Sustainability of (Open) Data Portal Infrastructures

Summary Overview and Key Recommendations

This study has been prepared by the University of Southampton as part of the European Data Portal. The European Data Portal is an initiative of the European Commission, implemented with the support of a consortium led by Capgemini Invent, including Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, 52North, Time.Lex, the Lisbon Council, and the University of Southampton. The Publications Office of the European Union is responsible for contract management of the European Data Portal.

For more information about this paper, please contact:

European Commission
Directorate General for Communications Networks, Content and Technology
Unit G.1 Data Policy and Innovation
Daniele Rizzi – Policy Officer
Email: daniele.rizzi@ec.europa.eu

European Data Portal
Gianfranco Cecconi, European Data Portal Lead
Email: gianfranco.cecconi@capgemini.com

Esther Huyer
Email: esther.huyer@capgemini.com

Written and reviewed by:
Johanna Walker
Email: j.c.walker@soton.ac.uk

Elena Simperl
Email: elena.simperl@kcl.ac.uk

Luis-Daniel Ibanez
Email: l.d.ibanez@soton.ac.uk

With Laura Koesten, Mark Frank, Jacqui Ayling, Peter West, Eric Costa and Sarah Hewitt (University of Southampton)

Last update: 02.03.2020
www: https://europeandataportal.eu/
@: info@europeandataportal.eu

DISCLAIMER

By the European Commission, Directorate-General of Communications Networks, Content and Technology. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use, which may be made of the information contained therein.

Luxembourg: Publications Office of the European Union, 2020
© European Union, 2020

OA-02-20-162-EN-N ISBN: 978-92-78-42145-8 doi: 10.2830/086159

Note: this document is part of a series of research reports developed on the topic of "Sustainability of (open) data portal infrastructures", all of which are available on the European Data Portal at https://www.europeandataportal.eu/en/impact-studies/studies .

The series is made of the following reports:

1. A Summary Overview
2. Measuring Use and Impact of Portals
3. Developing Microeconomic Indicators Through Open Data Reuse
4. Automated Assessment of Indicators and Metrics
5. Assessment of Funding Options for Open Data Portal Infrastructures
6. Open Data Portal Assessment Using User-Oriented Metrics
7. Leveraging Distributed Version Control Systems to Create Alternative Portals

Table of Contents

1. Introduction

1.1 What is Portal Sustainability?

Data portals are intrinsically enmeshed with the idea of open data. Before Directive 2013/37/EU encouraged governments to open their public sector information to the public as open data, there was very little need for a place to which they could promote their data, and where prospective users could find it. Open data, in removing the friction of financial costs, of various copyrights and, importantly, of the necessity of fostering relationships with the data owners in order to attain access, has most certainly been one of the driving reasons for the huge growth in interest in and development of data-utilising businesses and activities. Today's landscape looks very different to 2009, when aporta.es (now datos.gob.es) was launched, or 2011, which saw the launch of data.gouv.fr. Portals have proliferated - Open Data Portal Watch monitors 278 globally.[1]

In this potentially crowded market, there are some internal issues with portals that may challenge their sustainability. There is remarkably little variation in functionality, with the vast majority aimed at meeting the needs of publishers, rather than the potentially wide variety of users. Some of the choices made in publishing may be problematic – for instance, there is often more transactional than reference data on portals. Many users require the reference data to understand the transactional data. [2] This leads to the first aspect of sustainability – the operational. Portals must continue to be fit for purpose and future-proofed. This ability to internalize the ability to develop and grow is a key aspect of sustainability.

Additionally, the external landscape has changed. In addition to open government data portals at national, regional and municipal levels there are a (very limited) number of non-governmental open data portals, a growing number of commercial data marketplaces and even social enterprises providing accessible data infrastructure for a combination of open and paid-for data.

Further while open data advocates and adherents may continue to be purists, the vast majority of data users and potential users do not have similar concerns. They may prefer open data in some situations, they may prefer paid-for data in others, in particular where there are major data quality issues. By and large, they do not reject or select data on the basis of its open nature or otherwise. Open data, as noted above, has brought us to this point, but the future is more agnostic and embracing of the full data spectrum.

Another aspect of sustainability is, naturally, funding. While economists such as Open Knowledge founder Rufus Pollock have clearly stated the macro argument for the benefit of open data[3], the micro-economic arguments remain confounding. If open data is both free as in 'free speech' and, due to the almost zero cost of replication over the web, free as in 'free beer', how is revenue acquired? And without revenue, who is responsible for the costs – which are, in some cases, substantial?

[1] http://data.wu.ac.at/portalwatch

[2] As a very simple example, if a transaction – an event of some kind - recurred against a code ABC, the reference data to explain that code would be required to make sense of the whole.

[3] https://www.repository.cam.ac.uk/bitstream/id/534321/0920.pdf/;jsessionid=134629A629D53FA963C4EC7B34D28B89

Governments at the national, regional and municipal level have largely funded this initial step but these pockets are not endlessly deep. Further, a perception of continued state funding of portals may invite criticism if benefits are not clear. Lastly, the European Commission (2006), states, “a project is sustainable when it continues to deliver benefits to the project beneficiaries and/or other stakeholders for an extended period after the EU’s financial support has been terminated.”[4] If the EDP aims to become fully sustainable, this is a relevant criterion to consider.

Therefore, there are two sustainability questions around funding of portals – the macro, or the proof of how portals positively impact on taxation, employment and other aspects that early supporters of open data posited, and the micro, or how portals are funded day to day. Both of these questions are addressed in this work, in Sections 1 and 3.

Portals, therefore must be sustainable both operationally and financially. In the context of this project, therefore, sustainability is understood as not only the effort to maintain and improve the Data Portal Infrastructures, but in particular to understand what the Data Portal Infrastructure needs to offer to the user, without which there is no reason for portals to exist. A sustainable project would therefore attract more data providers and users on the supply and demand side as well as increasing the availability, and crucially use, of high value data sets by seeking to understand its position both up and down the value chain.

1.2 Previous Work on Sustainability

The previous iteration of the European Data Portal has delivered several studies on sustainability, including ‘Recommendations for Open Data Portals: From Set up to Sustainability’ (2017) and ‘Ensuring the Economic Sustainability of Data Portals: Understanding Impact and Financing’ (2018). We build and expand on this work, starting from their main recommendations:

- measuring impact via new approaches;
- applying technical methods for tracking reuse;
- adapting the Common Assessment Framework; and
- exploring freemium funding models.

1.3 Gaps Identified

Our work is organised in four themes:

1. Indicators and Metrics for Assessing the Economic Impact of Portals - Measurement has always been key to portals. This theme largely focuses on quality of data (adherence to 5* of open data; timeliness; ‘cleanliness’ and other indicators), and also around portal-based activity around datasets (views, downloads). The research goes beyond this activity to understand the relationship between portals and re-use.

[4] European Commission Directorate-General Education and Culture (2006) “Sustainability of international cooperation projects in the field of higher education and vocational training - Handbook on Sustainability”. Luxembourg: Office for Official Publications of the European Communities

2. Automated Assessment of Indicators and Metrics - Automation is key to scalability; it ensures that assessing indicators and metrics can be carried out whenever needed on large samples of datasets and activities. Starting from the recommendations from theme 1, we devise a methodology to map high-level indicators and features to observable features in portals and platforms, assess values of indicators and metrics that correlate with re-use and predict their evolution.

3. Assessment of Funding Options for (Open) Data Portal Infrastructures - A key aspect of sustainability is that of funding – ensuring there is sufficient funding to not only sustain current activities but also the potential to fund growth activities. Work has already been done by the European Commission on Digital Infrastructure Sustainability Solutions Framework looking at wholly private, NGO sector and public ownership and further mixed options, and this should be leveraged for portals.

4. Exploring Alternatives to Portals - As Google Dataset Search[5] begins to show potential for a greater data discovery role, portals need effective guidelines for identifying and strengthening their other functions, and ensuring they are utilising the technologies that will facilitate this.

1.4 Report Structure

The resulting work addressing these gaps is published as 6 separate reports. In line with our aim to make this research more applicable to a wider group of users across policy, operations, technology and other stakeholders, this summary overview highlights the key findings and recommendations of each individual piece of work, while making the complete reports available separately in their entirety. These can be read - and in many cases implemented - in any order. While we appreciate policy makers require a broad overview of the entire landscape, many portal owners and developers are focusing on one aspect at a time, and we hope our approach supports their needs equally.

The summary overview is structured as follows:

Section 1 (Indicators and Metrics for Assessing the Economic Impact of Portals) provides an overview of the current landscape of measuring the economic impact of open data portals. This remains a formidable task. It falls into two main subtasks: deciding what attributes to measure and identifying good methods for measuring those attributes. We look at how portals can understand impact and in particular develop a suite of micro-economic metrics, along with low cost ways to measure them, plus provide guidance for developing further microeconomic indicators. Microeconomic indicators are often overlooked or given in insufficient detail, such as 'sales'. However, the previous work showed this is both a gap and also a relatively accessible and comparatively inexpensive field.

Indicators and Metrics for Assessing the Economic Impact of Portals - Full Reports	
1.	Measuring Use and Impact of Portals
2.	Developing Microeconomic Indicators Through Open Data Reuse

[5] datasetsearch.research.google.com

Section 2 (Automated Assessment of Indicators and Metrics) offers a way to identify and automatically assess metrics that indicate reuse. It does this by reframing the problem. Instead of trying to reduce the cost of existing methods such as surveys and case studies, or attempt to build on existing technologies in portals, it posits a third approach - that of assessing reusability within a repository, and using this as a proxy to automatically assess reuse.

Automated Assessment of Indicators and Metrics - Full Report	
3.	Dataset Reuse: A Method for Transforming Principles into Practice

Section 3 (Assessment of Funding Options for (Open) Data Portal Infrastructures) presents a toolkit that explores the various funding models (private/public/hybrid/self) that may be possible for portals and the various activities that might be required in order to provide value for these funding streams. Some ideas are posited in the previous report and this work builds on this and creates an actionable guide for portals to develop business cases for their data release and funding models.

Assessment of Funding Options for (Open) Data Portal Infrastructures - Full Report	
4.	Funding Portals: A Business Case Approach to Funding Model Longevity

Section 4 (Exploring Alternatives to Portals) provides both a system which portal owners can use to assess their own portal's robustness and a prototype of a possible alternative which improves on the areas in which most portals have been found to be weaker. It builds on the report 'The Future of Open Data Portals' which presents ten ways in which Open Data portals must evolve for sustainability and added value. These include publishing good quality metadata, and borrowing principles from ecommerce to organise for use. It operationalises these 10 aspects, by providing ways to assess to what extent portals are managing to achieve these goals. The co-location of tools is of particular import; by developing the correct tools to create communities around datasets, such communities can then contribute the necessary added value to the data, in a virtuous circle.

Exploring Alternatives to Portals - Full Reports	
5.	Open Data Portal Assessment Using User-Oriented Metrics
6.	A Distributed Version Control Approach to Creating Portals for Reuse

1.5 Disambiguation

This report uses the terms 'user-centric metrics' and 'reuse metrics'.

'User-centric metrics' describes our approach to creating metrics that measure a whole range of aspects of data. Many metrics are available to measure data. The most easily accessible are largely derived "top down", in that they assess properties of the portal or data. These provide a valuable perspective and are relatively easy to implement, however, these do not take into account the concerns of users, and thus are only are weakly linked to the impact of open data. User-centric metrics are derived from "bottom-up" methods for measuring the value of open data that are grounded in what users need from data to perform core functions. This method produces metrics that are therefore more directly related to the impact of the data[6].

A **reuse metric**, on the other hand, is a metric that specifically measures data reuse and reusability. These should still be derived in a user-centric manner.

[6] Frank, Walker and Thompson, 2015

2. Indicators and Metrics for Assessing the Economic Impact of Portals

2.1 Measuring Use and Impact of Portals

There are many motivations for measuring open data. It is needed to maintain quality of data and support; to justify investment; to focus resources to most effect; to compare progress between countries, institutions and portals and to set benchmarks for countries, institutions and portals. However, it is still difficult to know exactly what to measure and how to measure it. Once a metric is decided, it becomes the focus of both effort and observation. This can detract from other important aspects of assessment, and can also result in 'target chasing' - investing time and resources to affect a specific suite of metrics to the exclusion of others.

Even before any technical issues of measurement can be addressed, there are broader issues that affect any attempt to define a set of metrics, even in areas that appear to draw consensus. For instance, 'quality' is often used as a metric for data. However, this term can mean very different things for different types of data and different types of users. Is a good quality data set one that has all fields completed, or that perhaps has fewer fields but is more accurate? Is a good quality data one that has been cleaned, or one still contains the raw data, including outliers?

Secondly, there is often misalignment when deciding who 'the users' are that any metric should attempt to measure. Is it the primary (those who use data directly) secondary (those who use it through an intermediary) or tertiary users (those that use the product of the data) that need to be assessed? What particular aspect of reuse activity should or could be measured - downloading, integrating, creation of application or use - and how can this be addressed when different users perform different functions? Further there is not clear track between a portal and use - even if data appears in an app, it may have actually been collected from one of many sources - the original data, a copy, or via a catalogue.

2.1.1 Methodology

In this section, we conducted extensive desk research around the construction of metrics and indicators as well as reviewing existing and emergent methods for capturing these. In the previous report the Common Assessment Framework (CAF) was proposed as a basis for developing metrics. We held a workshop with national portal representatives in which we:

Presented the CAF to see if it effectively disambiguated Use and Impact in practice; Used the Impact section to guide participants in developing their own metrics and indicators for impact, based on the outcomes they desired; and developed metrics and indicators based on the showcase and use case corpuses.

Context	Data	Use	Impact
Legal	Licensing – how open Technical – format, APIs, documentation What data – core data, sectors represented Quality – up to date, complete	Type of users – researchers, entrepreneurs Purpose – reduce spending, ease congestion Activities – benchmarking, mapping	Environmental - reduced pollution
Organisational			Economic - increased jobs, growth
Political			Political - reduced corruption, better services
Legal			Social - greater equality, participation
Social			
Economic			

Figure 1: The Common Assessment Framework (selected aspects)

2.1.2 Results

The CAF gives clear guidelines for defining what should be measured. Further, the structure of the CAF is particularly useful for isolating and disambiguating ‘Use’ and ‘Impact’, as can be seen below. It is based around a loose ‘PESTLE’ format - a framework for ensuring that Political, Economic, Social, Technical, Legal and Environmental issues are addressed (see the boxes ‘Context’ and ‘Impact’. This provides a second way of ensuring that all users of the CAF are approaching the assessment in the same manner and increasing comparability across initiatives.

When measuring use and impact, the range of possible subjects to measure is vast. This leads to confusion over what needs to be measured. The impact indicators suggested at the workshop were highly diverse, including participation rate at elections, economic growth, improvement of reliability/efficacy/speed of public services, reduction in the number of homeless and increase in engagement in participatory budgeting.

2.1.2.1 Explore Showcase and Use Case

There are now a number of substantial corpuses of use cases and showcases. There are over 550 use cases on the European Data Portal, 1793 on data.gouv.fr and 232 on datos.gob.es. ODImpact.org is a site devoted to such use cases. While these were originally intended to understand how impact might be derived, the range of uses, and of course, to inspire, they are increasingly available in numbers that can be analysed quantitatively to create metrics. These might include:

- Number and quantity of data themes
- Number of types of reuses
- Log files
- Inferring user needs from quantifying areas of interest –If they need it, it should be impactful

- Measuring impact of hacks from apps developed
- Speed of addition to portal (as a proxy for rates of reuse)

Exploring individual applications at a more micro level can also help to measure impact. For example, a Journey Planner App for bikes might be able to demonstrate impact via the size of the installed and user bases. Obviously, these indicators have to be obtained, which can be facilitated by publishers and reusers working together. One suggestion to enable this was requesting that showcased app developers commit to report on a set of indicators in exchange for promotion of the app.

2.1.2.2 Develop Microeconomic Indicators

Microeconomic studies have similar econometric methodologies to macroeconomic studies but focus on specific publishers and data. An example is the report 'Assessing the Value of TfL's Open Data and Digital Partnership'. This identified direct benefits, realised in the form of revenues from market transactions and indirect benefits of positive externalities, for example, increased engagement with municipality and services. These are very reliable and comparative metrics.

While expensive, some of the cost of running such surveys can be defrayed by incorporating portal surveys into larger surveys run by business associations, such as that run by ASEDIE, the Spanish Multisectorial Information Association. They conduct longitudinal analyses of the information and data market, and the outputs of trade and governmental economic assessments can be useful to local assessments of the impact of open data in a local marketplace.

However, the key challenge with microeconomic metrics is that they usually are seen as only applicable to private reusers, and are generally limited to metrics such as sales, turnover, profit, jobs, and so on. However, there is a vast range of potential indicators out there to be developed. Below are the metrics used by ASEDIE in the previously mentioned study.

Metrics deployed by ASEDIE	
Subsectors of infomediary companies	Technical consulting, culture, directory service, economic and financial, publishing, market research, meteorological, geographic information, infomediation technology, tourism
Turnover	average, total, by subsector
Employee	total, by subsector, average turnover per employee, average expenditure per employee, average wage per employee
Share capital analysis	total, by subsector, average social capital
Profit and Loss	total, by subsector
Analysis of commercial risk	total, by subsector
Long term companies	sales evolution, employee evolution
Delisting	by motive (e.g. closure), community, subsector

Table 1 ASIDIE Metrics

Given this gap between availability of metrics and the potential usefulness, in the following report we develop a method of identifying and creating microeconomic indicators for public sector projects.

2.1.2.3 Focus on Primary Users

The keyway to facilitate measurement of use is to focus on primary users. This has not always been the case, especially with arm's length measurements such as macroeconomic surveys. This can be done by engaging more closely with users. Developing an increased relationship with the community implies a two-way dialogue that will ultimately be beneficial. In this way, the impact can be crowd-sourced in a number of ways.

Amongst current methods, the user survey is an attractive but hugely underutilised compromise for measuring user type and activity amongst primary users. These are aimed at the users of the open data around a portal and are intended to understand how open data is being used and thus assess its impact. They typically gather information using surveys and publicly available sources of data such as company registers. They provide an important different perspective from economic studies but can also suffer from being expensive to run and therefore hard to repeat. In addition, because they rely so much on self-reporting surveys, there are often concerns over whether they have obtained a reasonable representation of all users of the data. However, where they are used, they are some of the most impactful and useful studies.

A (comparatively) simple example is offered by the Irish national portal, which runs a continuous on-portal user survey offered to all users who download a dataset, which allows both engagement and measurement. In the future, new automated methods and social media analysis may avoid the need for compromise in some contexts, but this is currently some way off.

Whatever methods are being used, there are clear advantages to adopting consistent standards across the open data world and making those standards consistent with other relevant measurement programmes where practical. This increases efficiency by facilitating reuse and greatly enhances transferability and comparability of all methods. There is also significant potential in examining how methods can be usefully combined. For example:

- A microeconomic survey can be used to calibrate an automated method and thus increase the ongoing validity of the automated method.
- A user survey may be an efficient way of determining which user types should be the subject of a microeconomic study

2.1.2.4 Develop Site Analytics for Reuse

As open data is published online this has allowed the utilisation of site and related analytics for the measurement of some open data activity. For instance:

- **Page analytics.** These are similar to other types of web site, which record metrics such as which pages are accessed most often and the order in which they accessed.
- **Downloads**. Which datasets are downloaded and how often?
- **API metrics.** Where the portal enables users to access data through an API it is possible to record how often the API is used and some data about who is using it.

These automated assessment metrics are the longest standing portal level indicators, but they are considered limited in their application for assessing use. However, our research has demonstrated that, when combined algorithmically, they can be used to develop an accurate proxy for reuse. In Section 2 we describe a process for doing this and automatically assessing it.

2.1.2.5 Integrate Data Reusers and the Public into Portals

Currently, with one or two notable exceptions, users are not specifically encouraged to engage with data portals in a meaningful way. In order to more effectively track use, it is key to develop portals in the direction of more collaborative environments where the user is encouraged to engage with the portal (via other users) rather than extract the data and leave. Such an environment can be found in other data communities, such as those that use version control (VC), for instance, GitHub. Increased onsite activity would also mean the effort of finding links and improving data quality would be shared with data consumers, distributing the effort required to maintain and improve data quality among those benefiting from the data. As a side effect of using such technologies, data publishers would have access to more granular data on how their data is used, which in turn would allow them to identify high value datasets and ascertain the impact of open data. A further benefit of using VC is this community has already begun to consider the challenges around IoT data, for instance, how to manage extremely large files, such as 240 million rows of parking sensor data, and managing data aggregation.

2.1.3 Who Should Use This and How

Portal owners can action these insights to develop metrics and indicators that appropriately measure use and/or impact as desired, without conflating the two.

Ecosystem organisers, whether from the business or local government communities, can use these insights to engage with the right partners to enable reach to users who may not be directly engaging with portals.

Finally, these conclusions and directions for research should be of interest to the wider measurement community. This includes researchers, open data activists and funders, for example Luminate.

2.1.4 Lessons and Best Practices

Business groups should be encouraged to survey their members for open data use and impact.

Publishers should aim to engage with reusers identified in show cases/use cases to develop quantifiable indicators

Publishers and portal managers should share lists of metrics they have identified, in order to encourage larger catalogues of metrics

2.2 Developing Microeconomic Indicators Through Reuse

As noted above, microeconomic assessments are still underexploited as measurement tools for open data. Measuring the economic impact of open data falls into the classic problem set that plagues evaluation of any complex project; identifying the chain of causal links; devising appropriate tools and methodology for measurement, and having sufficient resources to carry out the evaluation.

The reuse of open government data creates value for the public and private sector in delivering services and insights, where the challenges of tracking value are in applying the appropriate measures to data re-use. Microeconomic indicators have been developed to track value creation and impact from the perspective of both government and of the private sector. Ongoing programmes like the annual Infomediary Sector Report produced by ASEDIE[7] (in its sixth iteration), track the progress over time of companies which have a data-based business model, using quantitative business metrics to evidence economic value. These indicators, such as sales and jobs, are generally familiar to people. Yet as well as these benefits for private companies, the public sector and citizens can also benefit, but in less obviously measurable ways. These might be efficiency and productivity gains; potential time and cost savings from innovative services running on open data, as well as improved public services

[7] ASIDIE, 2018

both in levels of service and reduction in costs to the taxpayer.[8] It is therefore vital to identify microeconomic indicators for these.

2.2.1 Methodology

To develop appropriate indicators, four cities were analysed that were creating solutions to public sector challenges with open data. They did this through open innovation projects: publishing their data openly and then co-creating public sector solutions with external SMEs. The aim was to understand what impact they were hoping to achieve and how they were assessing this. By following the open data innovations, a transferable method for identifying and measuring the impact of individual open data projects was developed.

2.2.2 Results

The first stage was to outline the process of development from the definition of the problem, to measurement of the effectiveness of the solution. This process is replicable and allows for creation of other metrics using this process.

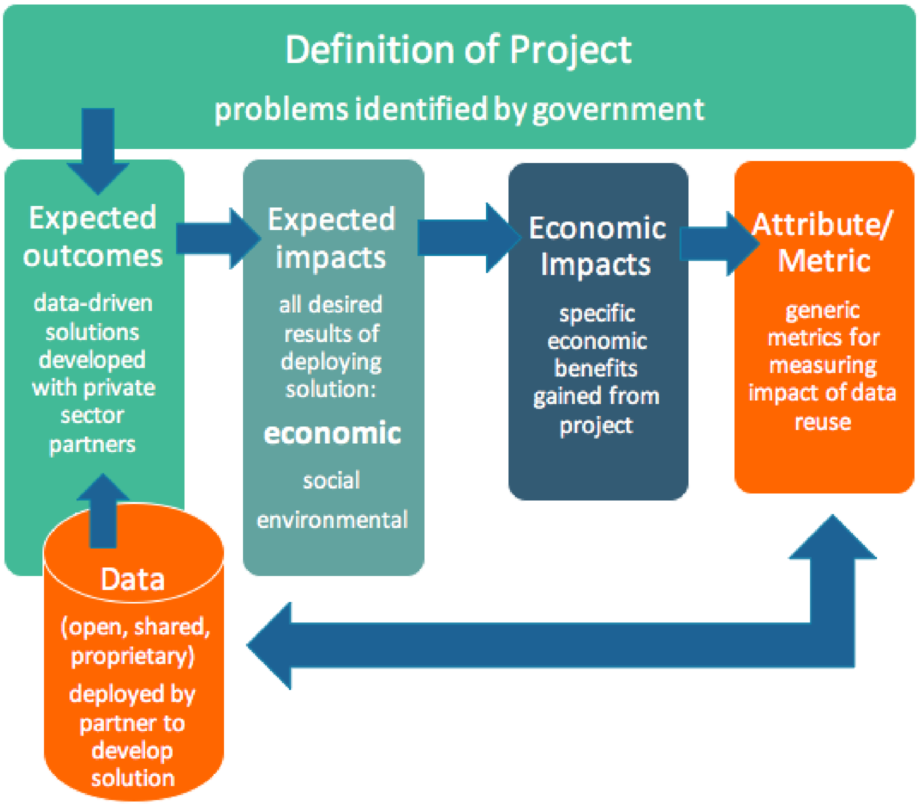

Definition of Project
problems identified by government
Expected outcomes
data-driven solutions developed with private sector partners
Expected impacts
all desired results of deploying solution:
economic
social
environmental
Economic Impacts
specific economic benefits gained from project
Attribute/ Metric
generic metrics for measuring impact of data reuse
Data
(open, shared, proprietary)
deployed by partner to develop solution

Figure 2: Framework for Identifying and Measuring Impact in Open Data Projects

From the projects that were being undertaken, we co-developed 23 metrics for impact the table below shows the economic impact, a metric that can be used to measure this, and possible sources for those metrics. The metrics are categorised according to how accessible they are to the portal owners; in-house data, related public services data and private sector data. None are inaccessible.

[8] Koski, 2015; OECD Digital Government Studies, 2018

2.2.2.1 Operations Data (In-house Data)

Operations data is held in-house by the cities as it is derived from their own management and financial records. This data is generally available across organisations and domains as it is routinely collected for the purposes of financial and budget reporting, and for monitoring and delivering services. Surfacing this data via a dashboard would have a range of benefits internally for the organisation, and allow analysis and visualisation of data for a variety of purposes, including specific project evaluations. Aggregated data could also be surfaced for a citizen-facing dashboard to enable transparency in budgets and spending, and business intelligence insights.

2.2.2.2 Related Public Services Data (Held by Other Public Services)

Drawing on related public services data presents more of a resource challenge for evaluation purposes as it requires more time, effort and what could be complex interpretation and analysis of data sources, and building models for inferring impacts of open data reuse projects. This is a barrier to a data owner engaging in more complex project evaluation due to restrictions on resources.

2.2.2.3 Private Sector Data (Financial and Business Data)

While public authorities which are co-creating solutions with individual organisations can easily access data impact on that particular organisation, it is harder for them to access wider data. City governments themselves are unlikely to engage in local assessments of the data marketplace and SMEs, but this kind of analysis is of interest at a regional and national level, and as previously noted, could be collected through partnerships with business associations.

Economic Impact	Attribute/Metric	Project Evaluation Did the data-driven solution work? How can we tell?
Operations Data - collected in-house by government		
Labour costs/ productivity	Wage costs No. of hours worked	1. Management data from service delivery teams to track impact on staff hours 2. Workflow efficiencies (e.g. smart routing)
Service delivery	Service level outcomes Service delivery costs	1. Management reporting data 2. IoT data – e.g. smart bins 3. Citizen complaint levels – data gathered from citizen app, website, email and telephone to City re: services 4. Citizen reporting - data gathered from citizen feedback app, civic website, email and telephone re: conditions (e.g. weather, road conditions, cleanliness)

Resources	Mileage Fuel Consumption Grit/salt consumption Water consumption	1. Management data from service delivery teams to track impact on vehicle use (mileage, hours of running, repairs) 2. Purchase records - fuel, salt, grit 3. Water meter readings
Procurement	Contract pricing	1. Management data – service levels and costs
Traffic congestion	Traffic monitoring Journey times	1. CCTV and traffic light data 2. Travel app data 3. Citizen reporting
Related Public Services Data		
Road traffic accidents	Frequency of accidents Emergency call outs Injury statistics	1. Road traffic accidents statistics 2. Emergency services records 3. Hospital admissions/treatment records 4. Insurance claims
Health	Health statistics Air quality Exercise levels	1. Hospital admission/medical treatment records for respiratory disease, asthma, cardio-vascular disease, children's fitness/obesity 2. Air quality monitoring data from national data collection and local monitoring 3. Travel app statistics (journey length, route, frequency)
Private Sector Data		
Data services marketplace	SME no. SME turnover SME profit/loss SME sustainability SME employment Data Products/Services	1. Industry surveys (data services sector) 2. Financial reporting data 3. Investment data 4. Sector employment figures 5. Market survey data (products and services)

Table 2 Microeconomic Indicators and Metrics

2.3 Who Should Use This and How

Public sector portal owners should be encouraged to collect and publish these metrics; not only can public sector portal owners increase public spending and efficiency transparency for their citizens, but as more data is published it can be used cumulatively to assess impact over a broader area and timescale.

Policy makers should encourage the creation and sharing of these metrics, possibly by making it a requisite of funding of such data reuse projects.

2.4 Lessons and Best Practice

The Impact>Attribute>Evaluation method can be applied to any public sector and/or smart city open data project. By applying it to other re-use projects the list of possible microeconomic indicators can be extended. These should be published and shared.

Building the assessment of these metrics into the original project plan reduces the cost of collecting them.

3. Automated Assessment of Open Data Use

An increasing amount of data is published openly on the web, ideally with the aim of reuse. One of the key challenges to its' uptake is supporting formats and capabilities to make it useful in as many contexts as possible (Shadbolt 2012). Reuse is more common in some domains than in others: Scientists reuse data of their peers to repeat previous experiments, propose new solutions, and derive fresh insights. Data is recognised as an asset in itself, cited and archived just like scientific literature. Developers define benchmarks and gold standards that everyone can use to establish to compare related approaches. They reuse such datasets to ensure that approaches remain comparable. Supervised machine learning, one of the most successful types of AI is dependent on the availability of relevant datasets to train algorithms. In this case, reuse is an economic necessity –deep learning architectures need to be pre-trained on large amounts of data and generating new datasets is too costly for most machine learning applications.

Reusability is stated as one of the four FAIR principles, a compilation of high-level best practices for making data findable, accessible, interoperable, and reusable. The "R" in FAIR gives guidelines on reusability include the following points, all focusing on metadata: (i) meta(data) are richly described with a plurality of accurate and relevant attributes, (ii) (meta)data are released with a clear and accessible data usage license, (iii) (meta)data are associated with detailed provenance, (iv) (meta)data meet domain-relevant community standards. The EDP in itself can be understood as a tool to improve the FAIRness of the over 1 million open government datasets it harvests.

While the FAIR metrics provides exemplary metrics for the FAIR principles, measuring FAIRness is not an established practice. There are also a variety of best practices and guidelines detailing data sharing and reuse principles, including the W3C best practices for data on the web or SharePSI or metadata standards for different purposes: general purpose standards such as Dublin Core2 or DCAT3, focusing on specific elements such as provenance (PROV4) or data quality5 as well as domain specific extensions or standards.

Despite these efforts, portal owners and data publishers do not measure reuse routinely. Existing guidelines, indicators and metrics cannot be trivially mapped to observable features in the technical architecture of the publishing platform, which could be tracked and assessed automatically. Previous work [citeEDP1report] has suggested several solutions, including pixel tracking, dataset citations, and enforcing log-ins. These solutions have important limitations:

· Pixel tracking, and similar methods, operate at a granular level, and findings depend on the front-end design of the platform rather than on how useful the dataset is. More importantly, translating pixel-tracking insights into principles and practices to make datasets more reusable is hard, as the former is too low-level for the latter.

· Dataset citations, while an excellent idea, is not widespread outside scientific communities. While an incentives system for data citations is emerging in this space, it is unclear how it would transfer to open government data.

· The most used public sector datasets (such as urban transportation) often have excellent ecosystems that enable them to track usage in a less automated fashion (such as surveys, or app galleries). While the intense usage of their datasets, and the value of learning more about what features are most beneficial justify the cost of managing this tracking, this does not transfer to datasets that are less popular, as these cannot draw from a community of users for feedback. In the same time, the holders of these high-value data assets may not have the incentives to explore new tracking methods that would benefit other types of datasets.

· Finally, very few portals imply publish their own data – most provide a platform for data from a variety of sources, and some, such as the European Data Portal, are catalogues of datasets. Therefore, most portals are not in a position to implement tracking features such as log-ins.

Therefore, it is vital to address an alternative assessment approach, which focuses more on the reuse side of open data than the publishing, with automation support. This report presents such an approach. We introduce a method that helps a portal owner understand what makes a dataset more or less reusable, using engagement data they can track themselves. To apply the method, the portal needs to capture a minimum of engagement metrics, map higher-level dataset reuse indicators to such metrics and identify a subset that co-relate with reuse.

Automated assessment of reuse remains a substantial challenge. In an ideal world, a more end-to-end tracking of portal activities throughout the process would enable this. However, this requires new underlying structures, and while these may well be necessary eventually to ensure the sustainability of portals, the description of this goes beyond the remit of this report, which describes what can be achieved with the current technology, or with minimal adjustments. For these reasons, we have validated the method in a scenario which captures data about how people engage with datasets, for which such engagement data is easily available. We provide recommendations for portal owners to augment their publishing and portal design practice to support and enhance those features of a dataset that are quantifiably linked to higher engagement from users.

3.1 Methodology

The method consists of the following steps, to be carried out by teams managing open data portals:

1. Scope the assessment exercise, for instance by deciding the specific collection of datasets that will be considered.

2. Define reuse metrics. These depend on the capabilities of your portal and the underlying technical infrastructure. If you cannot define direct metrics, think about proxy metrics. Run a study to validate them with observable reuse indicators for datasets published on Github

3. Collect reuse metrics (or proxies). For this, you need technical capabilities which may be built into the publishing software you're using, or aggregated metrics derived from lower-level system logs.

4. Define reuse indicators. These need to be measurable and will be used as features in the prediction model. Below is an example list of observable metrics for datasets in Github. (The full report provides a list which can be used as a starting point, based on a comprehensive literature review.)

Category of Feature	Feature
Portal	Size of repository
	Number of all data files
	Licence
	Dominant data filetype (number of csv/etc)
	Description
	Ratio of open to closed issues
	Ratio of data files to all files in a repository
	Problematic files with respect to a particular library
Documentation/Metadata	Length of the documentation
	Unique URLs
	Language of the documentation
	Number of coding blocks (i.e. both inline and highlighting blocks)

	Number of images
	Broken URIS
Data Files	Number of rows and columns of each individual data file
	Missing values
	Data Type of HEADERS (i.e. check if headers are strings)
	Size of each data file
	Aggregated size of all the data files in the repository

Table 3 Possible Reuse Indicators

5. Analyse their distribution for the top-reused group of datasets.

6. Use a combination of those features to build a statistical model to predict reusability.

7. Derive recommendations to datasets and publishing processes.

3.2 Results

Based on these an approach was developed to predict the likelihood of whether a dataset from the platform will be reused. This was combined in a predictive model to estimate a data repository's reusability, based on the documentation and structure of the repository. While this was done on the example of GitHub, this methodology could, in theory, be applied to any engagement proxies relevant to a specific data portal.

The model uses features from all three layers (repository and description as well as the data file) to learn what makes a dataset reusable in this particular context. For our GitHub analysis, the repository features were found to be most predictive. The approach categorises a dataset repository into 1 out of 4 potential groups of reuse likelihood: Very likely to be reused; likely to be reused; moderately likely to be reused and unlikely to be reused.

Looking at a statistical analysis of those repositories that are very likely to be reused showed a number of interesting results. For instance, the textual description of the data repository was longer, the repositories have a lower number of problematic files (meaning they can be opened with standard configurations), and the age of the repository does not correlate much with its reuse status. There was also more "traffic" around the datasets visible, in terms of community engagement through opening and closing issues on the platform that notify others.

This work demonstrates the tension between calls for data reuse principles and actionable metrics and automated approaches facilitating data publishers and tools designers to implement functionalities supporting dataset reuse in an open collaborative environment. The findings point to a number of under-explored opportunities to encourage and facilitate dataset reuse on the web.

3.3 Who Should Use This and How

Even with current technologies, this approach can be used to inform system designers building functionalities to capture this information automatically; publishers in supplying certain information as metadata, and user experience designers, to inform the design of the interaction process between datasets reusers and the interface of a data portal. Portal owners can use this to inform their portal development, and open data users in the wider ecosystem can use these insights to help them identify the data sets that may be most useful to work with.

3.4 Lessons and Best Practice

Current portals that have functionalities to measure engagement and user interaction can develop bottom-up reuse indicators targeted to the user group of the platform, based on their real interactions with the datasets.

This work could be built on by integrating functionalities that measure engagement with datasets in an automated way. Portals could support the automatic assessment of a dataset at the time of publication and recommend features that would increase reuse probability according to the proposed model. This would allow to increase a datasets reusability before publication, focusing on not just the data itself but also on documentation and other potentially relevant features of a project.

4. Assessment of Funding Options for (Open) Data Portal Infrastructures

In the vast majority of cases, portals and related activities are funded by government departments. Nationally this has largely been based on transparency budgets or municipal IT departments. Few portals are financially sustainable, and some have no basis for becoming financially sustainable either.[9]

The 2018 Open Data Maturity report found that the cost of actually running portals is subsumed into wider open data strategy funding, and no national governments were identifying the cost of sustaining an open data portal as its own activity.[10] Further, no alternative funding models had been explored by portal owners.

As Member States are required to publish certain data, and it is therefore understandable as a regulatory cost, this somewhat explains this cross-state hesitation to explore funding from other angles. However, this 'compliance' approach obscures the possibility of understanding the funding of open data from a more sustainable point of view, which can be developed using a business case.

[9] Barbero et al. (2018)

[10] https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n4_2018.pdf

4.1 Methodology

The report, Ensuring the Economic Sustainability of Open Data Portals: Understanding Impact and Financing (2018), made wide recommendations over all areas of financing, from selecting a platform to the cost of training to various mechanisms that might be used for freemium model. Our approach has therefore been to operationalise the insights and recommendations in a structured manner that assists portals in making decisions, and conduct additional research that informed this.

Our research involved secondary desk research, a workshop with cities and regions who are in the process of opening up their data, and the development of mini-case studies to inform insights between the rationale of a portal and its potential funding strategy. Analysis was carried out on the 2018 reported survey answers to inform the development of a budgeting template with 20 costs considerations.

4.2 Results

This report identifies three types of business cases for data portals, then explores 4 funding models which can be used with these business cases. Finally, it covers 20 cost activities that should be included in the budget for portals. We also provide a budgeting template and case studies on the financing models.

4.2.1 Three Business Cases for Data Portals

Business Case	Project	Funding and Publishers
Direct Budget Savings	Helsinki Region Infoshare (Regional)	Initial Funding: SITRA, the Finnish Innovation Fund; Finnish Ministry of Finance municipality cooperation grant Current Funding: Cities of Helsinki, Vantaa, Espoo and Kauniainenc Publishers: Multiple departments across the cities
Citizen Participation	Data Mill North (Mixed public and private publishers)	Initial Funding; Cabinet Office Release of Data Fund Current Funding: Repository partners Publishers: 63 data owners and publishers across the north of England
Innovation in Products and Services	SCORE/SCIFI (City and regional)	Initial Funding: Interreg 2Seas programme, internal IT budgets, private companies

		Current Funding: N/A (still in initial phases) Publishers: Amsterdam, Aarhus, Aberdeen, Bergen, Bradford, Dordrecht, Ghent, Gothenburg & Hamburg, Delft, Mechelen, Bruges, West Flanders, Saint Quentin

Table 4 Business Cases for Data Portals

Direct Budget Savings

According to the City of Helsinki,[11] a partner in Helsinki Region Infoshare, opening up city purchasing data has resulted in budget savings of 1-2 percent. This 'total transparency' has engaged new audiences with the city administration and encouraged civil servants to ensure their procurement is fully fair and obtains the best value. Additionally, releasing and using open data via open APIs has saved time and staff effort.

Consequently, the annual cost of providing the service is relatively low compared to the benefits received, especially when secondary benefits such as increasing trust or providing or enabling better services for citizens are factored in. While the initial pilot stage, which lasted 2 and a half years, cost around 1 million euros, the annual cost is 60.000 euros, split across the 4 partners.

Citizen Participation

Data Mill North[12], a collaborative website originally set up by Leeds City Council, began life as Leeds Data Mill, which tried to bridge the gap between decreasing resources and increasing demand for public services. The aim was to enable citizens and organisations to become digital social entrepreneurs who were aware of the relationships between the city's services and businesses. This required open data from multiple sources to be combined in one site.

This led to a naturally collaborative approach. The site grew larger and extended to include nearby Bradford. As the pooled data grew, so did the idea of pooling other resources including funding. Eventually, the site was extended to include data from the entire north of England, from not only cities but government departments, charities, other public sector organisations, schools and private companies.

Innovation in Products and Services

The Smart Cities Open Data Reuse (SCORE) and the Smart Cities Innovation Framework Implementation (SCIFI)[13] projects used public-private innovation processes to create new services with open data. Data in SCIFI is published on the project hub (FIWARE) to enable cities without existing portals to participate in the innovation.

These business cases identified the following information:

- a brief description of the problem;

[11] https://hri.fi › en_gb

[12] datamillnorth.org

[13] smartcitiesinnov.eu

- KPIs that would be used to assess if the problem was solved;
- the root causes of the problem; who was affected;
- what was the scale of the problem;
- who the ‘problem owner’ was;
- who the political sponsor was;
- the stakeholders;
- who had been consulted about this;
- the link to the relevant part of the policy plan;
- the resources that could be committed.

In any business case there is a need to define a ‘do nothing’ scenario, to assess the comparative value of not investing. In these business cases, the leads were challenged to find other existing technical solutions, i.e., to see if the problem could be solved without actually opening data.

4.2.2 Four Financing Models for Portals

The business models can be used with a variety of financing models. However, not all financing strategies can be used with all business cases, and there are limitations of each, which are outlined in the Key Issues column below. Details of full implementations of each of these models can be found in the full report.

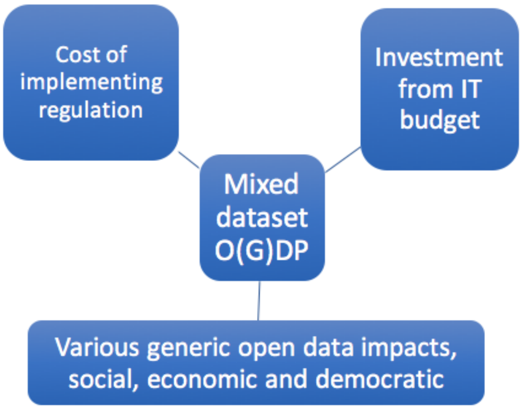
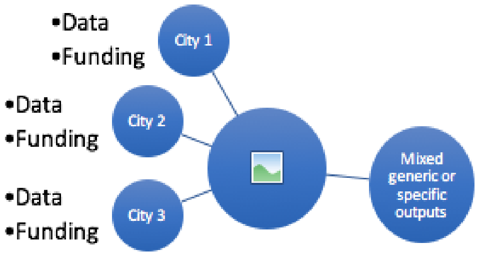
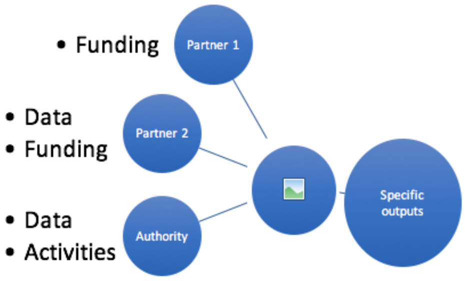
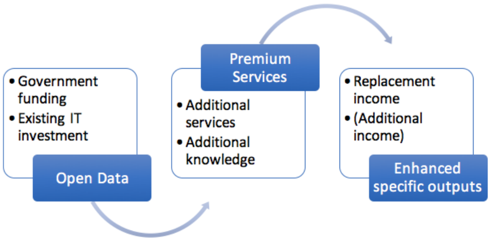
Funding	Model	Key Issues
Internal (public) financing	 Cost of implementing regulation Investment from IT budget Mixed dataset O(G)DP Various generic open data impacts, social, economic and democratic	- Often budgeted as a regulatory cost - Future funding needs are rarely planned for - Rarely manage to generate revenue
Co-funding	 •Data •Funding City 1 •Data •Funding City 2 •Data •Funding City 3 Mixed generic or specific outputs	- Sharing costs between data owners - Good solution for cities in close proximity to others - Secondary benefits include data pooling as well as reduced costs
External funding	 • Funding Partner 1 • Data • Funding Partner 2 • Data • Activities Authority Specific outputs	- Combines public and private sector data - Often driven by smart city initiatives - Can be part of larger public-private partnership
Self-financing	 • Government funding • Existing IT investment Open Data Premium Services • Additional services • Additional knowledge • Replacement income • (Additional income) Enhanced specific outputs	- Charge for either data, services, or tools using a freemium model - Requires data quality to be sufficient, and users who can both use the data and afford to pay for it - Highly skilled employees needed to deliver enhanced datasets

Table 5 Funding Models for Portals

4.2.3 Twenty Cost Considerations for Portals

Analysis of the previous sustainability report identified 20 cost areas which must be addressed, whichever business case or funding model is used. From this we created a template for developing a budget that covers all of the areas.

Activity	Cost	Details
Build	Development	This is generally seen as the largest cost, however, with a wide variety of catalogues and platforms available, the cost of development is reducing. A major decision is whether to develop (and then maintain and improve) in house or to contract out
	Infrastructure (incl. hosting)	This has a number of dependencies: is the portal a catalogue or will it host, totally or partially, the data sets? How will publishers and users access the datasets and how frequently? Answers to these and other technical questions will impact on the cost of the infrastructure, which might be minimal if only a few datasets are hosted, but extensive in the case of a large national portal
	Design and user experience	Again, this can vary extensively depending on whether the portal owner chooses to innovate or simply reuse an existing format
Strategy	Short, medium- and long-term goals	Setting aside budget to cover time for the setting of short, medium- and long-term goals, which often require the input of a number of stakeholders
	Prioritisation	Identifying time and resources for the development of business cases and associated funding plans

Activity	Cost	Details
Operations	Portal operations (incl. user engagement)	Portal operations include all the day to day activities that might include content management and social media, reaching out to users regarding updated data sets and liaising with data publishers
	Data provision	Identifying, locating, cleaning/redacting and preparing data for publication. This is a large part of ongoing budget spend. Specialist support with aspects such as metadata may be required, which should be reflected in the staffing.
	Staffing	This is likely to change with changing priorities and value-added services. This is the area portals frequently underestimate, both in ongoing requirements and hidden costs
	Outreach, training and support for publishers	The percentage of the budget that should be allocated to this will vary with the nature of the portal. For a portal focused on a data intensive area such as national mapping agencies, publishers are likely to already be highly skilled. For a national portal publishing data from multiple departments, this may require considerable investment
	Data analytics	If an external platform provider is being used this cost may be rolled up with the design and hosting
	Maintenance and improvements	For subsequent years. This element of the budget should not be reduced too much as it will limit the ability to respond to user need
Encouraging Use	Outreach, training and support for users	This is an important element to ensure take up. As above, it is often left to separate budgets, but for a consistent approach, should be included in the main budget

Activity	Cost	Details
	Incentivising use	While this may not be necessary in every budget, it is particularly important where portals are ‘eating their own dog food’, i.e. publishing data openly as an effective way to share it between departments or sub-departments
	Value-added services	Value-added services may include co-locating tools, improving documentation or enhancing metadata. They may also include more complex services that can be charged for
Measurement	Monitoring of use and impact	Ongoing assessment and monitoring should be implemented where possible
	Measurement of use and impact (research)	A small percentage of the budget should be reserved for an annual survey or other mechanism to understand how the site is being used and what impact this is having on the larger ecosystem
	Recording and management of activities, measurements and monitoring	Depending on the funding stream, this can potentially be a relatively onerous cost area. If reporting back to a central grant making body or project overseer is required it is important to apply sufficient resources to this task. Even where this is not required, ensuring that activities and impact are documented properly is an important part of sustainability
Income	Revenue	Provision of freemium services/sponsorship (if applicable)

Table 6 Cost Considerations for Portals

4.3 Who Should Use This and How

Portal owners at all levels from the national down to specialist areas can utilise these methods as a toolkit to focus and direct financing strategy and operations. By first building well thought-out business cases for opening data with colleagues, the public, business and academia, they can then begin to select a sustainable funding strategy. For instance, co-funding as a region makes sense if a city is focused on budget savings, but external funding would be less appropriate. Similarly, a region may find the provision of freemium services to be less successful than a single-issue portal, such as one focused on high value datasets such as forestry.

Portal owners can also use the business case approach to understand what data might be valuably utilised that should not necessarily be opened – i.e., which can or should be shared with users under a specific user policy or rationale, such as open innovation of public sector services.

Policy makers can use this to help to assess ways in which policy can be used to support movement away from a focus on open data as an IT cost and towards integration with other governmental activities. Finally, potential funders who are interested in supporting open data can use this to identify appropriate opportunities and ways to get involved with publishers across the spectrum.

4.4 Lessons and Best Practice

Portal owners should take a broad and deep view of the full cost of the portal including all cost activities, and ensure that the full budget is surfaced, to avoid hidden costs

Freemium services should be focused only on specific data areas, where both customers and staff are familiar with purchasing and supplying services

Portals cover a range of activities, and sustainable funding may come from a variety of different sources to cover this. While the hosting may remain an internal cost, portals require data and promotion, and commercial agreements could include covering the cost of these

Business cases should be created for the portal and datasets that will be published. A clear business case for the development and continued support of a portal will not only make it more sustainable but will also establish where to look for impact

Future alignment of the open data portal beyond being an IT or transparency concern should be addressed. Consider rolling it up with another aligned service to add value. For instance, if the aim is to create business innovation via open data, identify which business support activities the open data portal can become part of.

5. Exploring Alternatives to Portals

5.1 Open Data Portal Assessment with User-Centred Metrics

As the use of open government data enters the mainstream, it is necessary to think about what is next in publishing and using data. Google dataset search has arrived, which is already changing how users discover data and threatening to make portals redundant for discovery. That, however, assumes that the only important point about data is the ability to be found by users. Centralised portals for publishing were an early and necessary step in developing the open data narrative, but ultimately, portals must be a means to facilitate use and foster accountability and innovation. For the average citizen, it is what is done with the published data that is important. For the business analyst looking for the right information for their work, the challenge is mostly around finding and making sense of the sources and deciding which ones are most relevant in a given context.

Therefore, it is vital to ask, is the current technical approach to portals fit for the future? Firstly, it is important that the most crucial dimensions of portals are identified. This has been done in the 'Future of Open Data Portals' Analytical Report. By developing relevant indicators and metrics for each of these themes, the extent to which portals are currently fulfilling their potential can be assessed.

5.1.1 Methodology

The Analytical Report 'The Future of Open Data'[14] identifies 10 ways portals can organise for sustainability or add value to their offering. These are:

- Organising for use of the datasets (rather than simply for publication);
- Learning from the techniques utilised by recently emerged commercial data marketplaces; promoting use via the sharing of knowledge, co-opting methods common in the open source software community;
- Investing in discoverability best practices, borrowing from e-commerce;
- Publishing good quality metadata, to enhance reuse;
- Adopting standards to ensure interoperability;
- Co-locating tools, so that a wider range of users and re-users can be engaged with;
- Linking datasets to enhance value;
- Being accessible by offering both options for big data and options for more manual processing. Commercial exploitation may require Application Programme Interfaces, while citizen users may prefer to download a more human readable comma separated value files, This ensures a wide range of user needs are met;
- Co-locating documentation, so that users do not need to be domain experts in order to understand the data;
- Being measurable, as a way to assess how well they are meeting users' needs.

[14] https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf

We operationalised this research by either selecting from the literature or developing metrics that would allow these dimensions to be measured. We developed metrics using relevant literature and established guidelines. Once these metrics were established, ten portals were assessed as to how well they met the 10 user-oriented sustainability principles, in order to establish the size of the gap between the current situation and the ideal. The list included EU government data portals from different stages of Open Data Maturity (trend-setters, fast-trackers, followers, beginners), as well as some other specialist open data portals:

- Cyprus National Data Portal (trend-setter)
- Avoindata.fi (fast-tracker)
- Data.gov Belgium (fast-tracker)
- Data.gov Slovakia (fast-tracker)
- Dados.gov Portugal (follower)
- Island.is (beginner)[15]

plus

- EU Open Data Portal
- London Datastore;
- Geo Data Portal Luxembourg
- Open Data Trent

5.1.2 Results

5.1.2.1 Development of Metrics for Assessing Sustainability of Portals

Dimension	Metric	Existing/Developed
Organise for Use	1. Each dataset is accompanied by a comprehensive descriptive record (going beyond a collection of structured metadata). 2. An extract of the data can be previewed (for easier sense making). 3. The portal provides recommendations for related datasets. 4. The portal enables users to review/rate the datasets. 5. Keywords from datasets are linked to other published datasets	Based on Opquast Web Data Quality Checklist http://opquast.com/en/
Co-locate Documentation	1. Supporting documentation does not exist. 2. Supporting documentation exists but as a document which has to be found separately from	Intelligibility Metric, Walker, Frank and Thompson, 2015

[15] Categorisations correct at time of research

Dimension	Metric	Existing/Developed
	the data. 3. Supporting documentation is found at the same time as the data (e.g. the link to the document is next to the link to the data in the search). 4. Supporting documentation can be immediately accessed from within the dataset but it is not context sensitive (e.g. a link to the documentation or text contained within the dataset). 5. Supporting documentation can be immediately accessed from within the dataset and it is context sensitive so that users can immediately access information about a specific item of concern (e.g. a link to a specific point in the documentation or the text contained within the dataset).	
Be Measurable	1. Portal has No analytics. 2. Portal has Site analytics. 3. Portal has Use analytics. 4. Portal has Impact analytics.	Based on review of web analytics tools
Promote Standards	1. A permanent, patterned and/or discoverable URI/URLs is used for each dataset (e.g. URI/URLSs can be used as universal, unique identifiers by appending a serial number or other internal naming system to a domain). 2. The portal uses versioning of datasets (to maintain the history of a dataset). 3. Dates are available in a standard format (facilitates the automated exploitation of date-type data and their conversion according to specific needs or constraints). 4. Metadata associated with each dataset is available in a standard format (e.g. using VOID or DCAT) to enable automated metadata retrieval and import of metadata from other data catalogues. 5. The metadata catalogue can be retrieved using a standard protocol (e.g. automatic retrieval of the metadata catalogue using RDF or HTTP GET).	Based on guidelines from W3C eGov Interest Group and OpQuast.
Promote Metadata	★ Metadata Ignorance. ★★ Scattered or Closed Metadata.	European Commission 5-level maturity schema for metadata management

Dimension	Metric	Existing/Developed
	★★★ Open Metadata for Humans. ★★★★ Open Reusable Metadata. ★★★★★ Linked Open Metadata.	
Link Data	★ **On the Web:** Make your stuff available on the Web (whatever format) under an open license. ★★ **Machine-readable data:** Make it available as structured data (e.g. Excel instead of image scan of a table). ★★★ **Non-proprietary format:** Make it available in a non-proprietary open format (e.g. CSV instead of Excel). ★★★★ **RDF standards:** Use URIs to denote things, so that people can point at your stuff. ★★★★★ **Linked RDF:** Link your data to other data to provide context.	5 Stars of Open Data, Tim Berners Lee
Promote Use	1. The portal is connected with social media to create a social distribution channel for open data. 2. The portal provides users with online support for feedback, to request/suggest the publication of new datasets, and when problems arise during use (e.g. contact form, discussion forum, FAQs, helpdesk, search tips, tutorials, demos). 3. The portal provides a way for users to keep informed of updates to the data (e.g. news feed). 4. Datasets are accompanied by links or resources that provide user guidance and support. 5. Examples of reuse (fictitious or real) are provided (e.g. information contributed by other users, last reuse, best reuse, data stories).	Based on range of literature
Be Discoverable	1. The publisher/owner of the data has an open data portal (or similar search mechanism). 2. The publisher/owner of that portal publishes an updated, searchable list of datasets.	Discoverability Metric, Walker, Frank and Thompson, 2015

Dimension	Metric	Existing/Developed
	3. The publisher/owner of that portal publishes an updated, searchable list of datasets with synonyms. 4. The publisher/owner of that portal publishes a list of datasets which are known to exist but are not currently available (limiting the time wasted on abortive searches).	
Co-locate Tools	1. The portal does not provide visualisation or collaboration tools for users to engage with the datasets. 2. The portal provides visualisation tools to enable users to engage with the datasets. 3. The portal provides visualisation and collaboration tools to enable users to participate in the governance of the portal (e.g. dataset rating) but the engagement with other users is limited or mediated by the administrator. 4. The portal provides visualisation and collaboration tools to enable users to collaborate innovatively with other users.	Based on range of literature
Be Accessible	1. The portal uses human and machine-readable and non-proprietary formats (e.g. CSV, XML, RDF-based formats). 2. The portal provides different types of formats for the same dataset. 3. The mechanisms for accessing and interacting with datasets are documented. 4. Multilingual support is available on the portal. 5. The portal supports the visually and hearing impaired.	Based on Web Content Accessibility Guidelines (WCAG), Version 2.0, from the World-Wide Web Consortium (W3C) Uses 'always, sometimes, never' scale rather than being cumulative.

Table 7 Metrics for Sustainability of Portals

5.1.2.2 Assessment of Portals with Metrics

Portal	OUse	PUse	Disc	Meta	Stan	Co-D	Link	Meta	Co-T
EU Open Data Portal	3	4	2	5	4	3	4	2	2
Dados.gov. Portugal	3	5	2	2	1	2	3	3	4
London Data Store	4	4	2	2	1	2	3	1	3
Cyprus National Portal	3	4	2	5	4	2	4	2	2
Open Dat Trento	3	5	2	4	3	3	3	1	2

Geo Data Portal Luxembourg	2	5	2	2	1	3	3	3	4
Data.gov Belgium	1	3	2	2	3	2	4	1	1
Avoindata.fi	3	5	2	4	3	2	3	2	3
Data.gov Slovakia	4	5	2	4	4	3	4	1	3
Island.is	3	3	2	4	3	2	3	1	1

Table 8 Results of Assessment

The numerical results for 9 categories are shown above. The results show substantial variation on most dimensions. This is particularly true for 'Organise for Use', 'Promote Standards' and 'Co-locate Tools' which vary from 1 star to 4 star ratings, and 'Publish Metadata' which varies from 2 to 5. None of the portals had managed to achieve the 5th star of open linked data, despite this being the most well-established open data metric.

Unsurprisingly, none of the portals achieved particularly high scores on 'Be Measurable'. However, they all scored the same on 'Be Discoverable', in that none of the portals reviewed publishes synonyms or lists of non-available datasets to reduce search overheads. There was also little variation in 'Co-locate Documentation'.

5.1.3 Who Should Use This and How

These metrics can be used by portal owners to assess the current sustainability of their portal. They can also be used as a guide to areas which can be improved, and therefore in helping focus development choices. The metrics also describe exactly what should be implemented to improve value, which reduces the search costs of portal improvement.

5.1.4 Lessons and Best Practice

Portal owners should consider a regular assessment of their portals along these dimensions using the metrics above

Areas with particularly low diversity of results, such as Be Discoverable and Co-locate Documentation, should particularly be addressed. Are there technical or social barriers preventing the implementation of improved solutions?

Recent research has looked at ways to automatically assess some of these metrics and analysed a subset of (CKAN-based) portals indexed by the EDP.[16] Among other things, the analysis showed that current technical realisations of portals do not lend themselves well to a continuous, detailed monitoring of data use, which in turn means that portal owners have limited insight into the impact of their publishing effort. Further work should be invested in this area.

[16] Dix, 2019

5.2 Distributed Version Control Approach to Creating Community Data Spaces for Reuse

Previous studies on the future of open data portals, including the work above, suggest that co-location of tools and promotion of use are two of the aspects where current portals struggle the most. Many portals, including harvesters such as the EDP, cover large numbers of datasets which are heterogeneous in terms of size, format, quality and publication environment. In addition, publishers do or cannot make any assumptions about the scenarios in which the data will be used and tailor their processes and technologies accordingly. This leads to a trade-off. In theory, by not privileging some scenarios over others, publishers maximise use; in practice, this means that releasing the data follows a one-size-fits-all approach, which often creates substantial overheads down the data value chain when data has to be transformed and curated for particular applications or skill sets.

To make portals more useful, portals need to set up a participatory ecosystem to manage this trade-off. Users are the ones with the best knowledge about how the data could and should be used - they have used tools to handle the data, and could share some of their experiences with others. They are also much better placed than publishers to promote data reuse, leveraging on their experiences in working with the data.

A new technical concept is needed to facilitate the creation of such ecosystems around datasets. Figure 3 below presents the status quo, and then presents an alternative proposal.

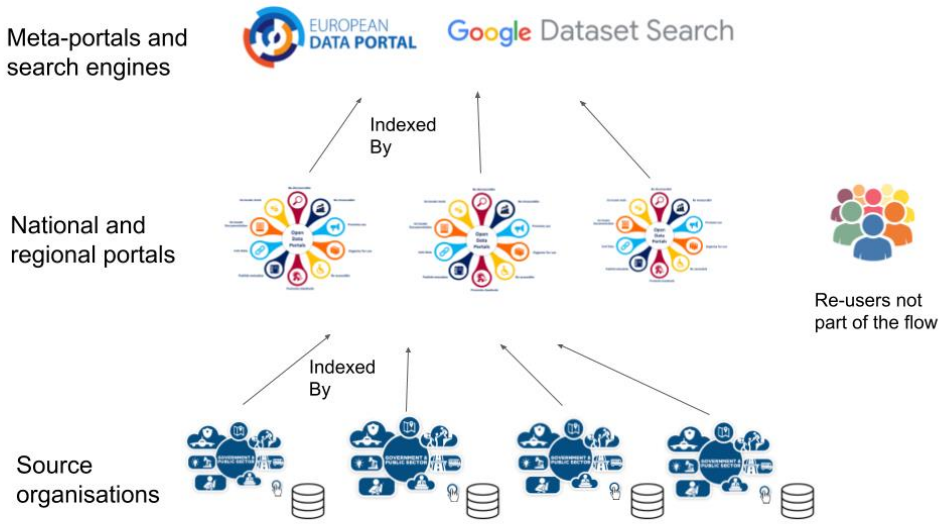

Meta-portals and search engines
EUROPEAN DATA PORTAL
Google Dataset Search
Indexed By
National and regional portals
Re-users not part of the flow
Indexed By
Source organisations

Figure 3: Open Data Portals Today: Centralised Portals, Indexed by Search Engines and Meta-portals

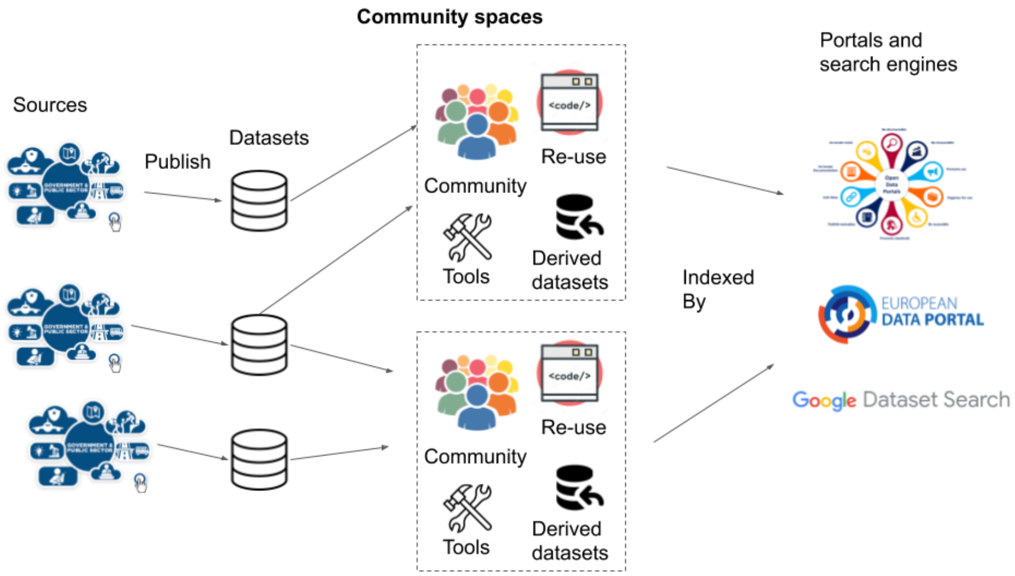


Figure 4: Users Join Spaces Organised Around Datasets and Share Tools, Develop Services and Apps, and Derive Further Datasets

In this model, data owners publish their datasets online on their own servers. If they would like their data to be discovered and used by others, they ensure that they remain compliant with technologies used by search engines such as Google Dataset Search or the EDP to index datasets. Datasets are maintained, reviewed, reused and enriched in so-called ‘community data spaces’, which co-locate tools, revisions, and derived datasets. The spaces could be hosted by data owners or current open data portals (centralising the technical infrastructure), or by the communities themselves (decentralising the infrastructure). In addition, they would be equipped with metadata generation capabilities to allow spaces and their work around datasets to be discovered by crawlers and harvesters such as EDP and maintain provenance links between original datasets released by government and derived datasets from the community. They would also have metadata interlinking capabilities to foster cross-fertilisation among communities.

A basic version of this concept can be seen today on Kaggle or on data.world.[17] In the former, each dataset published on the platform is linked to so-called ‘kernels’, which consist of code that works with that data with no installation or pre-processing needed, and notebooks, which document specific data projects.

5.2.1 Methodology

This work is in two parts, the design and building of a solution based on Distributed Version Control (DVC) tools, and testing of the prototype. DVCs are, by design, implementing several of the user-oriented principles laid out in the “Future of Open Data Portals” report[18] and are popular in other technology areas for their community-fostering capabilities. They include functionality for sharing, collaboration and reuse of resources created and used by multiple parties. We innovated on this in

[17] data.world

[18] https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf

that these resources are traditionally code and associated documentation, whereas in this variant they focus on data.

This concept extends an existing DVC tool with capabilities to publish, document and store data.

5.2.2 Results

The result is a new type of open data portal that is future oriented as a community data space, as it:

- Promotes use: by giving users the possibility of sharing how they have modified the dataset, and giving visibility to their contributions. Data stories and prominent case studies of the data could also be hosted by the portal
- Co-locates documentation and data: Each dataset has its own wiki and issue management section, which is one of the major bottlenecks in data reuse[19]
- Is measurable: version control technology has built-in metrics on the interaction of users with repositories, and allow the tracking of collaboration. This complements web analytics such as the ones used to assess portal use above
- Co-locates tools: for software development, version control tools go beyond storage and versioning of code and support continuous integration and development. In the case of datasets, tools for format conversion and data linking could be mirrored on the portal. This enables contributors that process datasets with the tools to participate in the collaborative development of the tool.

As noted earlier, DVCs were designed for code, and not datasets. Therefore, several extensions need to be implemented to facilitate data discovery, sensemaking and interlinking:

- Be discoverable: search engines in VC environments are optimized for source code and would not be effective for datasets. DVC requires enhancement as follows:
  - Automatic generation of the DCAT data catalogue of the portal, so it can be served from a SPARQL endpoint or consumed by a metaportal or search engine.
- Publish metadata: To encourage the improvement of metadata by the community, generate issues on missing metadata and missing properties in metadata (either mandatory or recommended). Open issues can be used to encourage contributors to solve them. Improved metadata also improves Google Dataset Search's ability to locate the data sets.
- Link data: Each dataset repository can contain a link-set to other datasets. Linksets can be automatically generated by a linking tool integrated into the portal, or be contributed by dataset consumers. Contributions by data consumers will be managed through the version control infrastructure. Linksets can then be consumed by clients or loaded into a SPARQL endpoint to enable queries and data stories. This is a crucial ability as very few portals currently enable this linking as measured by the 5 Stars of Linked Open Data, and yet it is probably the most well-known metric for assessing data.[20]

[19] https://dl.acm.org/citation.cfm?id=3025838

[20] http://5stardata.info/en/

5.2.3 Who Should Use This and How

Portal owners, especially at regional and national levels, can use this to inform the development of user-oriented features. Similarly, platform developers can be guided when adding further features into their offerings.

The open data use community can use this as a base to engage with portal owners.

5.2.4 Lessons and Best Practice

Community data spaces can create a 'virtuous circle' as improvements and changes made to the datasets by one user can be accessed by another, increasing the value of the datasets.

Hosting CDSs in a portal requires an extra investment in storage, as datasets need to be copied to enable full functionality. In terms of computing power, needs are a function of the number of expected users and of the requirements of other tools that a portal would like to co-locate.

Big datasets are difficult to include in this paradigm, however, most of the large open datasets are geospatial or satellite images that have their own infrastructure and their own set of specialist tools and community

Data that is accessed by an API won't take advantage of versioning capabilities, however there is still room to co-locate ways to access the data (call the API), and its usage.

A way to take further advantage of the investment in storage and cloud to implement this approach is to use it to include smaller communities and organisations that may not have the resources to run their own local portal

6. Conclusion

After two years of research across all of the four areas discussed above, two clear and cohesive themes have emerged which affect portal sustainability. The first is the necessity of building portals that allow users to do more with data. The second is the necessity of building ecosystems that will engage those users and create beneficial network effects.

Measurement is still one of the more complex challenges facing open data. Measuring data should not be left until after data has been published and used. Impact sought must be a starting point, and identified in business cases for data opening. Data should not be opened in the hope that it will prove useful, it should be opened purposefully, preferably based on evidence that it will be of use. Measurement should be more standardised and less ad hoc, in order to reduce the cost. No tool is perfect, although some are underused and more attention needs to be paid to exactly what is being measured, which often makes the ever-popular case study, which has low comparative value, not very useful.

However, what is clear is that, if we want to measure use and reuse, then portals need to encourage these activities, and encourage users to engage with each other to support these activities. Using technologies that support this by allowing individual users to benefit from and develop ecosystems, and to do more with data is vital.

Being able to do more with data and to build ecosystems again underpins the developing financing landscape. This report demonstrates that the portals that have successfully moved away from a simple publishing model to more complex user and publisher ecosystems are those that are able to develop more sustainable funding models. Portals which provide more consistent data, more formats, more updates and more expert advice on how the data can be used are portals which can move to freemium models. Portals which build extensive ecosystems around them can recover costs via budget reductions elsewhere or increased use of services.

Integrating new portal technologies that allow the integration of users with the data, such as distributed version control is therefore key for sustainability for open data portals.