ensure sustainability of the initiative. In this chapter, we begin by researching the types of plans that are already in use to manage data, and the elements that make up these plans. We compare templates from (Open) Science, Horizon 2020 and Open Government initiatives. We explore what elements should be included in a data management plan (DMP) to ensure Openness, FAIRness and sustainability. By reviewing DMP templates, H2020 template and Open Data publication plans from research, OGD and public sectors and combining with recommendations for sustainability of Open Data portals, we identify key elements that should be included in an Open Data management plan.

4.1 Data Management Plans

The Data Management Association (DAMA) defines data management as, *"the business function that develops and executes plans, policies, practices and projects that acquire, control, protect, deliver and enhance the value of data and information"*[35].

The importance of data management for Open Data initiatives is recognised in many sources. Among them, the European Data Portal's "Open Data Goldbook for Data Managers and Data Holders" recommends that an organisation should be clear on its data management practices before attempting to publish data as Open Data (Carrara, et al., 2018). Before publishing data as Open Data, the organisation should have an overview of the current data management structure and see how the data management is organised. Luna-Reyes identifies improved data management practices along the data lifecycle as a top requirement for realizing the benefits of Open Government Data (2019).

A key component of good data management practices is having a data management plan. Data management plans or DMPs are defined as a written document that describes the data you expect to acquire or generate during the course of a project, how you will manage, describe, analyse, and store data, and what mechanisms you will use at the end of your project to share and preserve your data. They have traditionally been associated with the

[35] DMBoK - Data Management Body of Knowledge (dama.org)

scientific and research community to ensure traceability and reproducibility of research results. They are required by funding bodies and institutions world-wide. The United States National Science Foundation requires that all proposals include a data management plan to facilitate future dissemination and sharing of research results[36]. The National Institute of Health (NIH) has issued a data management and sharing policy establishing the requirement of the submission of data management and sharing plans for research funded or conducted by NIH[37]. The European Commission expects project proposal submissions to address good research data management and include a Research DMP, updated at different stages in the project.

DMPs are also being introduced to ensure research data is findable, accessible and re-usable after the project ends. In other words, DMPs are an essential mechanism for research groups to ensure their outputs are FAIR (European Commission Expert Group on FAIR Data, 2018), but not necessarily open. For Open Data initiatives, Open Data publication plans and policies govern the data processes required for preparing and publishing the data. In the next sections, we will take a closer look at these different types of plans for managing data and what components do they include.

4.1.1 The DMP Review

In their 2017 study on DMPs, Williams et al. (2017) reviewed sixty-six data management/data sharing requirement documents from research funders across different types of organisations. The review found a low percentage of funders require a DMP with more emphasis on data sharing plans indicating a greater focus on data sharing and re-use than on upstream processes such as data collection and processing. The review also uncovered a large variation in the required or suggested DMP topics among funder requirements. In total, forty-three DMP topics were identified and categorized into seven related groupings (see *Table 26* in *Appendix: DMP topics*). Based on this list, the authors offer a list of topics for a DMP adding the requirement for traceability of data in research contexts, which they felt was lacking from the DMPs that were reviewed. Their suggested hybrid DMP structure (*Table 16*) is designed

[36] Dissemination and Sharing of Research Results | NSF - National Science Foundation
[37] NOT-OD-21-013: Final NIH Policy for Data Management and Sharing

to aid data managers in comprehensively addressing the management of research data (Williams, et al., 2017).

Component	Summary
Project Personnel	Personnel, their qualifications, roles and responsibilities, data access.
Description of all data sources	Describing the origin of the data.
Data and workflow diagram	Depicting the flow of data from source, through information systems and operations.
Definition of data elements	Data models, procedures and algorithms performed on data to ensure traceability.
Procedures for all operations performed on data	Details of any operations performed on the data for traceability.
Description of software and devices used for data	Configuration specifications, testing plans, security plans, backup plans for data.
Privacy and confidentiality plan	Steps to take to comply with privacy and information security policies.
Project management plan	Planned deliverables, schedule and milestones, reporting and controls, and resource estimates.
Data retention, archival and disposal plan	The planned final stages of the project; what to do with the data.
Data sharing plan	How to share the data.

Table 16: Data Management Plan Framework Components, adapted from (Williams, et al., 2017)

4.1.2 The Horizon 2020 DMP Template

Horizon 2020 is an EU Research and Innovation programme that funds research projects with the goal to ensure Europe produces world-class science, removes barriers to innovation and makes it easier for the public and private sectors to work together in delivering innovation[38]. Project proposal submissions to Horizon 2020 are expected to address good research data management and submit and update a living Research DMP during the life of the project. They suggest a template, summarized in *Table 17*, with a strong emphasis on making data FAIR (European Commission, n.d.).

4.1.3 Open Data Publication Plans

Open Data publication plans can be thought of as a close equivalent to DMPs but are for managing the opening up of data, in particular government data. While we did not discover one specific format or template in our research, we found that many sources specify how to

[38] What is Horizon 2020? | Horizon 2020 (europa.eu)

establish an Open Data publication plan (Kučera, et al., 2015), (Carrara, et al., 2018) and (Share-PSI 2.0, 2016).

Component	Summary
1. Data Summary	Describe the purpose of data collection/generation, data types and formats, data origin, data volume, and potential users.
2. FAIR Data	
2.1 Making data findable, including provisions for metadata	Describe data discoverability, what metadata will be provided or created and how; any metadata standards, use of naming conventions, URIs or DOIs.
2.2 Making data openly accessible	Describe what data will be made openly available, what software is required to access data, where the data, metadata, documentation, code will be stored and how access is provided if there are restrictions.
2.3 Making data interoperable	Specify (meta)data vocabularies, standards and methodologies to facilitate interoperability, use of standards and ontologies.
2.4 Increase data re-use (through clarifying licences)	Specify licences, terms of use of the data by third parties, describe data quality assurance processes.
3. Allocation of Resources	Identify roles and responsibilities for data management, estimate costs, one off and long-term maintenance, Identify source of funding.
4. Data Security	Address data recovery as well as secure storage and transfer of sensitive data.
5. Ethical aspects	Any ethics review of data needed or ethics deliverables.
6. Other Issues	Refer to any procedures in use for data management.

Table 17: Horizon 2020 DMP Template Fields, adapted from (European Commission, n.d.)

As presented in *Table 18*, these plans mainly describe what datasets will be published as open data, how open or what the target level of openness is, what roles are involved in the publication of Open Data and what responsibilities they have.

Component	Summary
Decide what data(sets) to publish	Planning what to publish, is there demand for the dataset, justify if not publishing.
Determine level of openness	From one to five star, for example.
Define roles and responsibilities	Identify roles in the process, in particular data ownership.
Define terms and conditions for use of data	Legal openness.
Outline an action plan	To include, the level of: Technical openness - what formats to take, and standards to adopt to publish data, e.g. in machine-readable formats FAIRness - Metadata standards to use.

Table 18: Open Data Publication Plan Components

The plan also describes what are the terms and conditions under which the open datasets should be published. Finally, the plan sets a roadmap for publication of the selected datasets. The roadmap typically includes the management of data quality, technical Openness, legal

Openness and metadata, specifying recommended formats, metadata schemas and standards for the data to be published.

4.2 Sustainability Planning

The definition of sustainable, according to the Merriam Webster dictionary, is *"capable of being maintained at length without interruption or weakening"*[39]. Susha et al. (2015) identified the sustainability of the Open Data initiative as a critical success factor for Open Data publication. Sustainability of Open Data portals can be summarized as the extent to which a portal can respond and adapt to challenges (Walker, et al., 2020). It is important, to consider the sustainability of an Open Data initiative, from the initial set-up and design, to ensure that the portal and the data it contains will be actively maintained and remain relevant over time and to avoid creating another data dump (Publications Office of the European Union, 2020).

"The Future of Open Data Portals" report enumerates ten ways in which Open Data portals must evolve for sustainability and added value (Simperl, 2020). *Figure 13*, extracted from the report, displays the ten approaches. We propose to include these approaches in our plan to ensure sustainability of an Open Data portal.

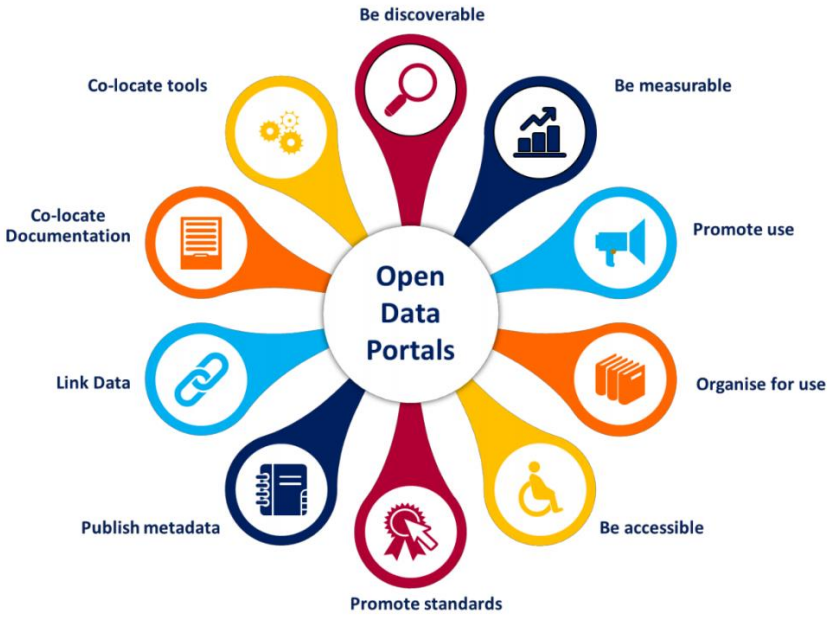

Figure 13: Ten ways to make a portal more sustainable: extracted from (Simperl, 2020)

[39] Sustainable | Definition of Sustainable by Merriam-Webster

Table 19 summarizes what is represented in *Figure 13* and describes the key approaches and recommended actions to follow to achieve Open Data portal sustainability (Simperl, 2020).

Sustainability Guidelines	Summary	Recommended Actions
1. Organise for Use	Design Open Data portal with the user experience in mind, understand the users' journeys and enrich the experience using online retail techniques.	Analyse user behaviour Consider focus groups, usability tests, A/B testing, eye-tracking and participatory design workshops to drive the portal direction.
2. Promote Use	Make it easier for consumers to use the data, create and share data stories.	Link related datasets from within and across other portals; Create a community of practice for feedback and sharing of experiments.
3. Be Discoverable	Ensure data is discoverable from the organisation's different channels including the public web site.	Enable data portals for search engines; Consider Schema.org for microdata mark-up; Be transparent on what datasets could be made available if demand exists.
4. Publish Metadata	Metadata should help users make sense of data and include relevance, usability and quality descriptions.	Ensure metadata is machine readable Use standard metadata schemes such as DCAT-AP Use metadata to make relationships between datasets.
5. Promote Standards	Use well defined common standards to enable parties to have a shared understanding of the data, attributes. licences and connections.	Use technical key standards for data and metadata, Use standard domain vocabularies.
6. Co-locate Documentation	Make supporting documentation immediately accessible from within the dataset to facilitate making sense of the data.	Allow comments and contributions from broader community.
7. Link Data	Link datasets to core reference data to allow cross referencing and analysis of multiple datasets.	Implement and maintain Linked Data via use of URIs and domain vocabularies or use algorithms to calculate similarities.
8. Be measurable	Portals should measure usage for publishers and quality for users.	Define metrics for both categories and how they will be measured.
9. Co-locate Tools	Make tools available with the data to enable users to explore data sets and determine relevance for them.	Provide basic mapping and visualization tools with the dataset.
10. Be Accessible	Ensure data can be easily accessed by both humans and machines.	Avoid publishing data in proprietary formats, e.g. PDFs.

Table 19: Approaches to achieve sustainability of Open Data portals, adapted from (Simperl, 2020)

Whilst these approaches are targeting portals, they have an impact on the upstream activity of data management and preparing the data for publishing, especially those related to the FAIR principles. Some emphasize the importance of managing the data post-publication (Organise for use, Promote use, Be measurable) to ensure the portal continues to be fit for purpose for both producers and consumers of the data.

4.3 Comparison of Plans and Approaches

Taking the components from the four plans presented, we list them all alphabetically and cross-check them against each plan (*Table 20*). Comparing the different types of plans and approaches outlined above, we see that each type of plan has a different focus. The hybrid DMP structure (*Table 16*) with such a focus on managing data from research projects, places heavy emphasis on managing the provenance, creation or collection aspects of data during a project life cycle. The publication or sharing of the data is a very small part of the overall plan. This template does consider the final stages of the data lifecycle and data disposition, something the other templates do not explicitly mention.

On the contrary, the Horizon 2020 DMP template *(Table 17*) has one component, the data summary, to cover the purpose, objectives, the data origin, types and formats and thus encompasses four of the components from the hybrid DMP. The bulk of the remainder of Horizon 2020 plan concentrates on ensuring the data generated from the research is designed to be FAIR. It alone mentions ethical aspects which might in practice be covered by the 'Privacy and confidentiality plan' component from the DMP template, although we consider ethical aspects to be broader. The component, 'Other issues', is a catch-all to refer to any other procedures for data management that are used.

The Open Publication plan (*Table 18)* is more a plan for organisations with existing data rather than focussing on data that is an outcome of a project, which explains why it alone has the decision of what data to open-up, and not so much emphasis on the data creation. It has the broad category of creating an action plan which in theory could subsume most of the data sharing/publishing components from the other plans. This approach of a plan within a plan is also present in the hybrid DMP, hinting at the cyclical nature of publishing and maintaining Open Data.

The sustainability plan *(Table 19*), written as directives, is specific to how to manage a portal to ensure sustainability. It overlaps most with the Horizon 2020 plan, with both having FAIR related components, thus highlighting the importance of FAIR data for sustainability.

Components	Hybrid DMP	Horizon2020 DMP	Open Data Publication	Sustainability Plan
Allocation of Resources	Project Personnel and project mgmt. plan	Allocation of Resources	Define roles and responsibilities	-
Be accessible	-	Making data openly accessible	-	Be accessible
Be discoverable	-	Making data findable	-	Be discoverable
Be measurable	-	-	-	Be measurable
Co-locate documentation	-	-	-	Co-locate documentation
Co-locate tools	-	-	-	Co-locate tools
Data and workflow diagram	Data and workflow diagram	Data Summary	-	-
Data retention, archival and disposal plan	Data retention, archival & disposal	-	-	-
Data Security	Description of s/w and devices used for data	Data Security	Outline an action plan	-
Data sharing plan	Data sharing plan	Making data openly accessible	Outline an action plan	Be accessible, Promote use
Data Summary	Description of all data sources and Definition of data elements	Data Summary	Decide what data(sets) to publish	Publish metadata, Co-locate documentation
Decide what data(sets) to publish	-	Making data openly accessible	Decide what data(sets) to publish	-
Define roles and responsibilities	Project Personnel	Allocation of Resources	Define roles and responsibilities	-
Define terms and conditions for use of data	Data sharing plan	Increase data re-use	Define terms and conditions for use of data	Promote standards
Definition of data elements	Definition of data elements	Data Summary	Outline an action plan	-
Description of all data sources	Description of all data sources	Data Summary	Decide what data(sets) to publish	Publish metadata
Description of Software and devices used for data	Description of Software and devices used for data	Data Summary	Outline an action plan	Co-locate tools
Determine level of Openness	Data sharing plan	Increase data re-use	Determine level of Openness	Be discoverable
Ethical aspects	-	Ethical aspects	-	-
Increase data re-use (through clarifying licences)	-	Increase data re-use	Outline an action plan	Promote use, Promote standards
Link data	-	Making data interoperable	-	Link data

Making data findable, including provisions for metadata	-	Making data findable	Outline an action plan	Be discoverable
Making data interoperable	-	Making data interoperable	Outline an action plan	Link data
Making data openly accessible	-	Making data openly accessible	Define terms and conditions for use of data	Be accessible
Organise for use	-	-	-	Organise for use
Other Issues	-	Other Issues	-	-
Outline an action plan	Data sharing plan	All the FAIR steps	Outline an action plan	-
Privacy and confidentiality plan	Privacy and confidentiality plan	Making data openly accessible	Decide what data(sets) to publish	-
Project management plan	Project management plan	Allocation of Resources	Outline an action plan	-
Project Personnel	Project Personnel	Allocation of Resources	Define roles and responsibilities	-
Promote standards	-	Making data findable, Making data interoperable	-	Promote standards
Promote use	-	Increase data re-use	-	Promote use
Publish metadata	-	Making data findable	-	Publish metadata

Table 20: Comparison of DMP, Horizon 2020, Open Data publication and Sustainability plans

It alone considers establishing metrics for measuring usage and quality (Be measurable), the former to inform data providers and the latter to indicate to the consumers what to expect from the data. It also is alone in encouraging a user-centric approach (Organise for use) and in co-locating documentation and tools, to assist consumers in exploring and making sense of the data.

4.4 Aggregated Components to form a Set-up and Sustainability Plan

We combine the related components from *Table 20* and list them in the order that we anticipate they should happen. Analysing the flow, we see four groupings emerge of components that are more plan-oriented, those that are more action oriented, those that are about review and those about taking follow-on actions. This is a natural fit with the Deming cycle, an iterative design and management methodology for continuous improvement of a process or product[40].

[40] PDSA Cycle - The W. Edwards Deming Institute

Stage	Component	Summary	Context
PLAN	Decide what data(sets) to publish	Planning what to publish, is there demand for the dataset, justify if not publishing	Dataset
	Determine level of Openness/Define terms and conditions for use of data	What level of openness from one to five-stars, what type of licence to be used to ensure it is legally open.	Dataset
	Description of all data sources/data summary	Describing the purpose, objectives, data types and formats, data origin, data volume, and potential users.	Dataset
	Project management plan/ data sharing plan/ action plan	Planned deliverables, schedule and milestones, reporting and controls, and resource estimates for the broader project and subsequent cycles.	Both
	Define roles and responsibilities/allocation of resources/Project personnel	Personnel, their qualifications, roles and responsibilities, data access for both the broader context and identifying personnel for each dataset.	Both
	Privacy and confidentiality plan	Steps to take to comply with privacy and information security policies. Establish a standard, but plan to check each dataset against the standard.	Both
	Ethical aspects	Steps to take to ensure ethical aspects are considered. Establish a standard, but plan to check each dataset against the standard.	Both
	Data and workflow diagram	Depict the flow of data from source, thru information systems and operations.	Dataset
	Definition of data elements	Data models, procedures and algorithms performed on data to ensure traceability.	Dataset
	Description of software and devices used for data	Infrastructure and configuration specifications, testing plans, security plans, backup plans for data.	Portal
	Data security	Address data recovery as well as secure storage and transfer of sensitive data	Portal
DO	Organise for use	Design Open Data portal with the user experience in mind, understand the users' journeys and enrich the experience using online retail techniques.	Both
	Making data findable/Be discoverable	Describe data discoverability, what metadata will be provided or created and how, any metadata standards, use of naming conventions, URIs or DOIs	Dataset
	Making data openly accessible/Be accessible	Ensure data can be easily accessed by both humans and machines	Dataset
	Making data interoperable/Link data	Link datasets to core reference data to allow cross referencing and analysis of multiple datasets,	Dataset
	Increase data re-use	Clarify data licences to make it easier to reuse	Dataset

	Promote standards	Use well defined common standards to enable parties to have a shared understanding of the data, attributes. licences and connections.	Both
	Publish metadata	Metadata should help users make sense of data and include relevance, usability and quality descriptions.	Both
	Co-locate documentation	Make supporting documentation immediately accessible from within the dataset to facilitate making sense of the data	Dataset
	Co-locate tools	Make tools available with the data to enable providers to easily publish and consumers to explore data sets and determine relevance for them	Dataset
STUDY	Be measurable	Collect and review statistics on usage for publishers and quality for consumers.	Both
ACT	Promote use	Make it easier for consumers to use the data, create and share data stories.	Dataset
	Data retention, archival and disposal	As part of data life cycle planning, manage the data post-publication	Dataset

Table 21: Aggregated components for a set-up and sustainability plan

Also known as the Plan-Do-Study-Act cycle, it fits the cyclical nature of data management, data publishing, assessing the impact and iterating. We group the components by a stage in the Deming cycle, resulting in *Table 21*. To organise our plan, we identify that some components or activities belong to the broader initiative of setting up the portal and establishing a data publication roadmap, whilst others are very specific to performing activities on individual datasets. Then there are others that can be applied in both contexts. We label the components if they apply to the 'Portal', 'Dataset' or 'Both' contexts.

We now apply this approach in our use case for first planning the open data portal and then managing the publishing of datasets to the portal on a cycle.

5 Open Data from Set-up to Sustainability Use Case

This chapter describes a concrete deployment of the data principles aggregated in chapter 2 *(Table 8*), the data portal feature list from chapter 3 (*Table 15)* and the data set-up and sustainability components from chapter 4 (*Table 22*) in a practical use case of establishing and sustaining an Open Data portal. OrgX are exploring how to create a single, unified Open Data portal that enhances discoverability of datasets, promotes use through ease of storytelling, empowers business owners to respond rapidly and take ownership of their data sources and engages and responds to user needs. A data portal that includes metadata, clear licencing agreements, encourages standards and accessibility while enhancing interoperability.

Figure 14 is a high-level representation of how the project was designed to run iteratively to continuously build on what the previous cycle had delivered. The project started with planning the portal and managing identified data (PLAN). The portal was then implemented, and a subset of the data was prepared and published (DO). Then, metrics were put in place to measure usage and feedback channels are established. The data from these sources was analysed (STUDY). Follow-on actions were determined, including responding to the feedback and usage statistics, and promotion of the portal or published datasets (ACT). That brings us back to the start, the next planning stage, incorporating what surfaced from the previous stage and planning the next batch of data for publication and tasks for sustaining the portal. This cycle ensures continuous reflection on what has gone before and planning for improvements over time, guaranteeing an evolving product that stays relevant.

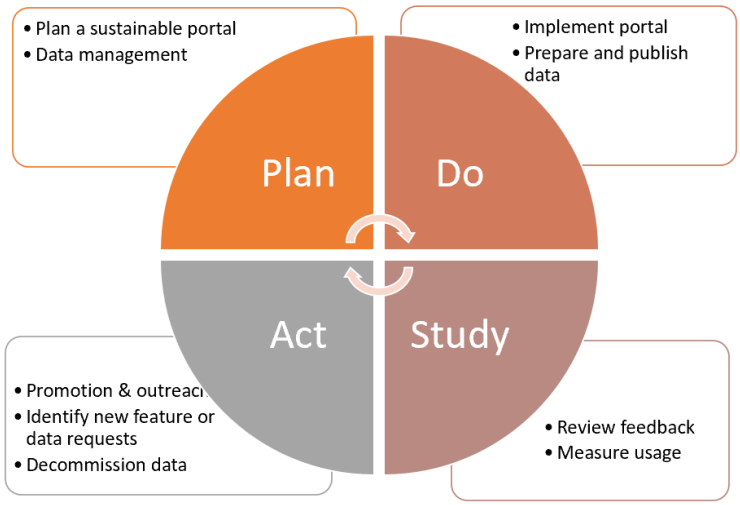

Figure 14: Approach to set-up and sustain an Open Data portal

At the time of writing, we have completed one full cycle of the framework by actioning the components, portal requirements and principles.

5.1 PLAN STAGE

Taking the components from the PLAN stage, guided us in the initial planning process and determined a list of actions to take (*Table 22*). This planning happened in parallel with planning the data processes. The portal planning focussed on defining and setting standards to follow and data planning tested and informed what the standards should be to ensure support for the different types of datasets.

To manage the data, we first created a data audit of OrgX's existing datasets. We analysed which of them could potentially be published as Open Data and what level of openness would be applicable. Discussions were held with the legal department to determine what terms of use could be applied to the portal and the datasets contained within. The ideal was to publish all datasets at the highest level of openness (5 Star), where possible, and to apply the CC-By 4.0 licence per dataset.

As each existing dataset was evaluated, a decision was taken on the appropriate openness level based on business need and the licence based on the source of the data. We gathered a standard set of details on each dataset and set priorities for moving them to the portal. For the first cycle, a selection of five datasets were chosen to publish based on their properties, complexities and ability to test specific features of the portal platform. For example, the datasets chosen had either specific security, visualization, sharing or format requirements and were of high priority. For the five chosen, a data workflow was designed, and the data model planned.

Once the initial plans were approved, the project team were then ready to implement the portal and selected datasets as part of the first pass through our process cycle.

Stage	Context	Component	Action Taken
PLAN	Dataset	Decide what data(sets) to publish	A Data audit was created as a comprehensive inventory of existing datasets. Applying set criteria, recommendations were made on how to manage and publish those datasets with assigned priorities.
	Dataset	Determine level of Openness/Define terms and conditions for use of data	Planned to support maximum level of openness per dataset; therefore 5 stars. Set the standard for publishing datasets as CC-By 4.0.
	Dataset	Description of all data sources/data summary	This was captured as part of data audit
	Both	Project management plan/ data sharing plan/ action plan	A high-level project plan was created as part of the initiation stage of the project, documenting high-level stages for establishing a portal and a process for data publication. This was heavily informed by a comprehensive data audit.
	Both	Define roles and responsibilities/allocation of resources/Project personnel	Roles and responsibilities were defined for the life cycle of the project as part of the project plan. At the individual dataset level, a data owner, maintainer and contact details were collected as part of the data audit.
	Both	Privacy and confidentiality plan	Part of the project initiation, setting a standard approach based heavily on the organisation wide policy on privacy and confidentiality
	Both	Ethical aspects	Part of the project initiation, applied the Data Ethics Canvas by ODI
	Dataset	Data and workflow diagram	Designed a data pipeline to specify how the data would be extracted from the source. Identified any transformations or operations required per dataset.
	Dataset	Definition of data elements	Planned a data model per dataset.
	Portal	Description of software and devices used for data	Described existing s/w as part of the as-is analysis and identified key features/requirements for the to-be project vision.
	Portal	Data security	Documented security requirements for selection of an Open Data portal. Planned risk assessment and security reviews of implementation.

Table 22: PLAN Stage - Components and actions list

5.2 DO Stage

This stage had two goals, to implement the base portal and migrate 5 datasets following the DO components from the set-up and sustainability plan (*Table 23)*. However, we found our DO components lacking in that they did not specify how to establish the portal initially and the components were more focussed on configuring an existing portal and datasets. So we took the requirements and desired features identified from chapter 3, namely *Table 15* and

used these to go about identifying, evaluating and selecting the base portal software to run the Open Data portal on.

5.2.1 Implementing the portal

To establish the portal, we first needed to determine what software to use as the base, by creating evaluation criteria to evaluate a number of options. The evaluation criteria were a mixture of functional and non-functional requirements and cost considerations. We expanded *Table 15* by adding the non-functional requirements that are standard for IT projects at OrgX, such as requirements specifying performance, availability, backup and recovery, security and infrastructure constraints.

We evaluated the options of taking a product-based approach, exploring CKAN, DKAN and ArcGIS, or custom building the solution. We had seen the popularity of CKAN with Open Government Data platforms in chapter 3. DKAN, also open source, offers similar functionality to CKAN. ArcGIS[41] is a geographic information system which offers solutions connecting data to maps. The team who built the original web applications proposed developing their code base into a platform, offering a custom build approach. We applied the evaluation criteria to all four options and the CKAN product scored the highest and was selected and procured. The team then set about configuring the portal and implementing the datasets.

5.2.2 Implementing datasets

The implementation of the portal features to support the datasets was guided by the requirements list and the components from the DO stage of our plan and lead to the listed actions taken (*Table 23)*. Using a portal product like CKAN enabled the implementation of many of the features and requirements as simple configurations of the underlying platform.

A challenge arose in determining how to make the data interoperable and whether there was a requirement for linked data. For one dataset, the data owner was very interested in publishing the dataset as linked Open Data, but more time is required to establish what standards to use and what to link to. Technology is not the constraint but governance and policies. Making APIs available and publishing the datasets with unique identifiers added to the metadata sufficed for the first cycle of our plan. In future iterations, this will be revisited.

[41] ArcGIS Online | Web GIS Mapping Software for Everyone (esri.com)

Stage	Context	Component	Action taken
DO	Both	Organise for use	Completed a user journey mapping exercise for the standard user profiles. Designed and configured the portal to fulfil the journeys. Gathered usage statistics on existing web applications.
	Data	Making data findable/Be discoverable	Determined the metadata fields required per dataset; any deviations from the standard were justified and added if necessary. Explored use of DOIs for datasets linked to publications.
	Dataset	Making data openly accessible/Be accessible	The CKAN portal features ensure data can be easily accessed by both humans and machines.
	Dataset	Making data interoperable/Link data	No use case was identified in the first cycle.
	Dataset	Increase data re-use	Datasets published with clear (usage) licence per dataset. We ensured the data download and API functionality worked per dataset. Materials created for outreach and promotion on how to use the data.
	Both	Promote standards	Defined a standard set of metadata attributes per dataset, defined the licence type (CC-BY 4.0), and defined technical formats (CSV) and API standards (endpoints for querying data).
	Both	Publish metadata	Ensured the portal product selected could publish a flexible, extendable set of metadata with support for standards such as DCAT. At the dataset level, leveraged the CKAN data catalogue for publishing Metadata, extending it where necessary, included provenance, relevance, usability and quality descriptions.
	Dataset	Co-locate documentation & tools	Where available, provided links to supplementary documentation from the dataset. For consumers, made visualizations tools available for gaining insights into the data. Plans are in place for future addition of tools to assist data owners

Table 23: DO Stage - Implementing the portal and datasets

Screen shots of the implemented portal and datasets can be found in *Appendix C: OrgX Open Data Portal.*

5.3 STUDY Stage

There was one item in the set-up and sustainability plan for this stage in the cycle: 'Be measurable' (*Table 24)*.

Stage	Context	Component	Action taken
STUDY	Both	Be measurable	Employed Google Analytics on the portal pages and utilized the inbuilt statistics in the CKAN platform. Ran focus groups to gather initial feedback from both providers and consumers of the data.

Table 24: STUDY Stage – measuring and reviewing results

This component prescribes setting metrics for usage of the portal and for the quality of the data. For gathering data on usage, we employed Google Analytics on the portal pages and utilized the inbuilt statistics in the CKAN platform. However, the pilot portal has not yet been in production for long enough nor promoted and therefore has not generated enough traffic to generate insights into the usage.

For the data quality metrics, work is under way in defining how best to measure. The level of openness based on the 5 Star Open Data scheme is widely recognised as a quality metric on how open data is. We ensure that data owners can indicate on the portal how many stars their data deserves. We saw from our assessment of portals in chapter 3, that other data quality characteristics are difficult to determine from the portal view alone. We ensured that the metadata includes fields for dates indicating when the dataset was uploaded or last modified, at what frequency does the data change, plus a flag for the data owner to indicate the completeness of the dataset. We added a discussion board add-on to each dataset enabling end-users to leave comments and feedback on how they perceive the quality of the data.

Once the portal is launched to the broader public, we anticipate data coming from these channels to help us measure the data and the portal on an ongoing basis.

To get more immediate feedback on the pilot, we organised two-hour focus group sessions involving data owners and consumers for individual datasets who were familiar with using the existing web applications to get access to data. Comparing the older individual websites against the portal approach, we conducted a usability and feedback focus group to assess if the current and future needs of both groups are being met with the new approach. The focus groups took a structured approach and all responses were documented. The feedback from these sessions, was reviewed and evaluated. Overall, the responses from the focus groups, from both data owners and consumers, were positive, although many interesting discussions were generated and ideas for enhancements raised, which were used during the ACT stage to determine next steps.

5.4 ACT Stage

This stage took the feedback and reflections from the focus groups and determined what to do about it. From the set-up and sustainability plan, there were two components of possible action in this stage (*Table 25)*.

Stage	Context	Component	Action taken
ACT	Dataset	Promote use	From the focus groups, we saw the benefit of meeting directly with consumers to demonstrate use. It was identified that a full outreach campaign needs to be planned in the next cycle.
	Dataset	Data retention, archival and disposal	From this cycle, none of the data was flagged as requiring decommissioning. Feedback was raised to add an automated system to check the last modified data of datasets to detect if any are going stale. This will be part of a future cycle.

Table 25: ACT Stage - determining next steps

Missing from the list is the activity of identifying what from the STUDY stage to implement and adding it to the plan, in addition to promoting or decommissioning datasets. To borrow a term from the Scrum Agile development methodology, what we found ourselves doing was backlog grooming. Gathering and documenting feedback and suggestions, prioritizing them and lining them up for the next stage in the cycle, back to the PLAN stage, where plans are put in place for the next subset of the backlog. Two items that raised a lot of discussion and feature heavily in our backlog were data presentation and linked data.

5.4.1 Data presentation

In the focus groups, an area where disparate views surfaced was in how the data should be presented. In one case, the data owner was concerned that the new user interface of the portal was not as visually attractive or engaging as his previous stand-alone website could be. He preferred the freedom that custom development afforded him in allowing him to brand the web page as he wished, embedding an interactive map and power BI reports and enabling him to present data aggregated as he deemed fit. However, the consumers of the data indicated in their feedback, a strong preference for access to raw data in an open format over online filtering tools/visualizations. Their need is to download the data and use it in their own advanced modelling software, so any effort in making the web pages attractive, does not contribute to their task. They preferred the structured, consistent approach offered by the portal and the ability to download the data via APIs. They did express a request to label the

individual data columns with more meaningful names or have a data dictionary available to explain.

5.4.2 Linked Data: Taking data from * to *****

Another area of discussion was how to leverage linked data. Our initial goal was to elevate the five selected datasets from their existing one-star status to five-star linked Open Data. However, during the implementation, four of the five data owners could not see a business benefit from their perspective in investing in that extra effort and were happy to publish their datasets at 3 or 4 stars. One owner expressed a strong desire to publish as linked Open Data, but it was not clear which domain vocabularies or metadata standards to use or what to actually link to. This will be further investigated in the next cycle.

With a new list of requirements, we then were ready to start the planning stage again, adding the new requirements to the backlog and preparing the next set of datasets for publishing.

6. Conclusion

We began this thesis with a practical business problem in mind, how best to establish an Open Data portal for an organisation with limited resources and at its infancy in data management strategy and practices so as to reap the benefits of transparency, consistency and efficiency. Our objective was to create an Open Data Portal Management framework that could be followed for the initial set-up of the portal but could also be applied for the ongoing support and sustainability of the portal and management of the data.

The overarching research question was what should this framework look like? What elements or components should be included? Guided by what principles, best practices, or plans?

To answer this, we first had to familiarize ourselves with Open Data; the terminology, the history and the guiding principles for opening or sharing data. We discovered Open Data specific principle schemes and the broader FAIR principles for sharing data. Comparing and contrasting the principle schemes, we combined and consolidated these to produce an aggregated list of actionable principles, to ensure openness and FAIRness of the data for our project.

Next, we explored how these principles have been applied in practice by assessing a collection of existing Open Data portals. We reviewed portals of both multilateral organisations and governments to assess how they uphold the principles. To assess the portals in an objective way, we defined a set of metrics based on the measurable principle categories, namely *technically open, legally open, level of openness* and *(meta)data management*. The outcome of the assessment was a better understanding of the functioning and maturity of existing portals and a derived set of requirements and features for our portal implementation.

To answer our final research question on what the key elements or components of plans for data management and sustainability are, we researched different approaches to follow to ensure best practice in managing and sharing data. We compared existing plans for data management, Open Data publication and sustainability. By aggregating and combining the elements of these plans, we produced a comprehensive set-up and sustainability plan.

Our framework is tested by bringing together and implementing the principles, portal features and plan components in different stages of the iterative PLAN-DO-STUDY-ACT

process. Following a full cycle, going through all four stages and acting on the components in the plan, we delivered a portal infrastructure with five datasets migrated. The framework is now in place to follow through on migrating the remaining datasets identified in the data audit. It can also be applied for new datasets.

To conclude, this thesis has led to the creation of an Open Data framework, that can be adopted by organisations facing similar data challenges to OrgX. The framework is comprised of the following resources:

- An aggregation of Open Data principles into assessment categories (*Table 8*);
- A comprehensive listing of desired portal features and requirements (*Table 15*); and
- An aggregation of components for a set-up and sustainability plan (Table 21).

These are this thesis's contributions to the research community.

6.1 Limitations

One of the largest limitations we faced was time. To fully evaluate the framework, more iterations of the process are required, whereby more datasets are published, measured and promoted. With more cycles, we could continue to grow the portal by running the cycle for the remainder of the existing datasets, and officially launch a portal outreach and promotion campaign.

Another limitation, but with the framework, is that it does not specifically target data quality but instead implies that by following good sustainable data management practices and principles, quality data will be an outcome. If more attention to data quality is warranted, then this could be addressed by researching and consolidating data quality plans into the set-up and sustainability plan.

6.2 Future Work

An area to explore further is the role automation could play in ensuring sustainable data management. There is an opportunity to automate more around the process of collecting, cleaning and publishing the data, which in turn makes the initiative more sustainable, by relying less on manual processes. A topic that surfaced from the research but was not

explored further due to time constraints is the concept of machine actionable DMPs (Simms, et al., 2017). The concept is that a DMP could be automatically generated and shared, by embedding steps into the workflows that create the data or by facilitating interoperability and exchange of data between systems that manage the data. There are principles already published (Miksa, et al., 2019) that could be explored further for amalgamation into our principle list.

Another area for future work, is to explore how linked data could be leveraged by an organisation like OrgX and what additional benefits could be achieved. This would depend on more standards being adopted around (meta)data, to facilitate interlinking. Adopting metadata standards promises the added benefit of improved data usability (Mana & Sasiprabha, 2019). Researching which metadata standards to adopt to facilitate linking data is something that could be pursued now that the base portal is in place, with support for extensible metadata and linked data formats.

Thanks to the cyclical nature of our implementation plan there is a process in place to continuously improve on the existing framework, portal and data management, based on any future research.

Bibliography

The Open Government Working Group, 2007. *Open Government Data Principles.* [Online]
Available at: https://public.resource.org/8_principles.html
[Accessed 06 07 2021].

Attard, J., Orlandi, F. & Auer, S., 2016. *Data Driven Governments: Creating Value Through Open Government Data.* Berlin, Springer-Verlag.

Below, B., 2015. *Shedding light on government, one dataset at a time.* [Online]
Available at: http://oecdinsights.org/2015/05/29/shedding-light-on-government-one-dataset-at-a-time/
[Accessed 10 05 2021].

Berners-Lee, T., 2009. *LinkedData.* [Online]
Available at: https://www.w3.org/DesignIssues/LinkedData
[Accessed 30 June 2021].

Berners-Lee, T., 2012. *5 ★ OPEN DATA.* [Online]
Available at: https://5stardata.info/en/
[Accessed 09 05 2021].

Boulton, G., 2014. The Open Data Imperative. *Insights: The UKSG Journal,* 27 (2)(DOI: http://doi.org/10.1629/2048-7754.148), pp. 133-138.

Carrara, W. et al., 2018. *Open Data Goldbook for Data Managers and Data Holders.* [Online]
Available at: https://data.europa.eu/sites/default/files/european_data_portal_-_open_data_goldbook.pdf
[Accessed 02 05 2021].

Charalabidis Y., Z. A. A. C. J. M. L. T. F. E., 2018. The Objectives of Open Data. In: *The World of Open Data.* s.l.:Springer, Cham, pp. 1 - 9.

Christodoulou, P. et al., 2018. Data Makes the Public Sector Go Round. In: P. P. e. al., ed. *Electronic Government. EGOV 2018. .* s.l.:Springer, Cham., pp. 221-232.

Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S., 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data,* p. 54.

European Commission Expert Group on FAIR Data, 2018. *Turning FAIR into reality,* s.l.: European Commission.

European Commission, n.d.. *Horizon 2020 Online Manual - Data Management.* [Online]
Available at: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[Accessed 30 05 2021].

Fraizer, A., 2021. Data Helps World Go Round. *The Journal of Emergency Dispatch.*

G8 Science Ministers, 2013. *G8 Science Ministers Statement.* [Online]
Available at: https://www.gov.uk/government/news/g8-science-ministers-statement

G8, 2013. *G8 Open Data Charter and Technical Annex.* [Online]
Available at: https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[Accessed 30 04 2021].

Hammer, C., 2019. Open data stakeholders: Multilateral organisations.. In: *The state of open data: Histories and horizons.* Cape Town and Ottawa: African Minds and International Development Research Centre. http://stateofopendata.od4d.net, pp. 406-417.

Hasnain, A. & Rebholz-Schuhmann, D., 2018. Assessing FAIR Data Principles Against the 5-Star Open Data Principles. In: A. Gangemi, ed. *The Semantic Web: ESWC 2018 Satellite Events.* s.l.:Springer, Cham., pp. 469-477.

Hevner, A., March, S., Park, J. & Ram, S., 2004. Design science in information systems research. *MIS Quart., 28:,* pp. 75-105.

Higman, R., Bangert, D. & Jones, S., 2019. Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights,* 32(1), p. 18.

Hook, D., 2021. The Rise and Rise of Open Data. *Information today,* 34(8), p. 7.

International Open Data Charter, n.d. *Open Data Charter.* [Online]
Available at: https://opendatacharter.net/
[Accessed 30 05 2021].

Janssen, M., Charalabidis, Y. & Zuiderwijk, A., 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29 4, pp. 258-268.

Knippenberg, L. v., 2020. *Open Data Maturity Report.* [Online]
Available at:

https://data.europa.eu/sites/default/files/edp_landscaping_insight_report_n6_2020.pdf
[Accessed 15 05 2021].

Kucera, J. & Chlapek, D., 2014. Data, Benefits and Risks of Open Government. *Journal of Systems Integration, Vol 5, No 1.*

Kučera, J., Chlapek, D., Klímek, J. & Nečaský, M., 2015. *Methodologies and best practices for Open Data publication.* Czech Republic, CEUR-WS.org, pp. Vol 1343, pp 52-64.

Luna-Reyes, L. & Najafabadi, M., 2019. The US open data initiative: The road ahead. *Information Polity,* pp. 163 - 182.

Mana, S. C. & Sasiprabha, T., 2019. A Study on Various Semantic Metadata Standards to Improve Data Usability. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS),* Issue doi: 10.1109/ICCIDS.2019.8862044, pp. 1-4.

Martin, C., 2014. Barriers to the Open Government Data Agenda: Taking a Multi-Level Perspective. *Policy & Internet,* 6(3), pp. 217-240.

McAfee, A. & Brynjolfsson, E., 2012. Big Data: The Management Revolution. *Harvard Business Review,* pp. 61 - 68.

McDermott, P., 2010. Building Open Government. *Government Information Quarterly - GOVT INFORM QUART,* pp. 401 - 413.

Miksa, T., Simms, S., Mietchen, D. & Jones, S., 2019. Ten principles for machine-actionable data management plans. *PLOS Computational Biology,* 15(3): (e1006750).

Mons, B. et al., 2017. Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. *Information Services & Use,* 37(1), pp. 49-56.

Obama, 2009. *MEMORANDUM on Transparency and Open Government.* [Online]
Available at: https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government
[Accessed 30 04 2021].

ODI, 2012. *About the ODI.* [Online]
Available at: https://theodi.org/about-the-odi/
[Accessed 26 05 2021].

OKF, 2015. *Version 2.1.* [Online]
Available at: https://opendefinition.org/od/2.1/en/
[Accessed 28 05 2021].

Open Data Charter, 2015. *Open Data Charter Principles.* [Online]
Available at: https://opendatacharter.net/principles/
[Accessed 30 04 2021].

Open Knowledge Foundation, n.d.. *Open Knowledge Foundation.* [Online]
Available at: https://okfn.org/
[Accessed 10 05 2021].

Parycek, P., Höchtl, J. & Ginner, M., 2014. Open government data implementation evaluation. *Journal of Theoretical and Applied Electronic Commerce Research,* Vol. 9(No. 2), pp. 80-99.

Publications Office of the European Union, 2020. *Recommendations for Open Data Portals: From Setup to Sustainability.* [Online]
Available at:
https://data.europa.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf
[Accessed 23 05 2021].

Ramachandran, R., Bugbee, K. & Murphy, K., 2021. From Open Data to Open Science. *Earth and Space Science,* 8(5), pp. 2333-5084.

Share-PSI 2.0, 2016. *Best Practice: Develop an Open Data Publication Plan.* [Online]
Available at: https://www.w3.org/2013/share-psi/bp/odpp/
[Accessed 01 July 2021].

Simms, S., Jones, S., Mietchen, D. & Miksa, T., 2017. Machine-actionable data management plans (maDMPs). *Research Ideas and Outcomes,* 3:(e13086).

Simperl, E., 2020. *Analytical Report 8: The Future of Open Data Portals.* [Online]
Available at: https://data.europa.eu/sites/default/files/edp_analyticalreport_n8.pdf
[Accessed 02 06 2021].

Strohmaier, R., Schuetz, M. & Vannuccini, S., 2019. A systemic perspective on socioeconomic transformation. *Journal of Industrial and Business Economics.*

Sunlight foundation, 2010. *Ten Principles for Opening Up Government Data.* [Online]
Available at: https://sunlightfoundation.com/wp-content/uploads/sites/2/2016/11/Ten-Principles-

for-Opening-Up-Government-Data.pdf
[Accessed 15 05 2021].

Susha, I. et al., 2015. Critical Factors for Open Data Publication and Use: A Comparison of City-level, Regional, and Transnational Cases. *JeDEM - eJournal of eDemocracy and Open Government.*

Tauberer, J., n.d. *The Annotated 8 Principles of Open Government Data.* [Online]
Available at: https://opengovdata.org/

UN, 2020. *Data Strategy of the Secretary-General.* [Online]
Available at: https://www.un.org/en/content/datastrategy/images/pdf/UN_SG_Data-Strategy.pdf
[Accessed 30 05 2021].

Vaishnavi, V. & Kuechler, B., 2004. Design Science Research in Information Systems. *Association for Information Systems.*

Vaishnavi, V. & Kuechler, W., 2008. *Design Science Research Methods and Patterns. 1st Edn.,.* s.l.:Auerbach Publications, Taylor and Francis Group, Boca Raton, FL., .

Walker, J., Ibanez, L.-D. & Simperl, E., 2020. *Sustainability of (Open) Data Portal Infrastructures Summary Overview and Key Recommendations.* [Online]
Available at: https://data.europa.eu/sites/default/files/sustainability-data-portal-infrastructure_0_summary.pdf
[Accessed 08 05 2021].

Wieczorkowski, J., 2019. Open Data as a Source of Product and Organizational Innovations.. *Proceedings of the European Conference on Innovation & Entrepreneurship,* p. 1119–1128.

Wilkinson, M. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3(160018).

Williams, M., Bagwell, J. & Zozus, M. N., 2017. Data management plans: the missing perspective. *Journal of Biomedical Informatics,* pp. 130 - 142.

Zeleti, F., Ojo, A. & Curry, E., 2016. Exploring the economic value of open government data. *Government Information Quarterly,* Vol. 33 (No. 3), pp. 535-551.

Appendix A: Multilateral Portal Assessment

No.	Organisation	Technically open	Legally open	Level of openness	Metadata mgmt	Multilingual	No. datasets	No. publisher	No. topics	Tech	Download formats	Download	API	Linked	(Meta)data mgmt	Legally Open comments	Comments or Features
1	FAO	3	3	3	2	UN Lang	76	1	15	Custom Build - OpenSource	CSV, XLS, XML, JSON	bulk	No evidence	No evidence	Catalog of data searchable by topic, filter by country, elements, items, years, metadata but	Terms of use (Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO)	Nice feature of comparing datasets with visualizations, FAQ, Aggregates from National Stats offices
3	ILO	4	3	3	3	EN/FR/FP				Custom Build - OpenSource -R package	CSV, XLS, GZ	yes	Yes, JSON, SDMX,	No evidence	searchable by freq and DB. Fields include Last update, Description. No mention of licence or	For non commercial must cite, commerical needs to get in touch - not machine readable licence	topics, and promotion of events and community building content, Lots of tools available on site = R package (RILOStat). No count of datasets
4	IMF	2	3	2	2	UN Langs				Custom Build - Microsoft	TSV,XLSX,	yes	No evidence	No evidence	Catalog of different databases, very little metadata - Description,	Nothing on datasets, users may download, extract, copy, create derivative works, publish, distribute obtained from IMF Sites, Like a CC-Sharealike	Catalog more centered around publications or websites, Need to dig to find the data. No preview of data unless in PDF pub, download only
6	IOM	2	4	2	1	EN				Custom Build - Microsoft	XLS,	yes in bulk	No evidence	No evidence	No catalog, no URIs, or data previews, found metadata from clickiong on map to get data on incicdents	CC-Attribution 4.0	PowerBI visualizations, Data is aggreagated and presented as a story. Only bulk download available as one huge batch. No metadata or catalog. Licence is applied site wide. Web content.
8	OCHA	4	5	4	3	EN	19106	1460	254	Built on CKAN	Many formats - CSV, XLS, no RDF/linked data formats	yes	only for 2 datasets	No evidence	Catalog filterable by location, org, format, tags, licence; metadata contains source, freq., licence, methodology	Many licence types - majority are CC-by Attribution, 1000 are public domain.	New feature to display archived datasets, lots of tools for working with data, datasets with quick charts, showcases, URL to the CSV file visible on preview page. Nice to show volume on home page
9	OECD	3	3	3	2	EN/FR/JA	7771		12	Custom Build	CSV	yes	yes, developer API in JSON, XML	No evidence	catalog filterable by type, topic, hard to find metadata, no licence, URI	Terms of use for Data. Not machine readable. Share alike	Find, compare and share the latest OECD data, browse by topic or country, a lot of data in PDF publication format, but possible to find datasets in tabular format. Lots of visualizations. Multilingual for lables only, not content
10	UN	4	5	4	3	EN	22			Custom Build - Microsoft	XLS,	yes	yes REST/SOAP web service	No evidence	SDMX standard, catalog of datamarts, metadata shows source, short desc and last update	free of charge and may be copied freely, duplicated and further distributed provided that UNdata is cited as the reference. Not machine readable	Confusing terminology - Datamart vs dataset vs database. Data harvested from other sites, with view and preview function, and metadata per database. Like that they show site usage
11	UN Habitat	4	0	4	3	EN	139		16	Custom Build - Microsoft + ARCGIS	CSV	yes	yes, GEOService and GEOJSON	No evidence	catalog filterable by content type, last updated, tags, metadata contains summary, dates, attributes, tags	No licence provided - Request permission to use	Thematic search, Featured applications &content, highlights and tweets. Nice imagery. No Licence provided, ASPX pages, can see related datasets - nice
12	UNCTAD	3	3	3	2	EN/FR	1791		14	Microsoft - ASPX Pages	ZIP	yes bulk	No evidence	No evidence	Urillable catalog (folders), metadata includes title, author, category, keywords, desc, source,- no licence, uri	CC Attribution 3.0 on website	Accessible via ASP pages but needto download bulk to get CSV, documentation included, nice infographics
13	UNDP(1)	4	0	3	1	UN Lang				Custom built open source	CSV, XML, JSON, HTML	yes	Dquery access to data	No evidence	Catalog of projects, hard to follow metadata	No mention of terms of use or licence. Need to goto source	Beta site pulling data from IATI registry (CKAN) and displaying in visual appealing way, huge no. of filters in the search
14	UNDP(2)	4	4	3	3	EN	14k			Custom built html	JSON, ZIP (CSV), XML	yes bulk	Could use JSON links	No evidence	Searchable catalog of records on project data, metadata focus, recipient, donors, year, SDG	Creative Commons Attribution 3.0 IGO License (CC-BY 3.0 IGO)	Nice feature of being able to embed visualization of a specific record into another site
15	UNDP(3)	4	4	3	2	EN			14	Custom built	CSV, PDF	yes	REST API JSON format, requires registration	No evidence	Search by topic(dimension) and then select dataset, see source and desc.	Creative Commons Attribution 3.0 IGO	Documentation on site provides details on how the data is aggreagted (from sources), methodology, transformation.
16	UNECE	3	0	3	1	EN/RS	48		8	Unknown	save table as CSV, XLS,Jstat	yes	No evidence	No evidence	No clear catalog, metadata includes contact info, footnotes,	Nothing visible	Drill down into each topic with multiple filters
17	UNECLAC	3	0	3	2	SP/EN/PT	unknown			Custom build - ASPX	HTML, XLS, XML	yes	CEPAL STAT REST API JSON, require	No evidence	No clear catalog, more a drill down search into databases but metadata contains source,	Terms of use do not explicitley state how data can be reused	Has a menu item Open Data/API which gives coprehensive descripiton of API but no licence info. Search does not indicate no. of results.
18	UNEP(1)	4	0	3	3	EN/FR	654		30	Custom build PHP	CSV,HTML,XLS,ESRI Shapefile	yes	REST/JSON webservice, SOAP, WFS yes	No evidence	Catalog of dataset with rich metadata, status, update freq,topic,keywords, contact, method	Nothing visible	Extensive search, drilling down options to view data on map,graph,table,metadat,download. But no terms or use or licence info

No.	Organisation	Technically open	Legally open	Level of openness	Metadata mgmt	Multilingual	No. datasets	No. publisher	No. topics	Tech	Download formats	Download	API	Linked	(Meta)data mgmt	Legally Open comments	Comments or Features
20	UNESCAP	4	0	4	3	EN/FR/SP/IT/ND/DE/2 asian	210		17	OpenSource .STAT Suite Platform	CSV, XLS	yes	yes SDMX std, RESTFUL URL, JSON	No evidence	Catalog of SDG data organised by SDG/sub group and then data by country back to 2015,	Nothing visible	Opensource s/w, API, open formats but have to provide email for access to URL+embed code of data snapshots, metadata described on another site, Lang does not change data
21	UNESCO	4	2	3	2	EN/FR	340		5	.STAT technology	CSV, SDMX, XML	yes	yes SDMX	No evidence	Drill down by theme, sub-theme and datasets. Metadata shows source, no URIs/licence info	Terms of use - personal or public non-commercial use (not clear, machine readable)	Access limited to browsing by theme or popular queries,limited metadata, lengthy documentaiton on how to use site, can chart data
22	UNFPA	1	1	2	1	EN/FR/ES				Custom Build - Microsoft	no	No	no evidence	No evidence	Series of dashboards - no catalog	Terms of use - copy Materials for User's personal, non-commercial use, without any right to resell, redistribute, compile or create derivative works	Data presented in nice vizualizations, download in bulk to PDF-no access to raw data
23	UNHCR	3	4	2	1	EN/FR/ES/AR				Custom Build - POWER BI reports	XLS from power BI	yes - limited	Registration required	No evidence	No catalog, metadata not clear	Creative Commons Attribution 3.0 International License.	Series of webpages with info per topic - documents, tweets, map, dashboards. Registration required for API, raw data not accessible- data visualizations, dashboards and infographics published as PDF
24	UNICEF	4	4	3	3	EN	51		38	Custom build	XLS, CSV	yes	REST JSON SDMX server	No evidence	Catalog searchable by type, topic, year, SDMX metadata	Creative Commons Attribution 3.0 IGO	Querying UNICEF data warehouseRich metadata and URI not obvious, url access. SDMX web service API. Area about Open Data
25	UNIDO	2	1	2	1	EN				Custom Build	ZIP	yes	No evidence	No evidence	catalog of databases with some retricted, no real metadata	Complicated terms of use, not clear where and how can use data;	Page linking to 12 databases, 8 requiring authetication. Limited access to some databases, need to auth and pay, uses Data viz
26	UNODC	2	1	2	2	UN Lang	65		12	Custom with Tableau	XLS, PDF	yes	No evidence	No evidence	Catalog of datasets searchable by topic, country,city,year; no licence nor URI	Copyright refers to terms of use - link not working. Then on another page: Material for User's personal, non-commercial use, without any right to resell or redistribute them or to compile or create derivative works	Data presented aggregated in tableau, bulk download in excel only, microdata available for researchers only on request
27	UNOPS	3	0	3	2	EN			12	Custom Build - Tableau for visualization	CSV	yes	no evidence	No evidence	Catalog of projects, metadata at project level	Terms of use page gives an error	UNOPS project data Linked on Country, but download in bulk only, Labelled Open Data. Could see a list of contracts and list of projects
28	UNSD	3	0	3	2	EN/CH/FR/ES/RS				Custom Build - Microsoft ASPX	XLS,CSV, XML, PDF country snapshots	yes	no evidence	No evidence	Folder drill down,rich metatdata but separate to data,separate drill down. Includes source,method,proce	Nothing visible. No terms of use	MDG indicators, news feature but last news item from early 2016. Latest data also from 2016 so perhaps obsolete site. Metadat is printable rather than machines readable
29	UPU	1	1	1	0	EN/FR				Unknown	TXT	Kind of - select export	no evidence	No evidence	No catalog nor metadata visible;simply querying a database of stats	Copyright -None of the materials provided on this website may be used, reproduced or transmitted, without permission	Organised around a query of database of postal statistics. Query by request type and attributes such as year
30	WB	5	5	5	4	EN/FR/ES AR/CH	19490 items in catalog			Unknown	EXCEL, CSV	yes bulk download	Yes XML	Yes, 52 datasets	filterable by country/licence/ datatype/resource/rating, Metadata as RDF.catalog contains more than	different types of licence marked per dataset, some are CCO,majorit CC-BY 4.0 and some special microdata research licence	and open access to global development data.News, impact stories, what you can do with open data section, uses 5 star data rating, share on social media platforms, shows citations. PHP
31	WFP	3	0	2	2	EN			5	Custom Build -Dataviz	CSV, PDF	yes	no evidence	No evidence	cataolg of reports with minimal metadata	Nothing mentioned	explorers displaying visuals, not data, no licence/reuse info - data
32	WHO	3	3	3	2	UN Lang			35	Custom built data sync platform, Microsoft Power BI visualizations	PDF,CSV	yes	Yes OData (Open Data Protocol), ATHENA	No evidence	Search by themes and get related topics, drill down and get to page with visualizations, data, metadata, related info and related indicator tabs; follows indicator metadata registry list	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Intergovernmental Organization (CC BY-NC-SA 3.0 IGO) licence. But then separate terms and conditions about data that cntradict	Searchable data collections, each has its own 'website' differetn terms and conidtions link, look & feel, structure. open access policy - From 1 January 2021, all WHO-authored and WHO-funded articles that are submitted for publication in peer-review journals must be published in an open-access journal or on an open-
33	WIPO	2	3	3	2	UN Lang				Custom Build	CSV	yes	no evidence	No evidence	Link to IP statistics and annual report on country sites. Data centre is searchable by indicator. Can then chart or download the result.	Open Access Policy, (CC IGO) licenses. But then: WIPO's statistical data, users agree not to republish or commercially re-sell WIPO's statistical datasets.	metadata not visible, apart from filter criteria, contradcitory terms of use,no API, not machine readable

Appendix B: DMP Topics

Aspects of data values	Aspects of data elements	Aspects of the dataset	Applicable to data values, data elements or datasets
Accuracy	Data definition	Audience	Access to data
Attribution of entries and changes	Instrumentation and calibration	Description of types of data captured	Workflow
Contemporaneity of data collection	Data standards	Relationship to existing data	Dataflow
Legibility		Data ownership	Data quality assurance
Original		Intellectual property rights	Measuring data quality
Traceability		Privacy and confidentiality	
Data integrity		Organisation of data	
		Organisation of shared data	

Aspects of systems used to manage data	Aspects of data disposition	Project management aspects
Data storage and back-up	Preservation	Data stewardship responsibility
Security	Curation of data	Milestones and timelines
Software validation	Formatting data for reuse and redistribution	Required resources
Load assessment	Retention of data	Project specific functionality needed
	Archival of data	Specified timing of data submission
	Archival of derived datasets	Reporting project status
	Plan for disposal of data	Information products/other resources
	Essential documents	Legal requirements

Table 26: DMP topics categorized by seven aspects, extracted from (Williams et al., 2017)

Appendix C: OrgX Open Data Portal

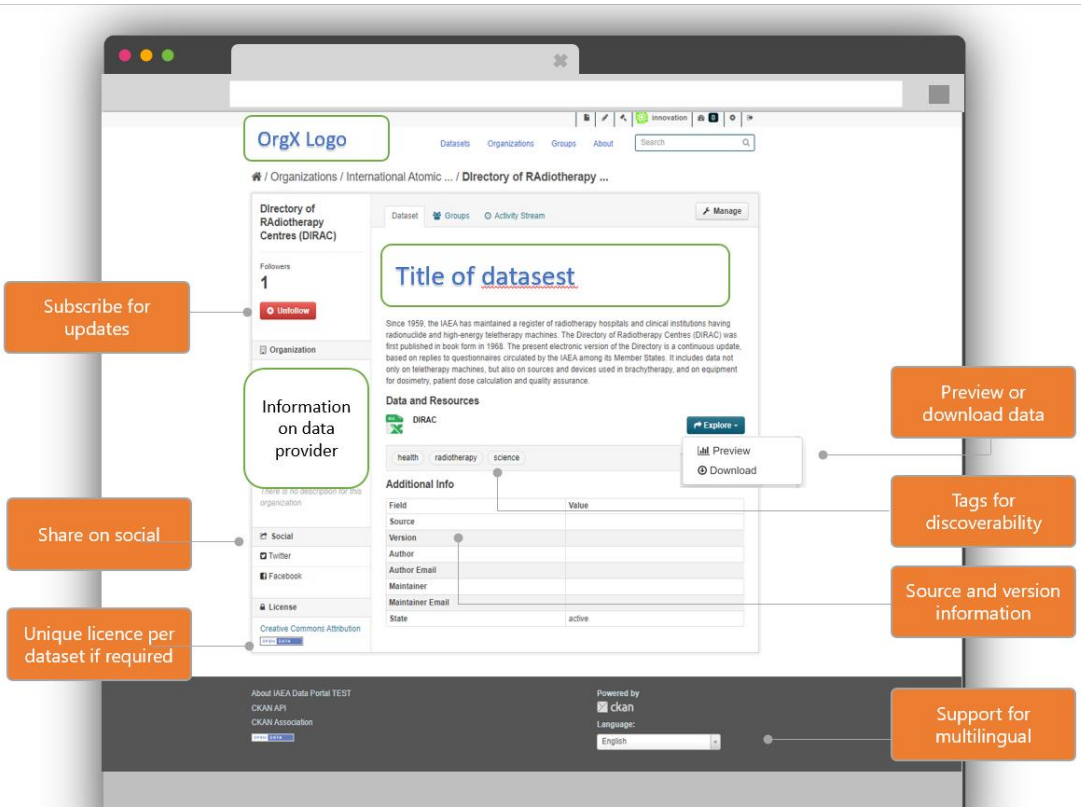

Figure 15: Portal features for an OrgX dataset

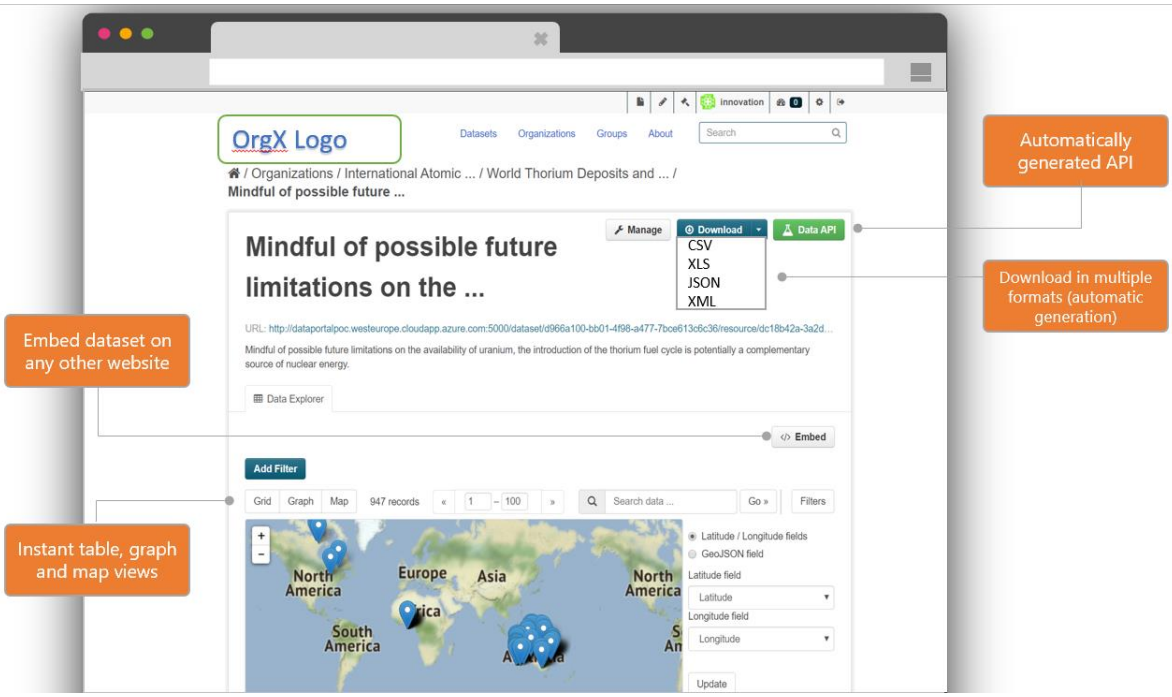

Figure 16: Options for data exploration from OrgX portal