Ensuring the Economic Sustainability of Open Data Portals:
Understanding Impact and Financing

This study has been prepared by the Open Data Institute 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[i] 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:

Written by:	**Reviewed and with contributions by:**
Jamie Fawcett	Wendy Carrara
Gillian Whitworth	Heleen Vollers
Lucia Chauvet	
Luis-Daniel Ibáñez	

DISCLAIMER

---

[i] At the time this report was first issued the consortium consisted of: Capgemini Invent, Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, the Open Data Institute, Time.Lex, and the University of Southampton.

Abstract

This is the second of two reports which explores the Sustainability of Open Data Portals across Europe. Open Data Portals are a critical part of our data infrastructure: they connect data holders with data users, who in turn create services that citizens and businesses benefit from and rely on.

The overall maturity of portals across Europe continues to increase as more and better quality data is published leading to more innovative products and services. However, as Open Data initiatives become an accepted, regular function of government, portal teams have increasingly been asked to justify spending and prove their impact. Portals are becoming increasingly ambitious, at the same time facing the budgetary constraints imposed on other mature functions of government. Ensuring that countries create the environment for sustaining portals, is more than ever critical for securing the economic, social and environmental benefits of Open Data into the future.

This report sets out how portals can create such an environment by improving how they monitor Open Data use and impact, at the same time as ensuring efficiency in the delivery of the many functions government's need to carry out. It provides practical recommendations in these areas based on current best practices outlined by portal owners and experts across Europe through a survey and in-depth interviews. Building on these best practices, it lays out potential approaches championed by experts to help build a sustainable financial future for Open Data portals in Europe and beyond.

Résumé

Ce rapport est le second d'une série d'études sur la pérennité des initiatives Open Data en Europe. Les portails Open Data offrent accès aux données et mettent les fournisseurs de données en relation avec les utilisateurs, qui à leur tour créent de la valeur pour les citoyens et les entreprises. Les portails Open Data sont, en ce sens, une infrastructure critique.

La maturité des portails en Europe augmente, permettant l'accès à plus de données de meilleure qualité et qui sont de plus en plus utilisées pour le développement de services innovants. Toutefois, alors que les initiatives Open Data se développent, ces activités sembleraient laisser place à nouveau au quotidien. De plus en plus d'équipes responsables de portails nationaux ou locaux se retrouvent obligées de justifier leur activité ainsi que leur budget, mettant en péril l'amélioration continue du portail ainsi que de la qualité des données publiées. L'enjeu est désormais de pérenniser ces initiatives Open Data et notamment les portails de données afin de sécuriser les retombées économiques, sociales et environnementales promises par l'ouverture des données publiques.

Ce rapport propose plusieurs pistes qui permettent de créer l'environnement nécessaire à l'évaluation de l'utilisation de l'Open Data, son impact ainsi que l'efficience qu'une utilisation plus accrue permettrait au sein de l'administration publique. Grâce à des entretiens avec les équipes responsables de portails nationaux et avec des experts européens, le rapport présente une série de recommandations pragmatiques. Ces recommandations, dérivées de cas d'usages et de bonnes pratiques européennes et internationales en la matière permettront aux acteurs de l'Open Data d'envisager l'avenir de leurs initiatives de façon plus pérenne.

Table of contents

List of Figures

Executive summary

The data landscape is rapidly evolving. Open Data now underpins a variety of products and services that have contributed to economic growth, allowed new businesses to thrive and improved the lives of citizens across Europe.

As Open Data is primarily made available through Open Data portals – i.e. catalogues and hosting solutions which make Open Data easier to find. A portal's ability to be sustainable – i.e. respond and adapt to challenges and secure ongoing financial support – is critical for maintaining and growing the benefits of Open Data into the future.

In the second of two reports on portal sustainability, we examine two areas that are increasingly important in the portal landscape: the question of **monitoring Open Data use and impact** and **creating the financial environment to sustain portals.**

With Open Data initiatives becoming an important part of government infrastructure, portal teams have increasingly been asked to justify spending and prove their impact. Likewise, with many portals suffering from falling government budgets, ensuring that countries create the environment for sustaining portals, both in providing sufficient funding and a useful network within and beyond government, it is critical for securing the economic, social and environmental benefits of Open Data into the future.

Monitoring Open Data use and impact

This report finds that, in order to ensure the long-term sustainability of Open Data initiatives, portal owners and data publishers need to be able to effectively understand and communicate the progress and impact of their initiatives. Measuring portal performance and monitoring the use and impact of Open Data are serious challenges that portals across Europe are facing. These challenges are inherent to Open Data given its non-rivalrous nature, the lack of requirements to declare use and the manner in which it generates positive network effects.

While lots of studies have been carried out and different methodologies tested, many of the more comprehensive studies remain out of reach of portals - often due to the level of resource and expertise required. In addition, there has yet to be a push for standardisation of methodologies for measuring use and impact, and impact studies are often limited by geography or time period. While individual portals can learn best practices from the work that has been done to date, there is a wider need for a more joined up approach between the various methodologies being currently employed.

The current array of macroeconomic and microeconomic studies, business population surveys and showcases must be more readily joined up through sharing and opening up the detailed methodologies and the data that underpins them. This will help Open Data portals across the continent in demonstrating and improving their impact. At the same time, portals and others should be looking to understand how the various methodologies can be employed in parallel. By employing

multiple methods, the benefits of each approach can be maximised and their constraints can be overcome, whilst also reducing costs.

On top of improvements to current approaches and taking a more joined up approach, portals must also explore methods to automatically track use. Only 26% of portals have attempted to automatically track reuse, despite this potentially lowering the resource burden of identifying users and minimise self-reporting bias. Over half of portals think it is not feasible. This report scopes several potential low-cost options, including tracking API usage, implementing version control systems and tracking through web search, that have high potential for portals in demonstrating impact.

Creating the financial environment to sustain portals

Additionally, as portals in Europe become more widespread, mature and advanced, the question of financial sustainability is becoming increasingly important. Many portals are facing a potential slow-down or loss of funding at the same time as they are becoming more ambitious. Here, we find that the vast majority of EU portals are offering Open Data on a marginal or zero-cost model, covered by government funding. It is unclear how the sustainability of these portals, and the Open Data initiatives and innovation they support, would be influenced by a sudden loss of financial support.

In this report, we find that although 94.4% view their funding model as sustainable, there are many hidden costs beyond just the technical site. These are concentrated in three main areas: maintaining portal infrastructure, engaging and encouraging publishers and building awareness, engagement and innovation.

Within each of these dimensions, portals are making decisions which have significant costs and impacts for their activities: whether they choose to adopt open source software or a third party hosted platform, what model they adopt for engaging publishers, and how far they can engage the community beyond government. Furthermore, national and sub-national portals live in very different economic realities when it comes to their funding structure and costs, and therefore the choices they are able to make – especially around supporting engagement and innovation. Some countries are adopting innovative approaches, like sharing the maintenance and governance of portals, to build stronger financial sustainability; encouraging collaboration within and beyond government appears to be the strongest way for supporting the costs and development of Open Data portals into the future.

Cementing portal sustainability

Looking to the future of portals, countries must consider how meeting and encouraging demand and providing value may impact the way that portals provide Open Data in future. Evidence suggests that in the future Open Data may take a more distributed, API-based approach to publishing. Such an approach might lead to rapid changes in the nature of portals technical infrastructure – for instance where portals function as a front-end catalogue for datasets accessible elsewhere, instead of the current centralised hosting model that is prevalent in many European countries. Whilst the technical infrastructure for Open Data publishing might change, it is important that countries, regions and cities maintain the considerable expertise of portal teams through sustainable funding, given their wider role in driving the publication of Open Data and the subsequent innovation. Whatever the outlook for

the technical process for publishing, the sustainability of portal teams remains an important consideration.

From setup to sustainability, portals across Europe are changing rapidly. There is more to be done to secure the future of portals, both in tracking, measuring and encouraging impact, as well as implementing funding models and relationships that will ensure financial sustainability in the long-term. This will enable the benefits of Open Data portals, and the Open Data they provide, to be shared by all.

Recommendations

Recommendations for portals: ensuring impact

Measuring impact through existing approaches
- Adopt and adapt the **Common Assessment Framework** to measure portal performance, identifying and using relevant existing metrics around Data and Context/Environment.
- Ensure **macroeconomic and microeconomic impact studies** provide **clear, detailed and repeatable methodologies** and **publish underlying data and tools**, which allow these calculations to be repeated
- For **business populations and user surveys,** partner with other organisations, **examine existing studies** and pose consistent questions, **publish the underlying data** as Open Data on the portal and make efforts to **automatically collect** and analyse the data on an ongoing basis.
- Establish **showcases and use cases** that allow users to submit their own re-uses, **encourage reporting** of re-use through community engagement, follow up with showcase re-users on a regular basis, **link use cases to the specific datasets** that are used and **collect more structured data** that could be linked.
- For **automated access metrics**, use page analytics and **track downloads** at the dataset level, keep APIs logs, and **publish access data under open licences**

Measuring impact through new approaches
- Use **holistic approaches** that focus on **use and impact at a dataset level** and examine approaches to **automating microeconomic analysis** based on the ongoing data collection approaches.
- **Share data** by **publishing underlying data** from studies, using methods from other portals to **infer and compare use** and impact, and **share metrics** for data not published on the platform.

Technical methods for tracking re-use
- For **automated approaches to re-use**, explore how **tracking APIs**, creating version control hosting and **web searching technologies** could be used to track use.
- Explore **data citation**, by developing **'model citations'** that indicate best practice, **guidance around granularity** of data for citation and guidance that **explains the purpose** of data citation tracking

Recommendations for portals: creating a sustainable environment

Current approaches to funding portals
- Choose either a third party hosted solution or build an instance of a popular open source platform such as CKAN or DKAN.
  - If choosing a hosted portal, ensure strong procurement processes are in place to identify the best suppliers.
  - If choosing to host an open source platform, ensure there are sufficient levels of technical knowledge on the team, either through hiring or upskilling existing staff.
- Look at ways to share the cost burden of technical hosting by partnering with other portals or governments to develop features, partnering with other governments to share a platform or bringing in technical expertise through fellowship programmes and other partnerships.

Engaging and encouraging publishers
- National portals should aim to have individual named contacts within each government department, if resource permits.
- Portals should aim to reduce costs by engaging multiple publishers through the provision of workshops and training.
- Portals should explore the use of self-service tools that help engage publishers such as dashboards, usage reports and rankings.

Building awareness, engagement and innovation
- Invest in events like hackathons as a method for both encouraging innovation among re-users and greater publishing of Open Data by government ministries
- Look for opportunities to create shared innovation strategies and funds across cities and regions
- Form partnerships with organisations beyond government, such as universities, and invest in joint ventures that may produce long-term impact

Alternative funding models
- Explore how to adopt a freemium model to datasets that are not yet open

1. Introduction

The data landscape is rapidly evolving. Open Data now underpins a variety of products and services that have contributed to economic growth, allowed new businesses to thrive and has improved the lives of Europe's citizens.

Open Data is primarily made available through Open Data portals – catalogues and hosting solutions which make Open Data easier to find. They enable more and better products and services to be built to benefit citizens. Consequently, their ability to be sustainable – respond and adapt to challenges – is critical in maintaining and growing the benefits of Open Data into the future. The teams that run Open Data portals are at the heart of most Open Data programmes; ensuring their sustainable access to resources is key to the success of Open Data initiatives. Portals and the teams running them must be flexible and durable enough to adapt to the changing data landscape.

The European Data Portal (EDP) was established in 2015 by the European Commission. It harvests metadata from the publication of Open Datasets in national, regional and local portals across the European Union, and seeks to improve the accessibility and usability of EU Public Sector Information. As well as operating as a portal, the EDP provides training materials and guidance for Open Data publishers and re-users, as well as promoting and showcasing data re-use around Europe.

This is the second of two reports which explores the sustainability of Open Data portals across Europe. Open Data portals are a critical part of our data infrastructure: they connect data holders with users, who in turn create services that citizens and businesses benefit from and increasingly rely on. Consequently, we must ensure portals are fit for purpose, now and in future.

1.1. The importance of sustainability

This report builds on our findings and recommendations from 'Recommendations for Open Data Portals: from setup to sustainability' published in February 2017.[1] For our first report, we conducted in-depth interviews with European Data Portal owners and civil society representatives, as well as building on practical experience and secondary research, to surface recommendations and best practices for portal owners to help ensure the sustainability of their portals.

Previously, we found that a sustainable Open Data Portal is one that continues to respond to core challenges and has the governance, finance, architecture and operational models to do so. It also uses metrics effectively to drive improvements in data quality and usefulness, and attempts to measure the impact from Open Data re-use as part of a support base for continued funding. At the time of writing in early 2017, we saw that although portal owners had taken some measures or made ad hoc upgrades to their portals, none had developed models or approaches to sustainability which encompassed all these areas.

1.1.2. Understanding sustainability of Open Data portals

In our previous report, we examined what makes Open Data portals sustainable in four key areas: **governance, financing, architecture and operations**. We also explored how **metrics** can be used to monitor and evaluate progress, and the feasibility of automating these metrics.

---

[1] European Commission (2017). Recommendation for Open Data Portals: from setup to sustainability. Available at: https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

Regarding **governance**, we discovered that portals are often established in the early stages of an Open Data initiative. They are often created without the need for a comprehensive business case or user research to help make the case for funding. This was primarily because Open Data was a strategic political objective, and administered separately to digital functions and strategies for government. Although these features have enabled the rapid progression of Open Data initiatives in their early stages, sustainability requires portal governance to be embedded in 'business as usual' government functions so it can adapt to changing government priorities.

Sustainable Open Data portals also require ongoing **financing**, for infrastructure and maintenance, as well as any outreach, training and support for publishers and re-users of data. A sustainable financing model needs to allow the team operating the service to maintain current operations, plan its strategy to work with a known budget, and have confidence in its longevity by accounting for updates and enhancements to the portal as they are required. Portals with a funding strategy give publishers and re-users confidence that it is going to be a sustainable mechanism for accessing Open Data into the future. We explored how portal owners could take responsibility for setting funding strategies and budgets, and how potential revenue streams could be generated to enhance these.

We found that the sustainability of **architecture**, in particular software, is not usually a key consideration in choosing solutions. As Open Data initiatives mature, and data services and technologies evolve, ensuring portal architecture is still fit for purpose and able to withstand service and funding changes becomes harder. Through discussions with stakeholders, we recommended that selecting open source software solutions can help avoid 'lock-in' situations, while researching user needs and their preferred data formats can drive data improvements and building links to other data portals can increase relevance for local users.

Operations were another challenge for ensuring sustainability. The operations of a sustainable Open Data portal involve a portal owner running it day to day, ensuring that it continues to be responsive and useful for portal users. Portal services often extend beyond the technical infrastructure underpinning the portal – such as its servers – to the practices and procedures that ensure it provides access to useful, high-quality, discoverable Open Data. Furthermore, the nature of Open Data portals mean that portal owners are often required to be more open, transparent and adaptive to end-user requests than may be expected of other services. These factors make it challenging to create sustainable portal operations, leading us to issue a number of recommendations (see table below).

Finally, the first report examined how **metrics** can be used to drive improvements and demonstrate impact. For an Open Data portal to be sustainable, the data accessible through it needs to be relevant, timely and usable, and there must be evidence of use and impact to argue for continued investment. However, portal owners are often not the publishers or maintainers of data on their portal, and the extent of the data's re-use can be difficult to track, making it hard to measure for impact. This led us to explore how metrics can assess both the outputs - the amount of work produced - as well as the outcomes - impact and sustainability - of Open Data, using automated means.

The recommendations from the first paper can be found below:

Recommendations from the first edition of the Sustainability of Open Data Portals report

Governance
- Have a business plan and clear governance structure in place
- Bring publishers and data users together to address specific challenges, using Open Data from the portal
- Build responsiveness to government priority changes into your governance structure
- Create hard levers to set and enforce data quality and metadata standards, and pursue data updates from publishers
- Create a non-ministerial leadership role to champion data publication and respond to issues

Financing
- Be open about your funding strategy, so that people publishing and accessing data from the portal can identify future needs, use cases and potential funding shortfalls
- To maximise scope for portal improvements, and reduce the impact of funding cuts, ensure your priorities (training, support for publishers, user engagement) align with those of your funding source(s)
- Ensure that your own role as portal owner includes responsibility for setting funding strategies and budgets
- Perform, commission or identify research into the impact of your portal's current or potential activities, to develop and support a business case for future funding.

Architecture
- Select open source software solutions, and solutions that offer archiving/downloading options for all data published via the portal
- Contribute to the development of standard APIs, that could be used across all Open Data platforms, for sharing, summarising and presenting data
- Build links to data held in other portals into yours, where they could be relevant for your local users
- Even if not responsible for the publication and maintenance of data, research your user needs and their preferred data formats to drive data improvements
- Build upon recognised standards to foster interoperability and comparability of metadata across Europe

Operations
- Manage publication operations to support different types of publishers from small- to large-scale, enabling automation where possible
- Manage technical operations to include effective monitoring and reporting systems for inaccessible data, preferably through publicly accessible lists for users to track progress
- On-board new end-users, publishers and monitors with effective User Experience design, clear publication processes, feedback loops and training
- Automate functions to ensure seamless integration of a diversity of data sources, increase user friendliness and limit overheads for stakeholders
- Capture and share lessons learned, and be open to best practices and standards developed by other portal operators

---

> **Automated metrics**
> - Choose metrics that help to benchmark data publisher performance, but do not rely on one metric e.g. quantity. Combine quantity metrics with data quality and engagement measures
> - Choose metrics that help potential data users find data that is suitable for them to use. Evaluate whether the metrics chosen are meaningful or potentially misleading
> - In measuring quality of Open Data, consider metadata and contextual information to increase user understanding and engagement
> - Overcome challenges in automating metrics by adopting standard language and terminology, publishing processes and metrics for data quality and re-use

1.2. Building on the lessons learnt from the first European Data Portal sustainability report

Since the first report, the European Data Portal has continued a programme of peer learning and capacity development. In 2017, the EDP conducted 6 tailored support workshops in Slovenia, Madrid, Malta, Romania, Sweden and Ireland, attended by 19 national Open Data Portal teams. In May 2017 and hosted by the ODI, the EDP conducted a three-day 'European Data Champions' programme to strengthen peer networks and build capacity. In addition, in October 2017 and again hosted by the ODI, delegates from Germany, Finland, Poland, Malta and Scotland attended the Open Data Leaders Network, which allowed those who manage or contribute to portals to learn from their colleagues around the world as also non-European delegates attended the workshop. These activities have allowed the authors to gather informal feedback from portal owners regarding the content of this second report.

We also solicited feedback on the first paper from portal owners. Of respondents to our survey, 79% who answered said they found the recommendations useful. One respondent, representing the Bath:Hacked community data portal said the report was a "useful summary of a variety of factors that will help contribute towards making our portal sustainable and generate impact". Respondents said it was useful to have an overview and summary of current practices in Europe, with a proportion of them using it to confirm their existing approach and measure themselves against the approaches taken by other portals. A number of respondents picked up on specific sections they found particularly useful, primarily those about governance and operations. The respondent from Latvia's Open Data programme which was in the process of launching their Open Data Portal, explained "we already had the specification and budget set in the time of the report coming out, but the organizational and legal conclusions were useful for addressing ongoing challenges". In response to the report, the Spanish national Open Data Portal, datos.gob.es[2], examined each of the topic areas providing open feedback[3] on the report.

We also solicited further feedback from respondents in interviews. Respondents from Luxembourg's national Open Data Portal and the Helsinki Infoshare portal both highlighted the particular importance of political support as a part of the governance recommendations. The representative of Luxembourg's portal went into depth about the importance of making Open Data a non-partisan issue, highlighting how new legislation in Luxembourg was codifying Open Data into law. The same representative also raised the importance of greater collaboration between data portal owners, in

---

[2] Datos.gob.es. Available at: http://datos.gob.es.
[3] Datos.gob.es (2017). Recommendations for Open Data portals: from setup to sustainability. Available at: http://datos.gob.es/en/noticia/recommendations-open-data-portals-setup-sustainability.

particular sharing lessons and technology decisions. They highlighted the importance of their relationship with France's established Open Data Portal programme in helping them to set up their programme effectively. The EDP has been working to encourage both formal and informal peer networks between portal owners to help address this point.

For this report, we engaged portal platform providers in order to understand the lessons they had learnt and gather their expertise on implementation and feasibility of any proposed actions.

This report was also informed by the results of the 2017 Open Data Maturity in Europe report, produced and released in November 2017 by the EDP.[4] This allowed the authors to understand the latest activities of European portals and changes over the last year, and since the first sustainability report was released.

1.2.1. Gaps identified in portal sustainability

Over the last few years, the Open Data portal landscape in Europe has been rapidly changing. With the 2017 Open Data Maturity report finding that the number of trendsetters in the EU28+ countries has more than doubled between 2016-17: countries are engaging in a 'race to the top' when it comes to Open Data, and investing in initiatives to achieve their goals. In countries across Europe, portals are now a regular part of the structure of government, a tribute to the work of Open Data activists and civil servants in embedding Open Data initiatives within governments. Open Data is producing social (and environmental), economic and political impact.

However, these changes have also brought questions of using public funding appropriately and sustainably to produce impact from Open Data now and into the future. As portals become more mature, teams are increasingly asked to justify their allotment from a limited pool of public funding by demonstrating the impact of their initiatives. Those portals who are struggling to attract public funding may have to adopt different cost structures or activities to raise money that may affect their survival into the future.

In this context, this report examines two areas that are increasingly important in the portal landscape: the question of **monitoring Open Data use and impact** and **creating the financial environment to sustain portals.** In this section, we explore the evidence behind an increasing focus on impact and sustainability among portals in Europe, and establish the scope of these issues that will be covered in the report.

Although many other interesting gaps were identified, we were not able to address them all directly in this report. For instance, specific gaps around the technical aspects of data catalogue harvesting and DCAT-AP standards were highlighted by some respondents. While these issues require attention, they were deemed to be too limited in scope for this research. Furthermore, training and other activities provided by the EDP are more focused on responding to these concerns.

1.2.1.1 Open Data use and impact

In preliminary research and discussions for this report, the need to address the question of monitoring Open Data use and impact was clearly established.

---

[4] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

The issue of Open Data use and impact clearly builds on the Automated Metrics section of the previous sustainability report, as well as the recommendation to *"perform, commission or identify research into the impact of your portal's current or potential activities to support a business case for future funding."* In a survey of portal owners, the respondent to the survey representing Spain's national Open Data Portal suggested "the most relevant point for the next study would be: what efforts are or could be made to measure the re-use and impact of Open Data automatically, and the technical and financial feasibility of setting up such metrics." Another respondent representing Austria's national Open Data Portal highlighted the need to input from existing monitoring methods such as the "Open Data Barometer or Open Data Index".

Evidence from interviews with prominent actors in the sector underscored the need for portals to receive more support around tracking impact. Several portals cited tracking impact as one of the core challenges they faced – including representatives from Belgium, Helsinki and Luxembourg. The potential issue of not being able to evidence their impact was keenly felt by many, who felt it was hard to justify extra resource or expenditure without a clear picture of the return. Several respondents alluded to falling government budgets as a key concern for portals and for wider data publishing. Being able to provide economic figures was raised by a number of respondents, although others suggested other metrics were also useful. The need for systematic efforts which were relatively low cost or required little resources was also a key need identified amongst respondents.

Furthermore, the Open Data Maturity in Europe 2017 report demonstrates that many countries in Europe are only just beginning to formally assess the economic impact of Open Data. In 2017, only 32% of EU countries had conducted studies assessing the market value of Open Data in the past 2 years; 64% had yet to undertake any assessment. 50% of countries have also not undertaken (or are unaware of) any additional studies to measure any kind of impact in the past 2 years. Consequently, evidence indicates that tracking and measuring the impact of portals and Open Data initiatives in general is an important point for examination, and an opportunity to learn from the best practice of those that are implementing such methods.

This report addresses the concerns by examining the current methods of tracking and measuring impact, and technology that may help improve these methods in the future. Given the focus of several respondents on economic figures, and the important ties of these figures to securing sustainable funding from government, we take a particular focus on demonstrating direct and indirect **economic impact.** In the 2017 Open Data Maturity in Europe index, the EU28 average on the economic impact of their Open Data initiatives constituted 54%, only a 3% point rise on 2016 figures, compared to greater rises in political and social impact on 2016 (figures for political and social impact rose by 11% points and 21% points respectively). Consequently, this report aims to support portals in becoming more confident in tracking economic impact.

1.2.1.2.   Creating an environment to sustain portals

Connected to concerns around potential future budget cuts, our preliminary surveys and research for this report identified the need to explore the wider environment required to sustain Open Data portals. This builds on the financial and governance sections in the first sustainability report. Creating the environment for sustaining portals, both in providing sufficient funding and a useful network within and beyond government, is critical for securing the economic, social and political benefits of Open Data into the future.

Primarily respondents, as with Open Data use and impact, were concerned with justifying additional resource costs required to create wider impact. What was especially identified was whether portals

believed their funding and purchasing models were sustainable for long-term development. Another key point made by respondents was an understanding of how to build relationships within government to help enable portals to be more sustainable.

Again, this report establishes a primary focus on the **financial sustainability** of portals, an aspect which has been a source of concern for many European countries. 71% of the EU28 countries cited financial concerns as a barrier to Open Data publishing, particularly around financial resources for Open Data and supporting the work of portals that are provided on a zero-cost model. Consequently, this report takes an important focus on this aspect of sustainability.

1.3. Methodology

This research took a targeted mixed methods approach using desk research, a short online survey and targeted in depth interviews to gather quantitative and qualitative responses around funding, relationships and existing attempts to track access, use and impact of Open Data. The desk research involved an extensive overview of existing research, methodologies and current and potential technical approaches. The online survey was aimed at 'Open Data portal owners' across Europe and managers who are responsible for data portals.

This short survey was launched in summer 2017, receiving 19 valid responses from portal owners from 14 countries across Europe. There were no more than two respondents from any one country. Respondents included 11 national portals (58%), 7 local/city portals (37%) and one community run portal (5%).

The portals represented in the survey are laid out in the table below.

Portal	Geography covered	Type of portal	City/Country Open Data maturity 2017
data.gv.at	Austria	national portal	Trendsetter
opendata-ajuntament.barcelona.cat	Barcelona/Spain	local portal	Unclassified/Trendsetter
data.bathhacked.org	Bath/UK	community-run portal	Unclassified/Trendsetter
opendata.government.bg	Bulgaria	national portal	Trendsetter
opendata.comune.fi.it	Florence/Italy	local portal	Unclassified/Trendsetter
data.gouv.fr	France	national portal	Trendsetter
hri.fi	Helsinki/Finland	local portal	Unclassified/Trendsetter
data.public.lu	Luxembourg	national portal	Trendsetter
opendata.paris.fr	Paris/France	local portal	Unclassified/Trendsetter
podatki.gov.si	Slovenia	national portal	Trendsetter
datos.gob.es	Spain	national portal	Trendsetter
dati.trentino.it	Trentino/Italy	local portal	Unclassified/Trendsetter
data.gov.uk	UK	national portal	Trendsetter
data.gov.be	Belgium	national portal	Fast-tracker
daten.berlin.de	Berlin/Germany	local portal	Unclassified/Fast-tracker

opendata.gov.cz	Czech Republic	national portal	Fast-tracker
data.gov.lv	Latvia	national portal	Fast-tracker
open.wien.gv.at	Vienna/Austria	local portal	Unclassified/Trendsetter
data.gov.ru	Russia	national portal	Unclassified

Figure 1 Portal maturity of survey respondents, as measured by the 2017 EDP Open Data Maturity in Europe report

The majority (74% or 14 out of 19 respondents) of respondents were from portals based in countries with national portals being classified as 'trendsetters[5]' in the EDP Open Data Maturity in Europe 2017 report. A small minority (21% or 4 out of 19 respondents) were 'fast-trackers[6]', and only one country, Russia (5%) was unclassified[7]. We chose to target 'trendsetters' in order to capture the lessons they have learnt to help other portals to understand what success looks like in different contexts. Importantly, it allows us to identify and highlight examples of emerging best practice.

These surveys were then analysed, and certain respondents were identified for targeted in depth interviews based on their responses. These interviews were then used to explore and develop the topics identified in the survey by portal owners to develop a greater understanding of current approaches to funding portals, building relationships and tracking access, measuring use and inferring impact of Open Data. Given that some portals are hosted solutions, where the technical portal infrastructure is built and managed by software companies as a paid service, we also wanted to include insight from portal providers around charging models. Given their high level of technical expertise and need to prove the value of their services, we also wanted to capture their processes and plans around measurement of use and impact. We also wanted to incorporate additional insight from others who had carried out impact assessments and other approaches to monitoring use and impact.

As such, we carried out 7 interviews with portal owners, 2 interviews with providers and 2 interviews with impact assessors. Interviewees were largely from 'trendsetter' countries: Spain, Finland, Luxembourg, Italy and the UK with one interviewee from Belgium, a 'fast-tracker' country and one interviewee from Russia. We chose our interviewees in a similar manner to our targeting of surveys, aiming to draw out in more depth the lessons and approaches taken by more mature portals who could share their experiences with other portals. We interviewed OpenDataSoft, an Open Data platform provider based in Paris, and Urban Tide, a young UK platform provider specialising in real-time data provision, to understand how they charge for services but also how they currently track use and impact, and what plans they have to respond to their customers demand for future tracking. We interviewed Transport for London (TfL), an early adopter of publishing open transport data, because they recently carried out an impact assessment. We also interviewed Amazon Web Services (AWS), who host public sector Open Data for free through their cloud services to understand how they track re-use as a commercially driven, highly technical organisation.

---

[5] Trendsetters are those who have implemented an advanced Open Data policy with extensive portal features and national coordination mechanisms across domains.
[6] Fast-trackers are those who have significantly accelerated their Open Data journey, having either a policy or a portal that is substantially developed, however they still face a small number of shortcomings.
[7] Followers are those who have successfully developed a basic Open Data policy and have brought in more advanced features on their portal while limitations still exist in terms of data release.

The portals and organisations represented are laid out in the table below.

Portal / organisation	Type	Geography covered	City/Country Open Data maturity 2017
Luxembourg data portal	Portal owner	Luxembourg	Trendsetter
Florence data portal	Portal owner	Florence/Italy	Unclassified/Trendsetter
Russian data portal	Portal owner	Russia	Unclassified
Belgium data portal	Portal owner	Belgium	Fast-tracker
Red.es	Portal owner	Spain	Trendsetter
Helsinki	Portal owner	Helsinki/Finland	Unclassified/Trendsetter
data.gov.uk	Portal owner	United Kingdom	Trendsetter
Urban Tide	Portal software provider	N/A	N/A
OpenDataSoft	Portal software provider	N/A	N/A
Amazon Web Services	Other	N/A	N/A
Transport for London	Other	N/A	N/A

Figure 2 Country Open Data maturity of interviewees, as measured by the 2017 EDP Open Data Maturity in Europe report

1.4. Report structure

This report is divided into two parts which aim to address the key questions raised in the introduction. Part 1 of this report examines the challenge for portals around monitoring Open Data use and impact. It begins by exploring the theoretical challenges posed by measuring Open Data use and impact, drawing on existing research, and sets out a conceptual model for understanding use and impact. Using this model, the report examines the current approaches to measuring Open Data impact, use and access being employed by portals and publishers. For each of these approaches, it draws out lessons and best practices from portals, providing recommendations on how to maximise the usefulness of these approaches. The section finishes by exploring several potential methods for improving impact evaluation in the future.

Part 2 of the report is focused on creating the environment in government to sustain portals. The first section focuses on the need to sustainably fund portals, examining the current funding models and approaches, as well as the associated costs of running Open Data portal programmes across six different areas – technical, activities, training, awareness, organisation and legal. It also examines alternative models for funding portals and finally draws out lessons, best practices and recommendations for portals. The second section focuses on the need to build relationships within government and the impact these have on sustainability. This section examines the current approaches and their impact on portal sustainability. The report concludes with a discussion of the requirement to sustain portals and a synthesis of existing recommendations and best practices to help guide portals' decisions in the future.

2. Part 1: Addressing the question of Open Data use and impact

Over the past ten years, many local, national and supranational organisations have invested in Open Data portals and publishing initiatives. One index[8] lists 524 data portals worldwide, while the European Data Portal harvests 73 portals in Europe[9] alone. While many initiatives at the beginning of the Open Data movement arose from a desire for increased transparency as part of the open government initiative[10], for many the focus has shifted towards using Open Data to create social and economic value. Similarly, as many Open Data initiatives have become more mature, portal owners and data publishers are increasingly asked to justify programme expenditure.

To ensure the long-term sustainability of Open Data initiatives, portal owners and data publishers need to be able to effectively understand and communicate the progress and impact of their initiatives. Portal owners and publishers who are unable to do this to the satisfaction of their funders may risk endangering their portal's sustainability: demonstrable impact is often central to maintaining political will and support, and continued funding.

Countries are investing in understanding the impact of the Open Data they publish and its value to the economy and society. The 2017 Open Data Maturity in Europe report found 13 countries (46% of the EU28 countries studied) have conducted studies monitoring the political impact of Open Data, including Austria, Finland, the Czech Republic and Latvia.[11] Over 50% of EU28 countries estimate the impact of Open Data on improving government transparency to be high. 50% of the EU28 countries have also published additional studies measuring the economic impact of Open Data from 2016 to 2017, including studies on business models in open government data in Austria and publishing transport data for maximum re-use in Belgium. These studies indicate that governments are actively exploring the catalysts and success factors needed for Open Data impact.

Building on our recommendation from the first EDP sustainability report to *"perform, commission or identify research into the impact of your portal's current or potential activities to support a business case for future funding"* and the discussion of automated metrics, we examine the options for portals when it comes to measuring their effectiveness.

In this chapter, we will examine the inherent challenges to measuring portal performance and Open Data, in particular, concerning their use and impact. We will then examine the current approaches taken by organisations to measure the use and impact of Open Data. Last, we explore how these current approaches can be improved and what new approaches might be adopted to help measure use and impact going forward.

---

[8] Data Portals. Available at: http://dataportals.org.
[9] European Data Portal. Available at: https://www.europeandataportal.eu/data/en/organization.
[10] Open Government Partnership. Available at: https://www.opengovpartnership.org.
[11] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

Recommendations for Portals: Ensuring Impact

Measuring impact through existing approaches

Adopt and adapt the Common Assessment Framework to measure portal performance.

Provide clear, detailed and repeatable methodologies for economic studies.

Partner with other organisations, build on existing studies and make efforts to automatically collect business population data.

Establish showcases and use cases, encourage reporting of re-use through engagement and link use cases to the specific datasets.

Use page analytics, track downloads and keep API logs at the dataset level.

Publish all access and use data and methodologies under open licences.

Measuring impact through new approaches

Focus on holistic use and impact measurement at a dataset level and examine automation of ongoing analysis.

Publish data from all studies, share detailed methods and metrics

Technical methods for tracking re-use

Explore how tracking APIs, creating version control systems and web search could be used to track use.

Explore data citation, through 'model citations', best practices and guidance.

2.1. Challenges to measuring the performance of Open Data portals

Several challenges are presented in measuring the performance of Open Data portals. While the measurement of most new technologies can prove difficult, Open Data has a number of characteristics that exacerbates these difficulties. In this section we explore the efforts to date which attempt to measure Open Data initiatives and examine what portals can learn from these efforts. We also review the key challenges portals face when trying to measure the use and impact of Open Data.

2.1.1. Approaches to benchmarking Open Data initiatives

Since the emergence of the first Open Data portals in the late 2000s, many efforts have been made to try to monitor the performance of Open Data initiatives. Most of these efforts have been focused on benchmarking - evaluating and ranking - the relative performance of countries, organisations and datasets. Often, the intention of these benchmarking efforts is to help implementers identify and implement best practices from other Open Data initiatives. They are also almost exclusively carried out by teams of independent researchers in global organisations, which enable them to be objective in their assessment. Below, we examine four of the most widely used efforts to understand the performance of Open Data initiatives and their relevance to understanding the performance of portals.

Example benchmarking initiatives

Open Data maturity in Europe - European Data Portal
The European Data Portal's Open Data Maturity in Europe[12], released annually, examines Open Data readiness and portal maturity in the EU28 countries and the European Free Trade Area (EFTA) countries Liechtenstein, Norway and Switzerland. The 2017 report also examines the EFTA country Iceland and the EU accession countries of Albania, Bosnia and Herzegovina, the Former Yugoslav Republic of Macedonia (FYROM), Kosovo, Montenegro, Serbia and Turkey.

The report assesses European Open Data progress and maturity from the perspective of public sector representatives. National initiatives are benchmarked, with index positions informed by Open Data readiness and portal maturity indicators (each of which have several sub-indicators). Each indicator is initially scored by researchers, then complemented with an additional survey to create individual country factsheets. Country representatives validate results before EDP clusters results to compare countries and publishes them online.

The index allows observers to assess the implementation of the EU's public information law. Understanding best practice for portals and Open Data initiatives also enables EDP to identify where group and specialised training for portal owners can help to build the capacity of initiatives. However, since the index's methodology does not look at the work of regional and local portals, its insights are mostly relevant to portals operating at the national level.

Open Data Barometer (ODB) - Web Foundation
The Open Data Barometer (ODB)[13], produced by the Web Foundation, analyses the readiness, implementation and impact of global Open Data initiatives across the world. The index aims to

---

[12] European Data Portal (2017). Open Data in Europe. Available at:
https://www.europeandataportal.eu/en/dashboard#2017.
[13] Open Data Barometer. Available at: http://opendatabarometer.org.

benchmark countries to illustrate 'the true prevalence and impact of Open Data initiatives around the world'. It is based on data from three sources:

- a peer-reviewed expert survey on policy, implementation, impact and assessing key Open Data categories
- government responses to the above survey
- secondary data from sources such as the World Economic Forum and World Bank

Researchers use this data to create three sub-indexes of readiness, implementation and impact, each weighted to build an aggregated ODB score. The scores are ranked, with higher scores indicating a better performance and a higher place. A report with accompanying data is then released annually.

The ODB is a useful tool for helping Open Data actors understand prevalent issues and draw common lessons from other Open Data initiatives. By measuring Open Data initiatives across the stages of readiness, implementation and impact, the index delivers a fairly comprehensive view of an initiative's maturity. However, the index's utility is limited for portals as the ODB methodology has no specific portal indicators, therefore making it harder for portal owners to draw learnings from their country's score. The Barometer also does not capture Open Data publication by the private sector, which may limit its usability as private sector data becomes more common.

Global Open Data Index (GODI) - Open Knowledge
The Global Open Data Index (GODI)[14] benchmarks the publication of open government data across the world. Using a crowdsourced survey, GODI measures the openness of government data from a civil society perspective to help governments understand their data gaps, make data more usable and create more impact. Run by the Open Knowledge Network, GODI currently contains 1410 dataset evaluations across 94 'places'.

The index's methodology contains four phases: submission (through a snowball sampling approach), review, quality assurance of review result and public dialogue. OKF work with volunteer researchers who assess the openness of 15 data categories 'that has proven to be useful for the public', such as budget and spending data.

The index's in-depth methodology encourages a high-quality standard for Open Data in its target countries. The innovative public dialogue phase and forum also promotes constructive discussion between key stakeholders in civil society and government, building partnerships and improving the reliability and accuracy of the final data. However, there are few explicit lessons that portal owners can draw from the data, and insights are limited to the data publication at the national level. Recently, some of these tools have been used to examine the openness of local and city data as part of OKF's Open Data Census[15] work.

OECD Open-Useful-Reusable Data (OURdata) index[16]
The OECD releases the OURdata index measuring the extent to which public sector data is available, easily accessible and usable by individuals and companies worldwide. Information is released in the *Government at a Glance* reports.

[14] Global Open Data Index. Available at: https://index.okfn.org.
[15] Open Data Census. Available at: http://census.okfn.org/en/latest/.
[16] OECD (2015). OUR Data Index: Open, Useful, Reusable Government Data. Available at: http://www.oecd-ilibrary.org/governance/government-at-a-glance-2015/our-data-index-open-useful-reusable-government-data_gov_glance-2015-70-en;jsessionid=1s8cfr9njxwh7.x-oecd-live-03.

The index uses survey responses gathered from government respondents, primarily Chief Data Officers, to assess government efforts on Open Data. This is measured against three core dimensions of 'good' Open Data practice: data availability, data accessibility, and proactive support from government for data innovation. Each country is assessed separately, with index scores comparable across countries.

The OURdata index is beneficial for portal owners as data availability and accessibility are explicitly based on national portal data. Consequently, portal owners can understand how their portal sits against international best practice (gathered from the G8 Open Data Charter). However, the index is clearly limited to scope to the public sector and does not consider the work of portals in collaborating with the private sector.

2.1.1.1.   Limitations of benchmarking initiatives for portals

Benchmarking initiatives are very useful, and each of the benchmarking initiatives that have been explored provide insight into one or more aspects of the performance of Open Data initiatives. However, as a method for understanding the performance of individual portals there are several limitations:

- First, benchmarking activities tend to focus on national level initiatives which makes them less useful for individual portals to gauge their performance. Even when portal specific metrics are included, they tend to focus on national portals or evaluate the national portal and lower-level portals at the same time.
- Second, comprehensive benchmarking activities are carried out by independent organisations and teams of experts. These tend to be very time and resource intensive, for both the researchers and respondents and means that portal owners are dependent on the schedule of updates to determine their performance.
- Third, benchmarking studies are inherently limited to comparisons between initiatives. These comparative studies rely on proxy measures for the use and impact of data but do not provide absolute measures of performance in terms of data use and social, environmental and economic impact. These high-level comparisons help to improve understanding of relative use and impact but not the overall use and impact of Open Data.

2.1.1.2.   Lessons learned for portals

Benchmarking initiatives are useful for high level comparisons, but they can also provide a good guide for how portals might seek to evaluate their own performance in more depth. Growing out of these benchmarking initiatives, several organisations involved – including Web Foundation, GovLab, ODI, OECD – came together in June 2014[17] to draft a *Common Assessment Framework (CAF)[18],* which attempts to provide a standardised methodology for a rigorous analysis of the supply, use and impact of Open Data. It aimed to build on many of the existing Open Data benchmarking tools and processes and help coordinate the efforts of researchers and organisations in designing comparable and complementary research.

---

[17] Davis (2014). Towards Common Methods for Assessing Open Data. Available at:
https://webfoundation.org/2014/06/towards-common-methods-for-assessing-open-data/.
[18] New York University (2014). Towards common methods for assessing Open Data: workshop report & draft framework.
Available at: http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf.

While the adoption and development of the framework appears to have faltered, it provides a useful set of four conceptual dimensions, which many performance initiatives adopt, at least in part:

1. **Context/Environment** – *the context within which Open Data is being provided. This might be national, in the case of central government's Open Data, or more specific, in a particular sector such as health, education or transport.*
2. **Data** – *the nature and quality of Open Datasets: their legal, technical and social openness, relevance and quality. The framework also looks to identify core categories of data that might be evaluated in assessments.*
3. **Use** – *the types of users accessing data, the purposes for which the data is used, and the activities being undertaken to use it.*
4. **Impact** – *the benefits gained from using specific Open Datasets, or from Open Data initiatives in general. Benefits can be studied according to social, environmental, political/governance, and economic/commercial dimensions.*

For portals, this framework might provide a generalised framework for understanding performance. In the case of individual portals, the Context/Environment dimension might be less relevant given they do not need to benchmark their starting point against another portal with a different background. There may be benefit in measuring their own performance in terms of community engagement metrics, for example how much outreach they do and events they hold.

In terms of existing work by benchmarking and other initiatives, metrics around Data – for example data availability and quality – can and are being re-used. In recent years metrics around the quality of data publication have been increasingly defined and standardised. For example, metrics used by tools such as the Open Data Monitor (ODM)[19] project and the ODI's Open Data Certificates[20] can be brought into portal evaluation. Many of these indicators and techniques have been included in benchmarking initiatives. For example, EDP's portal maturity assessment includes many of the automated metrics created for Open Data Monitor. Indeed, a number of the portal owners interviewed for this study had already defined metrics around their data publishing activities. Many of these developments were covered in the automating metrics section of the previous report.

However, formulating metrics and collecting data on the use and impact of Open Data are much harder. In 2015, the ODI published 'Benchmarking Open Data automatically[21]', in which they used the framework to rank each of the aspects against the feasibility and comparability of each of the four dimensions. They identified that data aspects were the most comparable and feasible to measure, while measurements of use and impact are the most difficult to measure empirically.

---

[19] Open Data Monitor. Available at: https://opendatamonitor.eu/frontend/web/index.php?r=dashboard%2Findex.
[20] Open Data Certificates. Available at: https://certificates.theodi.org/en/.
[21] The ODI (2015). Benchmarking Open Data automatically. Available at: https://theodi.org/guides/benchmarking-data-automatically.
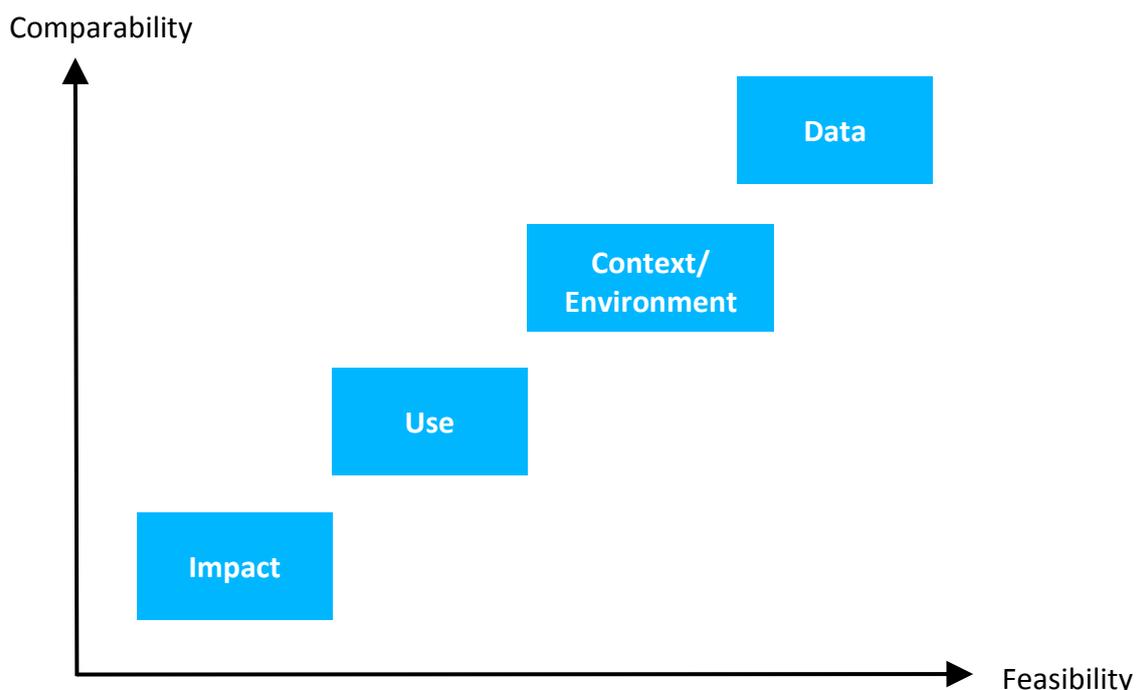
Figure 3 Feasibility and comparability of the four dimensions, under ideal circumstances[22]

Similarly, DataSF[23], which has taken a similar approach to measuring its performance around Open Data, identified use and impact as the most challenging to measure.

> **Recommendations for portals: monitoring performance**
>
> **Adopt and adapt the Common Assessment Framework to measure portal performance,** identifying and using relevant existing metrics around Data and Context/Environment.

2.1.2. The challenge of measuring use and impact of Open Data

It is widely acknowledged that empirically measuring the use and impact of Open Data presents significant difficulties. To understand why, we need to examine each of the challenges in depth. We will then use this understanding to describe a framework that we can use to understand the implications of these challenges for portals.

2.1.2.1. The challenge of measuring the impact of Open Data

One of the fundamental questions Open Data publishers face is how to measure the impact of data publication. Understanding the impact of Open Data is often framed in terms of the 'triple bottom line' model: social, environmental and economic impact.[24] Models like the Common Assessment

---

22 Source: 'Benchmarking Open Data automatically'[22] - Figure 3.1.

23 Data SF (2015). How to Measure Open Data. Available at: https://datasf.org/blog/how-to-measure-open-data/.

24 Elkington, J (1994) "Towards the Sustainable Corporation: Win-Win-Win Business Strategies for Sustainable Development," California Management Review 36, no. 2. p90–100.

Framework include impact on the political/governance system as an additional category. Using these four categories of impact - social, environmental, economic and political – we can begin to categorise the types of impact that might be attributed to Open Data.

The types of impact in each of the four categories tend to be broadly defined and overlapping, for example, better public health outcomes could be classified as social, economic or even political impact. In addition, there are a large number of variables that can be measured to describe each type of impact. Open Data presents a number of unique challenges which make it difficult to estimate its value. By its nature, data and particularly Open Data, is non-rivalrous[25], and generates network effects and positive externalities[26]. Studies have shown that accounting for the value of data[27] is particularly difficult given these properties. In addition, Open Data is a relatively recent development. The representative of Belgium's portal noted the parallel difficulties in estimating the value of Open Data with the emergence of widespread use of GPS over the past twenty years. They argued that it would have been exceptionally hard to estimate the value of GPS before its widespread adoption. Impact is also heavily dependent on who is using the data and what they are using it for.

Open Data in economic terms

Open Data is said to be non-rivalrous[28], and generates network effects and positive externalities[29]. These terms are defined as:

Non-rivalry: A non-rivalrous good is one that can be consumed simultaneously by many people at low or no additional cost[30]. Open Data is often viewed as "non-rivalrous" as anyone can access the same data with minimal cost to the provider, once the infrastructure is in place to provide that data. When someone uses this data, it does not stop another person from using it. As the World Bank states, 'the fact that governments (or others) have used the data for the purpose for which it was originally collected does not prevent that data being used for other purposes', either within or beyond government[31].

Network effects: A (positive) network effect is when increased use of the product or service increases its value for other users[32]. Open Data is said to generate positive network effects, in a similar way to open source software, through user input and feedback. For example, the users of Open Data can report issues with the data to the publisher – the more users there are using the data, the more likely it is that issues will be reported quickly. These corrections might not only benefit the users but the publisher themselves, who is likely to be using the data for their own operations.

[25] World Bank (2014). Open Data for Economic Growth. Available at:
http://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf.
[26] Konrad Adenauer Stiftung (2016). Open Data the Benefits. Available at: http://www.kas.de/wf/doc/kas_44906-544-1-30.pdf?160418125028.
[27] Cebr (2013). Valuing data on balance sheet vital for European businesses, economies. Available at:
https://cebr.com/reports/data-on-the-balance-sheet/.
[28] World Bank (2014). Open Data for Economic Growth. Available at:
http://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf.
[29] Konrad Adenauer Stiftung (2016). Open Data the Benefits. Available at: http://www.kas.de/wf/doc/kas_44906-544-1-30.pdf?160418125028.
[30] University of Pittsburg. Available at: http://www.pitt.edu.
[31] World Bank (2014). Open Data for Economic Growth. Available at:
http://www.worldbank.org/content/dam/Worldbank/document/Open-Data-for-Economic-Growth.pdf.
[32] Arun Sundararajan. Available at: http://oz.stern.nyu.edu/io/network.html.

Positive externalities: A positive externality is where 'production or consumption of goods and services [creates benefits for] others which are not reflected in the prices charged for the goods and services being provided'.[33] Open Data is said to create positive externalities from the aggregation of data sources to create innovative products and services.[34]

2.1.2.2. The challenges with measuring the use of Open Data

Understanding who is using Open Data (and for what purposes) can help us better understand the impact it is having on society, the economy and the environment. In addition, measuring use can help portals make strategic and technical decisions, for example by helping to identify preferred means of accessing and formats of datasets, e.g. download vs API, or preferences for types of data.

Tracking Open Data usage is inherently difficult. Even if you know who is accessing your data, it is often very difficult to understand if they are even using it. Several respondents mentioned this problem with reference to downloads of static datasets; users may misread the description or find the dataset to be inaccurate, meaning they do not use it. Conversely, the same dataset might be re-used by the same user for several different visualisations or mashups with only one download. As the representative of Florence's Open Data portal put it "we tried to measure data downloads, but the data download is not enough to measure the extreme variety of data usages. A single data usage that is very useful for a whole school can be more valuable than 10,000 downloads from a bot of a medium-interesting dataset." Likewise, one single download of a dataset which is then shared within a specific company where it leads to the creation of an innovative product will have a higher impact, even if the download only counts as one in the portal analytics statistics.

Even if you could know a dataset is being used, it would be hard to understand how it was being used from the way it was accessed. For example, a dataset might only need to be accessed once to have a big impact – this could be reference data which doesn't change but helps underpin the delivery of a service or it could be historical data which informs an important study. Alternatively, an API might have high volumes of calls which simply reflect poorly developed applications or services. Or a dataset might be rarely accessed because it is only used by a small community, but it might have a big impact on that community.

One method proposed that allows portals to better understand and measure access and use is by tracking who is accessing their data. However, some in the Open Data community have strong objections to this, with one survey respondent claiming it goes against the "philosophy of open" to track who is using Open Data. Other respondents echoed this concern with reference to the idea that Open Data should be accessible anonymously. Nevertheless, anonymity around Open Data is not necessarily guaranteed: it is absent from the popular open definition[35] and many licenses require attribution. These issues are explored further around the issue of data citation.

Those who have suggested automatically tracking access primarily focus on logins or user registration techniques. Most portals who implement these measures allow users to opt-in, often with the incentive of accessing additional features. However, the representative from OpenDataSoft pointed out that most API calls to the portals they hosted which have opt-in logins were still anonymous. In other cases, user registration is a requirement, particularly in the case of APIs which require API keys,

---

[33] OECD externalities. Available at: https://stats.oecd.org/glossary/detail.asp?ID=3215.
[34] Reitano (2013). Externalities. Available at: http://beautifuldata.ca/opendata/?p=349.
[35] Open Definition. Available at: http://opendefinition.org/od/2.1/en/.

for example those made available by the UK Met Office. The representative from Urban Tide said they were exploring requiring API keys for 'high-value', real-time datasets, however this has yet to be implemented.

In addition, some have argued that requiring user registration creates barriers to use. It has also been argued that user registration is not particularly effective for understanding use, given the lack of incentive to provide full and detailed information or to keep that information up to date.[36] Again, tracking data access is compounded by the challenge that access does not necessarily infer use. None of these approaches to measuring use provide an understanding of how data is being used or what impact it might be having.

One of the biggest challenges related to tracking use of data raised by respondents was the challenge of downstream use. Tracking downstream use is particularly challenging for Open Data, given that open licences allow the re-user to redistribute the data to others. This presents a particular difficulty in that even if portals are able to put in measures to track who is accessing data and how they are using it, they would be unlikely to be able to track the re-use of derived data.

A true assessment of the use of Open Data requires an understanding of the ecosystem that is using data not only from the portal but derived data from other sources. Even in the heavily-researched macroeconomic models of Open Data impact this has proved difficult. A UK study found that we may have underestimated the gains from lower prices of Public Sector Information[37] because of the difficulty in valuing the full effects of downstream and future activities. Some existing work has been done to try and map Open Data ecosystems. For instance, recent work done by the ODINE project[38] attempted to map the data value chain through textual analysis of submissions to the project. Mapping ecosystems[39] tends to rely on specialist knowledge and is often very complex.

2.1.2.3. A conceptual model for monitoring use and impact

Both the use and impact of Open Data made available through portals are very challenging to measure. There are a wide variety of current approaches put in place by portals and others to attempt to track them. To fully explore these approaches, their benefits and challenges for portals, we need an underlying understanding of what each is trying to measure and how. For this, we propose a simple conceptual model by which we can understand how data flows through the 'ecosystem' from publisher, through a portal to end users. Within this model we talk about access to Open Data, use of Open Data and impact of Open Data, where they are defined as:

- **Access to Open Data** – the act of retrieving data in a machine-readable format.
- **Use of Open Data** – the act of processing data in order to create information from which decisions can be made.
- **Impact of Open Data** – the outcome of decisions made based on information derived from Open Data.

---

[36] Dodds (2015). How can Open Data publishers monitor usage?. Available at: https://blog.ldodds.com/2015/11/25/how-can-open-data-publishers-monitor-usage/.

[37] Pollock (2008). The Economics of Public Sector Information. Available at: https://rufuspollock.com/papers/economics_of_psi.pdf.

[38] ODINE (2016). Data value chain database. Available at: https://opendataincubator.eu/files/2016/01/D4.6-Data-value-chain-database-Final.pdf.

[39] Mapping ecosystems. Available at: https://blog.ldodds.com/2017/03/13/some-tips-for-open-data-ecosystem-mapping/

Access has been added to help better understand the current approaches. We have specifically developed the model, so it recognises the challenge of downstream access and use of data. The model has been limited to three levels for our current purpose but conceptually could continue for any number of levels. Within our model an intermediary is defined as a person or organisation which accesses Open Data and cleans, enhances or otherwise produces derived data which others can access. The products and services they produce must contain data which can be analysed by a 'user'.
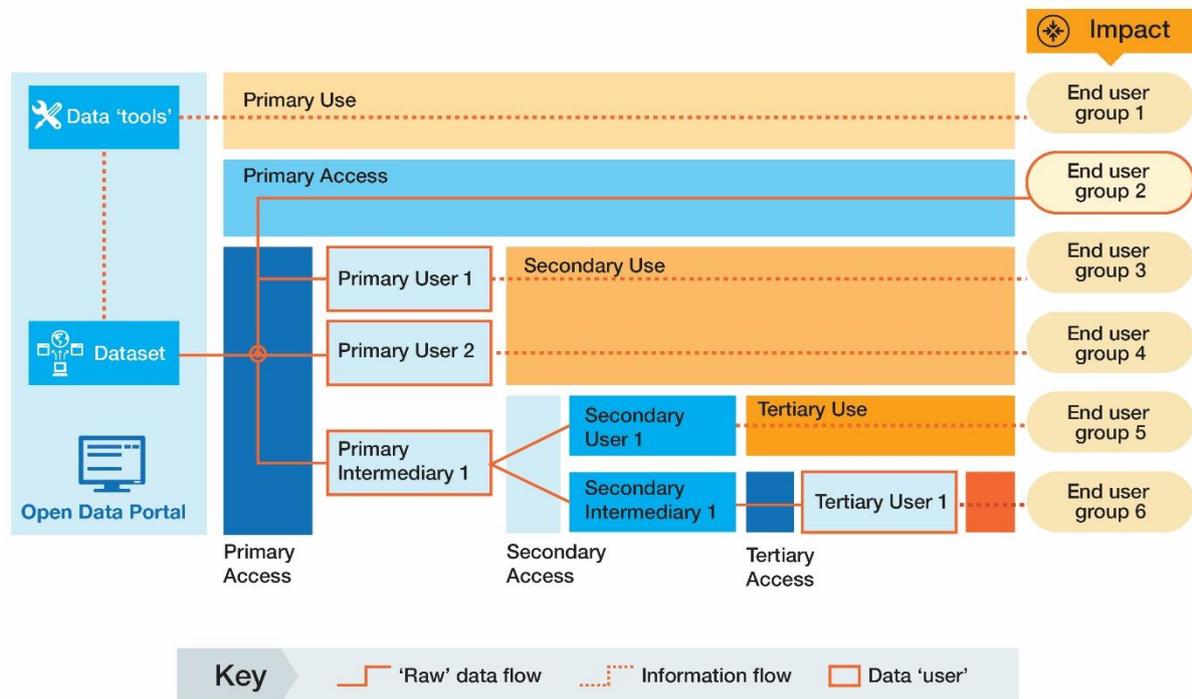
Figure 4 Conceptual model for monitoring access, use and impact of Open Data

The goal of analysing current approaches through this lens is to understand the difference in approaches and how they do or might fit together. We also need to understand the needs of portals when it comes to use and impact. Given the need for portals to be financially sustainable, we will assess the cost and resource effectiveness of each of the approaches. In particular, building on our work in the previous report and the ODI's benchmarking data automatically, we will look at the role automation can play in understanding use and impact. We will also examine the need of being able to connect impact to specific portal activities.

2.2. Current approaches to measuring Open Data impact, use and access

There are many different approaches utilised by portals and others to understand the impact, use and access to Open Data. Through our research we have categorised these into five different types, examining each approach and analysing the potential benefits and limitations of each approach, including examining best practices and lessons for portals. While the five categories vary in one way or another, there are several overlaps between them. They are roughly presented from highest level impact to most basic access measurement, however, often the lines between what they are examining are slightly blurred or they subsume other approaches into their own.

2.2.1. Macroeconomic impact studies

Macroeconomic impact studies are concerned with the performance, structure and behaviour of an economy as a whole. They have long been used to estimate the impact of Open Data and Public Sector Information. These types of studies use a variety of different econometric methodologies in an attempt to quantify the impact of Open Data in economic terms. They tend to identify new or existing examples of the various types of impact either witnessed or expected and use them to create economic models. They then take these models and apply them across a whole economy using existing data on the economy based on sets of underlying assumptions.

For example, in November 2015 the European Data Portal project published a macroeconomic analysis of the impact of the re-use of Open Data for Europe[40]. The study 'Creating value through Open Data' uses four key indicators to quantify the potential size of the Open Data market in the EU28+, being direct market size, number of jobs created, cost savings and efficiency gains. The proposed benefits of Open Data, from existing studies and tangible use cases, were then mapped to each of these areas. Using a wide range of existing data about Open Data, as well as contextual economic data, they put together a complex economic model which could then be applied to each of the EU28+ countries.

The report identifies that between 2016 and 2020, the market size of Open Data is expected to increase by 36.9%, to a value of 75.7 bn EUR in 2020. The forecasted number of direct Open Data jobs in 2016 was 75,000 jobs, with almost 25,000 extra direct jobs created by 2020. The forecasted public sector cost savings for the EU28+ in 2020 are 1.7 bn EUR.

Other examples

Around the world, Open Data is creating economic value at the local, national and international level.

The Omidyar Network have found that Open Data initiatives are contributing **1.1%** to the growth of G20 countries[41]. In 2016, Lateral Economics estimated that shifting to an open access regime would allow G20 countries to add an additional **0.5% to their GDP**.

In Europe, the EDP Creating Value through Open Data report (2015) found that between 2016 to 2020, the **EU28+ Open Data market size** is expected to grow annually by more than **8%** on average. It also found that Open Data could produce value for the public administration sector of up to **22.1 mill EUR**. Other studies have found that Open Data provides an estimated indirect economic benefit of 1.7% of the EU's GDP, and that tax revenue gained from Open Data activities in Europe amounts to 140 bn EUR annually.

A 2016 study in Denmark found that Danish society could gain between DKK 50-135 million (EUR 67-181 mill) annually in selected sectors, including electricity, district heating and agriculture, as a result of open meteorological data.

---

[40] European Commission (2015). Creating Value through Open Data. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.
[41] Omidyar Network (2014). Open for Business. Available at:
https://www.omidyar.com/sites/default/files/file_archive/insights/ON%20Report_061114_FNL.pdf.

2.2.1.1. Benefits and limitations

Macroeconomic models, which describe the behaviour of an entire economy, are the standard methodology employed to calculate the relative impact of policies and technologies in economies around the world. They are familiar to policymakers and civil servants, and allow comparisons between Open Data and a wide range of other technologies. Because of their ubiquity, the significant expertise that goes into them and the fact they have been around for a long period of time, they are generally viewed as being credible. This is particularly the case for studies produced by reputable firms.

While the value estimates differ – primarily due to when they were carried out, the methodological approaches use and data sources – there appears to be emerging consensus among the macroeconomic studies on the value that can be or is being unlocked by the release of Open Data, particularly in Europe. However, while each of these can be used by portal owners to help evidence the impact of Open Data, there are limitations to their usefulness in demonstrating the impact of individual portals.

- First, these studies tend to focus on a geographical or economic scale – global, supranational or national scale. They tend not to be associated with specific releases of data. As such, respondents noted the difficulty in using these figures to evidence the impact of their own portal.
- Second, there are large costs associated with carrying out these types of assessment, given they often require large scale data collection and interpretation by highly trained economic experts. Having the budget or resources to carry them out is typically beyond portal owners.
- Third, they only provide a snapshot in time and potentially some estimates for future growth. They are rarely followed up and so it is difficult to use them to measure actual performance of improvements.
- Fourth, while these studies often provide methodologies, they are not always fully transparent or accessible to their audience. By virtue of the expertise that goes into them the methodologies are often obscure to the user and this can be seen to damage the credibility of these high-level macroeconomic figures. One respondent argued that it was unclear where the different numbers came from and without access to the explicit methodologies and workings, it was difficult to use them persuasively.

2.2.1.2. Lessons and best practices

There are clearly benefits for portals in being able to refer to macroeconomic impact studies and they are clearly useful to other audiences, such as policymakers. The high-level figures can help justify the initiation of Open Data initiatives and portals, and provide momentum, but they are not as useful for measuring the performance of portals. By measuring whole economies, it is hard to demonstrate the direct impact of a particular portal, publisher or dataset.

If portals are able to carry out or commission such studies, it is important that they attempt to make them as repeatable and replicable as possible. Efforts should be made to provide clear methodologies which explain data collection, explicit calculations and assumptions. for example, both the EDP and Lateral Economics reports provide detailed methodologies around the calculations and approaches taken. Studies should also publish openly the code or tools used to calculate and underlying data on which the calculations are made. By doing this, portals should help reduce the cost of future studies, for themselves and others, and improve their ability to easily update the findings over time.

Recommendations for portals: Macroeconomic impact studies

If portals are able to carry out or commission macroeconomic impact studies, they should:

- **provide clear, detailed and repeatable methodologies,** which explain data collection, calculations and assumptions.
  - o Best practice examples: EDP and Lateral Economics
- **publish underlying data and tools,** allowing these calculations to be repeated.

If portals are unable to carry out or commission their own macroeconomic impact studies, they should:

- **examine how existing figures can be made relevant** to their context
- consider **partnering with publishers and other government agencies** to commission umbrella studies of all their activities

2.2.2. Microeconomic impact studies

Microeconomic impact studies adopt similar econometric methodologies to the macroeconomic studies, however they focus on specific publishers or datasets. They tend to be commissioned by data publishers themselves, or their funders, and carried out by external economic experts. They can also focus on specific sectors or types of data as a whole.

The value of open transport data: TfL in the UK

Transport for London (TfL) has been publishing Open Data about its services since 2010. In 2012, a microeconomic study commissioned by the Department for Business, Innovation and Skills (BIS) included a case study that estimated the value of the Open Data release[42] between £15m and £58m in 2012.

In 2017, TfL itself commissioned another review of the value of the open transport data[43] it had released. To calculate the value of the release they built a model around three groups who are identified as benefiting from the release: TfL passengers and other road users, London as a city and TfL itself. For each of these they identified ways they benefit and quantified the impact of the release of data, for example, customers benefit from more efficient journeys, London benefits from the companies reusing data and TfL benefits from savings associated with not developing the services in house.

To calculate the exact benefits, they used data generated from TfL's developer surveys which are carried out every six months to understand the apps that have been built using their data, currently standing at 600. This is primarily used to help understand the needs of users and drive the policy and technical developments of their platform. TfL also conducts a regular survey of TfL customers and citizens to understand how Londoners use these applications and the tools built by TfL themselves, most recently finding 42% of customers use one or more of the 600 apps built on TfL Open Data. Using this time-series data collection makes it easier to estimate overall impact when carrying out a specific study.

---

[42] Department for Business Innovation & Skills (2013). Market Assessment of Public Sector Information. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198905/bis-13-743-market-assessment-of-public-sector-information.pdf.

[43] Deloitte (2017). Assessing the value of TfL's Open Data and digital partnership. Available at: http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf.

The study revealed that the total value across all these impacts is estimated to be worth up to £130m per year[44]. The gross value add (GVA) from using TfL data for individual companies alone is between £12m and £15m annually.

Other examples

A 2015 study found that in Colombia, Open Data agricultural products were generating USD$3.6 million[45].

A 2015 study found that by opening up public transport data in San Francisco in the USA, the city was able to reduce the number of telephone queries and save more than USD 1 million[46].

In a 2014 study, Socrata estimated that the total value of products and services based on open weather data was USD$15 billion[47].

2.2.2.1.   Benefits and limitations

These econometric studies focused on particular data publishers or data publications have clear value to those who commission them. They often provide high level impact numbers based on relatively rigorous methodologies. They tend to build on and provide explicit examples of how the data is being used, as well as taking into account access figures. While these types of studies are very useful for portals and publishers, they have some of the key limitations of macroeconomic impact studies.

- First, costs and resourcing is seen as prohibitive to many portal owners. One survey respondent noted that "the biggest problem is that we would need more resources in order to accomplish this task". While there may be ways to conduct these studies internally another noted the lack of internal capability to deliver such work.
- Second, like macroeconomic studies they only provide a snapshot of impact at any given time.
- Third, a main concern with these studies relates to their re-usability and applicability beyond the publisher or dataset studied. One respondent noted that while there are groups of data which have estimated or proven value, it was difficult to use these figures to evidence and draw parallels with other datasets, even if the datasets are similar. Part of this is due to the differences in context, making it difficult to relate the figures calculated to other projects.

2.2.2.2.   Lessons and best practices

There are clear benefits to commissioning microeconomic impact studies for those with the resources to do so. However, these may be out of reach of many portals, especially if they want to track portal performance over time. Where portals are able to commission studies, similar to macroeconomic studies, the focus should be on developing replicable methodologies and publishing the underlying data openly. Portals and publishers who have already commissioned studies or have access might

---

[44] Deloitte (2017). Assessing the value of TfL's Open Data and digital partnership. Available at: http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf.
[45] GODAN (2015). New research shows the impact of Open Data in agriculture and nutrition. Available at: http://www.godan.info/fr/news/new-research-shows-impact-open-data-agriculture-and-nutrition.
[46] Ministry of Finance (2013). The Impact of Open Data – a preliminary study. Available at: https://www.w3.org/2013/share-psi/wiki/images/6/67/Impact_of_Open_Data_in_the_Public_Sector_Koski_2015.pdf.
[47] Cashman (2014). The Economic Impact of Open Data. Available at: https://socrata.com/blog/economic-impact-open-data/.

consider sharing, and if possible publishing, detailed methodologies of past studies to enhance reproducibility.

For the majority of portals who are unable to commission studies, efforts should be made to study the methodologies and determine how and if the figures can be made relevant to their own context. Portals might consider co-commissioning the development of open, replicable methodologies.

Recommendations for portals: Microeconomic impact studies

If portals are able to carry out or commission microeconomic impact studies, they should likewise:
- **provide clear, detailed and repeatable methodologies,** which explain data collection, calculations and assumptions.
- **publish underlying data and tools,** allowing these calculations to be repeated.

If portals are unable to carry out or commission their own microeconomic impact studies, they should:
- **examine the methodologies of existing studies to determine if figures can be made relevant to their contexts**
- **consider co-commissioning the development of open replicable methodologies**

2.2.3. Business population studies and user surveys

Business population studies and user surveys are used by portals and others to both identify use and understand impact. These types of studies attempt to better understand the ecosystem of Open Data users, either direct or downstream, and understand their overall impact. These studies differ from the microeconomic studies mentioned as they tend to focus less on specific datasets, publishers or sectors. In addition, they tend to focus on making a holistic assessment of the overall ecosystem – the nature of the businesses, and other types of users – including what types of data are most used and how that data is being used. They are also often less focused on quantifying the direct and indirect impact through econometric techniques, although many calculate a figure of overall direct business impact. Methodologically, they tend to identify users through publicly available sources and self-identification, usually using surveys and other data sources, such as company registers, to improve the self-reported data.

In 2017, the International Data Corporation (IDC) released an impact assessment of the European ODINE programme. It estimated that the combined revenue of these Open Data companies would contribute an estimated EUR 110 million between 2016-2020, in addition to creating 784 jobs. Importantly, the study indicated a positive relationship between a country's level of Open Data maturity and the number of successful ODINE applicants from that country. This demonstrates that producing impact and innovation with Open Data requires the creation of a strong data ecosystem. This study was enhanced by research done by the Open Data Institute and Fraunhofer on the business models of the same ODINE companies which looked at the types of data being used and published by the start-ups and how this played into their products and services.[48]

Of the portals included in the study, three of our respondents – Spain, the UK and Florence – had directly employed some type of survey or business population study. The most prominent example of

---

[48] ODINE (2016). Business Models, Lessons Learned, and Success Stories. Available at:
https://opendataincubator.eu/files/2016/01/D6.3-Final.pdf

these are the regular studies of the Spanish 'Infomediary Sector'[49]. These are produced every two years, starting in 2012[50] (actual report[51]) with two further reports in 2014[52] (actual report[53]) and 2016[54] (actual report[55]). They use desk research, outreach and surveys to put together a comprehensive understanding of businesses using Public Sector Information. The most recent study identifies 535 infomediary businesses and gathers data on their economic output, number of employees and sectors. They go into great detail on the usage of Open Data, not only from the national portal but regional, local and international sources, as well as private sector and academic data.

The comprehensive figures are used not only to drive policy and technology developments within government but also highlight best practices in the sector itself, aiming to help grow the overall size of the sector. The regularity means they have a historic record through which they can trace the growth of the sector and compare this to the work they have done. In addition, the overall impact of these businesses found they were generating EUR 1.7bn in revenue, with average turnover of EUR 2.68 mn per company[56]. This was an increase from EUR 330-550 million in 2012.

Other surveys have focused more narrowly on direct users of data platforms, for example the Florence Open Data Portal has recently conducted two surveys supported by a university, to understand how students and technicians use data from the platform. The goal is to understand the usefulness of Open Data usage among students and help guide the provision of data through understanding their needs. They are also trying to understand how the data is being used. Data.gov.uk, the UK's national Open Data Portal, has also carried out a user survey on the general usage of the site to help guide development but expect it will also lead to examples of how and why the data is being used.

> **Other examples**
>
> Many initiatives and studies have sought to understand Open Data's impact through mapping the ecosystem of re-use.
>
> The ODI's 2015 'Open Data means business[57]' study identified and analysed 270 companies that use, produce or invest in Open Data as part of their business, through desk research, surveys and interviews about their experiences. They found that these companies have an annual turnover of over £92bn, and over 500k employees between them, indicating the scale of Open Data's potential

[49] Red.es. Available at: http://www.ontsi.red.es/ontsi/es/estudios-informes?title=infomediario&field_mes_informe_value=All&field_anyo_informe_value=All&field_estudios_informes_tid=All.

[50] ontsi.red.es. Available at: http://www.ontsi.red.es/ontsi/es/estudios-informes/estudio-de-caracterización-del-sector-infomediario-en-españ-edición-2012.

[51] Datos.gob.es (2012). Characterisation Study of the Infomediary Sector. Available at: http://www.ontsi.red.es/ontsi/sites/ontsi/files/121001_red_007_final_report_2012_edition__vf_en_1.pdf.

[52] ontsi.red.es. Available at: http://www.ontsi.red.es/ontsi/es/estudios-informes/estudio-de-caracterización-del-sector-infomediario-en-españ-2014-información-del-0.

[53] Datos.gob.es (2014). Characterisation Study of the Spanish Infomediary Sector, Private Information Sector. Available at: http://www.ontsi.red.es/ontsi/sites/ontsi/files/ppt_private_infomediary_sector.pdf.

[54] ontsi.red.es. Available at: http://www.ontsi.red.es/ontsi/es/content/estudio-de-caracterización-del-sector-infomediario-2016.

[55] Datos.gob.es (2016). Characterisation Study of the Spanish Infomediary Sector. Available at: http://www.ontsi.red.es/ontsi/sites/ontsi/files/Presentation%20Characterization%20of%20the%20Spanish%20Infomediary%20Sector.%20Ed.2016_0.pdf.

[56] European Commission (2017). Open Data Maturity in Europe 2017. Available at: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

[57] The ODI (2015). Open Data Means Business. Available at: https://theodi.org/open-data-means-business.

value in business. As well as providing detailed case studies and a breakdown of the types of data being used.

Italy's Open Data Italy 200[58], was the first comprehensive study of Italian companies using Open Data to create economic and social value. The results highlighted that geospatial data is the most frequently used by Italian businesses. The research study also highlighted the innovative ways in which Open Data is being used across the country, from the Lombardy region to the City of Florence.

2.2.3.1. Benefits and limitations

Using large surveys and desk research, business population surveys provide a good overview of the ecosystems of re-users. Such surveys can be useful for portals to help understand their own ecosystem of users, including some downstream users. They can help to identify demand for different data types and datasets, how users are using Open Data and what types of businesses are involved. Representatives of Spain's portal highlighted the usefulness of conducting their study for guiding policy decisions and encouraging publishers. However, there are several limitations in their usefulness for measuring Open Data portal use and impact.

- First, as one respondent identified, such studies tend to be highly resource intensive given the difficulty in identifying and reaching out to potential re-users.
- Second, while they are able to capture some downstream re-users, they are often based on publicly available documents and self-reporting through surveys to identify use. Primary users of Open Data might not wish to report their use for competitive reasons and many downstream re-users may not be aware of their re-use of Open Data. This means that they are likely to underestimate the total size of the ecosystem.
- Third, these studies tend to only provide a snapshot of the ecosystem at a certain time. While several are conducted on an ongoing basis, allowing for comparisons over time, these intervals are normally quite large given the effort required not only by the researchers conducting the study but also by the users who have to respond.

2.2.3.2. Lessons and best practices

These types of studies are arguably some of the most comprehensive measures of use and impact being carried out. Portals can adopt these survey-based methodologies to help them understand not only their direct users but also downstream users. However, these studies are currently very resource intensive. To lower the burden on portals themselves, they can partner with other relevant organisations, such as universities, as the Florence portal does.

The methodologies vary - making them hard for some portals to replicate and not allowing for comparisons. Efforts should also be made by portals to identify standard sets of topics and questions, as well as methodologies which can be applied across different portals, to make them comparable and easier to carry out. Basing these methods on existing best practices such as the Spanish Infomediary research – which is regular and in depth - would be beneficial. These survey questions might be focused on those already asked in the studies in this section. They might also want to focus on questions of how and if companies sell or publish data that means it is used downstream - and if so, how these companies can be contacted to also take part.

---

[58] Open Data 500. Available at: http://www.opendata500.com.

In addition, where possible, portals should identify existing open or shared data sources which can aid in the analysis, such as company registers and economic statistics. Once surveys are carried out, the underlying data, suitably anonymised, should be published as Open Data on the portal, as was done with the Open Data means business study. In addition, any best practices around the identification of businesses and users should be published to help others. Finally, efforts should be made to automatically collect the data more often and ideally continuously, by making the survey available online permanently. Where possible, create mechanisms that allow for ongoing analysis which can help portals make decisions and understand their impact on a more regular basis.

Recommendations for portals: Business population studies and user surveys

- ◉ **Portals should seek to partner with other organisations,** such as universities and government departments who are also interested in the outcomes.
  - ○ Best practice example: Florence data portal[59]
- ◉ **Portals should look at existing studies and pose consistent questions,** efforts should be made over time to standardise the types of questions asked.
  - ○ Best practice example: Spanish infomediary study[60]
- ◉ **Once surveys are carried out, the underlying data, suitably anonymised, should be published as Open Data on the portal.**
  - ○ Best practice example: Open Data means business study[61]
- ◉ **Where possible, efforts should be made to automatically collect and analyse the data on an ongoing basis,** potentially by making the survey available online permanently and producing dashboards.

Direct vs indirect economic impact

Open Data can produce both direct and indirect economic benefits. Most of the approaches we have examined focus exclusively on direct economic benefits while not capturing the wider indirect benefits.

Direct benefits are benefits that are realised in market transactions in the form of revenues, Gross Value Added (GVA), number of jobs and cost savings[62]. For Open Data, these direct benefits are usually quantified for the companies who build products and services based on Open Data, for example the Spanish infomediary study or the EDP Open Data maturity studies both quantify the market for products and services based on Open Data.

However, there are many **indirect benefits** to the economy that are not captured in these terms. Indirect economic benefits may include new goods and services, time savings, increased efficiency in public services and growth of related markets. Often, these benefits are experienced further

---

[59] Florence Open Data Portal. Available at: http://opendata.comune.fi.it.
[60] 5th Edition of the Report on the infomediary sector (2017), Spain. Available at: http://www.asedie.es/assets/informe-sector-infomediario--2017.pdf.
[61] Open Data means business (2015). Available at: https://theodi.org/open-data-means-business.
[62] European Commission (2017). Open Data Maturity in Europe 2017. Available at: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

along the value chain or in populations not directly related to the organisation using Open Data[63]. These benefits are also often harder to quantify in economic terms because it is more difficult to associate outcomes explicitly with data release. Indirect economic impact may however be captured through both macroeconomic and microeconomic studies, as well as individual case studies.

For instance, the 2017 TfL study also captured that time saved for network passengers as a result of the greater certainty on when the next tube or bus would arrive. This saved time estimated at between £70m and £90 million per annum, which is likely to lead to higher productivity[64]. Likewise, the indirect impact of using US Landsat data, which provides high quality impacts of the Earth's surface, helped producers in Chile identify the right amount of water to use on their olive orchard and vineyard crops. This helped provide an $80 per acre cost saving in the energy used for irrigation on over 3,700 orchards each year[65].

Lastly, Open Data may have indirect economic benefits by creating a more inclusive economy for marginalised groups. For instance, BlindSquare[66] in Finland is an accessible GPS application which helps the blind and visually impaired travel through the city. The app is using Open Data to make the city more accessible. Likewise, SharePA[67] in France is helping people with reduced mobility understand which public services (such as parks and museums) are accessible, in order to help them plan their journey. Both these apps provide a dual benefit of social inclusion and indirect economic benefits by making it more likely that those with impairments will be able to engage with the city and its services[68].

2.2.4. Showcases and use cases

Use cases and case studies are one of the most used methods in demonstrating the use of Open Data and also its impact or potential impact. According to the 2017 EDP Open Data Maturity Report, 71% of the EU28 countries provide a section on their portals collecting use cases. From our survey respondents, only 58% explicitly mentioned collecting use cases and case studies as a key measure they employed, including France, Russia, the UK and Luxembourg's national portals and the Berlin, Florence, Trentino and Helsinki's local portals.

Portals collect these use cases in a variety of different ways for a variety of different reasons, but they tend to display these use cases in a 'showcase' on the portal, making them the most popular form of tracking use and impact. Given their popularity, all the portal providers we spoke to provided a means by which to create showcases as part of their standard offerings. For example, the popular open source data portal platform CKAN has a "showcase" plugin that allows owners to register visualizations, blog posts, journal articles and papers that use datasets in the portal. Other hosted solutions, such as OpenDataSoft, are flexible in allowing portal owners to display use cases on their site in a manner that suits them.

---

[63] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

[64] Deloitte (2017). Assessing the value of TfL's Open Data and digital partnership. Available at:
http://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf.

[65] USGS (2017). Vineyards and Apple and Olive Orchards, Chile. Available at: https://www.fort.usgs.gov/sites/landsat-imagery-unique-resource/case-studies/vineyards-and-apple-and-olive-orchards-chile.

[66] BlindSquare, Finland. Available at: http://www.blindsquare.com.

[67] SharePA, France. Available at: http://lexpress.github.io/SharePA/index.html.

[68] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

Showcases allow portals to provide specific examples that show how their data is being used and by whom. A number of portal providers and owners described the ability to link use cases to particular datasets as very important. The UK's national portal uses these links to generate reports for publishers about how their data is being used. This approach has led some UK government departments who are publishers to use the platform's reports about their data as their own internal metrics and evidence for publishing Open Data. A number of respondents also highlighted the usefulness of case studies and showcases in encouraging further re-use. The Urban Tide platform encourages publishers to produce visualisations or analysis alongside the publication of new datasets to 'whet the appetite' of re-users.

Gathering case studies for 'showcases' is one of the key challenges for portals trying to demonstrate the types of use to portals, publishers and re-users. Almost all the portals we surveyed and interviewed allowed data users to submit case studies about how they are using the data. This typically involves a short description of the application, a link and some standard set of attributes, for example the dataset(s) being used or the categories or sectors it involves. One example of this is the 1680 re-uses on data.gouv.fr[69] organised by type (visualisation, application, articles etc.), organisation and theme. The showcase also allows users to 'follow' re-uses they like and receive updates on them. The data.gov.uk[70] platform takes a similar approach and has provided a facility for submitting apps that use Open Data since 2012. Again, this is linked to the datasets used in the app and classified by sector.

While this approach appears to have worked to an extent, one respondent noted that relying on re-users to report their use may be 'wishful thinking', due to a lack of incentive for users to share their innovations in some cases. In an attempt to overcome this hurdle, the representative from AWS noted the importance of providing an incentive in the form of publicity for re-users. They explained that they were relatively successful in getting reports of use by putting effort into publicising individual use cases. However, this tactic might be difficult for portals with less resources.

A number of respondents to the survey explicitly emphasised the importance of reaching out and maintaining dialogue with the community of users – including representatives of Trentino, Luxembourg, Belgium and Slovenia. Some mentioned that it was often through existing organisations, communities and events that they discovered the best use cases and persuaded users to submit. The representatives of Luxembourg and Belgium in turn mentioned that specific existing communities were particularly useful – for example the OpenStreetMap and Open Knowledge communities. While this can be resource intensive, a number of respondents highlighted other ways in which they encouraged re-users to submit cases during other interactions - for instance, the representative of the Trentino portal said that when re-users approached them with questions about the Open Data they host.

This approach has been implemented in some portals by building the reporting of re-use into events hosted by the portal owners. For example, the annual gameofcode.eu[71] hackathon in Luxembourg requires participants to publish their re-use on the Luxembourg portal as part of the challenge. By requiring that teams also produce something that is 'coded' as well as reporting their use, they have managed to generate significantly more use cases for their showcase. The representative of Luxembourg's portal also noted that more people have begun to submit re-use cases since the event took place, indicating that this method increases the visibility of and provides an incentive to publish the showcase.

---

[69] National Open Data Portal France. Available at: https://www.data.gouv.fr/fr/re-uses/.
[70] National Open Data Portal UK. Available at: https://data.gov.uk/apps.
[71] Game of Code. Available at: http://www.gameofcode.eu.

Another method that is used to generate use cases is working with universities, several respondents noting that they either had or were planning to work with universities to generate and record use cases from students and research projects. For example, representatives from Helsinki and Florence both highlighted the importance of connections with local universities, and the representative of Belgium noted how they might seek to engage. In Helsinki, students on a specific university course were required to use Open Data from the platform. Write ups of these projects and any relevant materials could then be linked or hosted on the portal itself.

> **Other examples**
>
> Collections of case studies like those included on the EDP[72], the ODI website[73] and Govlab's Open Data impact page[74] illustrate a number of Open Data uses by governments, businesses and civil society across the world. For instance, Open Data is being used in the Netherlands[75] to help citizens and journalists keep up to date with political developments; likewise, Greek company Flex is using Open Data[76] to support people with dementia in completing everyday activities.
>
> Similarly, the Open Contracting Partnerships also collect use cases that show the specific impact of open contracting data. In the Ukraine[77], implementing an open contracting system that made data about government contracts and tenders available led to savings of EUR 1.2 million in the public budget in the first three months. In Slovakia, Transparency International used contracting data[78] to discover efficiency gaps in hospital procurement, finding that up to 50% of tenders had only one competitor. The diversity of use cases on the platform provide links to many projects and organisations active in the Open Data space.

2.2.4.1.   Benefits and limitations

Showcases and use cases are generally considered very valuable for describing use and impact. The representative from Belgium's portal highlighted their importance to politicians and publishers who are sometimes more interested in the tangible examples of re-use than the less specific overall impact figures. These sentiments were echoed by the representative of AWS, who mentioned that use cases are the most meaningful evidence currently available in helping people to talk about and understand Open Data in a tangible way. The representative of Urban Tide also emphasised their value as 'softer evidence' that provides context for Open Data published on the platform.

Showcases also tend to be very technically easy for a platform to implement, with almost all solutions providing functionality. Unlike one-off studies, showcases involve ongoing collection of data about re-use, meaning they tend to grow organically over time, especially where re-users can submit their own.

---

[72] European Data Portal. Available at: https://www.europeandataportal.eu/en/using-data/use-cases.
[73] The ODI. Available at: https://theodi.org/case-studies.
[74] Open Data's Impact. Available at: http://odimpact.org.
[75] European Data portal (2017). 1848. Available at:
https://www.europeandataportal.eu/sites/default/files/use_case_netherlands_-_1848.pdf.
[76] European Data portal (2017). Flex - Mobile Dementia Assistant. Available at:
https://www.europeandataportal.eu/sites/default/files/use_case_greece_-_flex_mobile_dementia_assistant.pdf.
[77] Open Contracting Partnership. Available at: https://www.open-contracting.org/why-open-contracting/showcase-projects/ukraine/.
[78] Marchessault (2015). Deals and dollars – The impact of open contracting data. Available at: https://www.open-contracting.org/2015/07/23/the_impact_of_open_contracting_data/.

However, there are several limitations in their utility for measuring Open Data portal use and impact:

- First, building large showcases can require a significant amount of outreach and engagement activity to encourage re-users to submit (although they are arguably one of the least resource-intensive options).

- Second, while the showcase as a whole may grow over time, individual use cases tend not to be updated. This can have a significant impact, as readers do not know whether innovations are still in use, or if any additional datasets have been included since the original use case was written, for example. Use cases also often lack a consistent structure or explicitly quantitative account of impact which can make them difficult to compare to one another, or to provide an overall understanding of the showcased use or impact.

- Third, they often rely on self-reporting which can limit them to direct users of the platform and often means they range in scope and detail. While they tend to be accurate, given the incentives in place, they are often not independently verified which might mean they contain more bias than other independent research approaches. A number of respondents emphasised the reliance on self-reporting as an inherently social and cultural challenge – including representatives from Spain's portal, OpenDataSoft and TfL. For example, several respondents noted that companies might not be willing or able to report their use, for fear of competition or existing legal agreements. There might be a greater disposition amongst Open Data advocates, freelance journalists or social-interest groups to report because of a lack of these factors, meaning that these showcases will tend to oversample them over potentially higher impact uses by large companies or emerging start-ups.

2.2.4.2. Lessons and best practices

Showcases are very useful for portals and many are already using them to evidence the use and impact of data on their platforms. There are a variety of lessons to be learned from the methods used to collect these use cases.

Although providing users with the ability to self-report uses clearly has an advantage for lowering the burden on portal owners, it is generally not enough to guarantee that users will take advantage. To encourage participation, portal owners should make efforts to engage diverse existing communities to encourage them to publish through working with existing interest communities who are already actively engaged, as well as embedding requests for submissions within their existing channels and outreach work as done by Trentino. Requiring hackathons and data challenge participants to record their innovations in order to be judged as is done in Luxembourg is an excellent example, as this provides a strong incentive to engage with the portal. Other incentives could be formed, such as explicitly promoting use cases in a way that helps re-users - portals and platforms could promote start-up use cases to venture capitalists. Incentives should also be created to encourage curation of use cases, including regular updates and follow-ups.

Linking case studies to the specific datasets that are used, like on the French or the UK's portal, is particularly important. This provides a good opportunity to evidence the utility and value of a particular dataset to other re-users and to publishers. Efforts could be made to collect more structured data around impact and the users themselves when submitting case studies - for example, the number of users of an application based on Open Data. Portals should also try to adopt standardised approaches to use cases, allowing them to compare similar uses from across different portals. This also provides the potential for users who use multiple data sources on different portals to report them easily to all portals they use.

> **Recommendations for portals: Showcases and use cases**
>
> - **Portal showcases should allow users to submit their own re-uses.**
> - **Portals should also encourage reporting of re-use through community engagement**, including partnering with universities and requiring hackathon entrants to submit responses.
>   - Best practice examples: Luxembourg, Belgium, Trentino, Florence, Helsinki.
> - **Portals should follow up with Showcase re-users on a regular basis,** asking them to update their cases.
> - **Showcases and use cases should be linked to the specific datasets that are used**.
>   - Best practice examples: France, the UK
> - **Showcases should focus on collecting more structured data and identify contextual data sources that could be linked**, standardising approaches should also be explored.

2.2.5. Automated access metrics

Data is accessed through a portal when it is viewed, downloaded or accessed through an API. While access to data does not necessarily provide the best approximation of actual use, it can still be insightful - access metrics often play a significant role in technology and policy decisions as they indicate the use of portals themselves. A core advantage of access metrics is that they are often automated when it comes to data collection, which makes it easier for portals to implement. When asked whether they track Open Data use, 47% of respondents explicitly mentioned techniques associated with automatically monitoring access, and 21% explicitly mentioned more than one method. The methods identified by respondents included monitoring page analytics, dataset downloads and API calls.

2.2.5.1. Page analytics

21% of our survey's respondents explicitly mentioned that they were tracking page analytics – including Trentino, Helsinki, Luxembourg and Spain. Given the availability and low cost of production-ready, user-friendly tools like Piwik or Google Analytics, the number of portals using these methods is likely much higher. These tools allow websites to track visitor behaviour and answer questions, such as: "what are the most seen pages?", "what are the pages that are frequently seen during the same session?", "the most shared?" and "what are visitors searching for in the search box?". They are used in pages across the web to understand the performance of particular sites and user journeys. In particular, digital content distributors and e-commerce sites have come to rely on the data generated by these tools to derive valuable insight about how users react to their pieces and products.

Open Data portals can leverage the same tools to estimate access to, and demand for the datasets they publish. In almost all Open Data portals, each dataset has its own page with a description and a link to download it. An analytics tool like Piwik could be configured, with minimal disruption to the portal infrastructure, to automatically track and report:
- The number of views that the page of a dataset has
- The approximate location of the viewers of the dataset page
- What other pages the viewer has visited
- How a viewer got to the dataset's page

This can help portal owners understand the relative interest in, or visibility of, a dataset when compared to other datasets on the platform.

Understanding how users navigate to a dataset can provide useful information for designing website navigation and inferring motivations. Like many other websites, Open Data portals tend to have a search box function where users can search for datasets they are looking for. Analytics, such as what query's results include the dataset and the number of times a dataset appears in a user's search result page, can provide useful information on the demand for Open Data. For example, a dataset that does not appear often as a result of a search might indicate that the demand for it is low, whereas a dataset that appears often as a result of a query but is not viewed might indicate either a problem with the search functionality (the dataset is irrelevant to the query), or that the dataset is less relevant than the other results returned.

2.2.5.2.   Downloads

Moreover, beyond dataset page views and user journeys, 37% of respondents also explicitly mentioned monitoring the number of datasets downloaded – including Helsinki, Florence, the UK, Russia, Bath, Spain, Paris. Portal owners interviewed highlighted that this was a relatively easy metric to automate, meaning the number of portals using this method is likely to be higher than explicitly mentioned. However, dataset downloads can be harder to track as standard web tools might not be set up to monitor it, meaning more effort is required for implementation. The representative of Florence's portal highlighted that their current infrastructure required them to develop new tools to measure access through downloads, which they chose not to do until they launched a new site which did this in an automated and integrated way.

Every time a dataset is downloaded, it is potentially being re-used. However, contrary to the case of e-commerce sites, where the "checkout" action can be always equated to the purchase of a product, a "download" does not always infer re-use. In addition, download patterns could also be used to more strongly infer use. For example, if a dataset has a high number of downloads each time a new addition is published, it might indicate that it is being re-used by a number of people. If users often download all historic datasets it might indicate they are more likely to be using them than if they download just a single dataset. While it is very difficult to make these assertions without additional evidence, portals might use these patterns to help them understand access.

2.2.5.3.   API metrics

While static bulk downloads provide a single record of access, datasets exposed through APIs offer the advantage that more information about their usage can be inferred. Of respondents to our survey, 16% explicitly indicated that they tracked API calls – including Luxembourg, Paris and Spain. The number of API keys requested, if they are required, can be used as an overall metric similar to downloads for static datasets, indicating that someone is interested in querying the dataset. For example, the respondent from TfL mentioned that they use the number of registered unique users as a KPI to indicate level of interest.

However, as discussed, there are barriers created by putting logins or requirements to use API keys in place. The respondent from OpenDataSoft highlighted that they had optional API keys and the majority of their users chose to access their APIs anonymously. In cases where API keys are not tracked, other unique identifiers may be logged which are not used to identify the user but can still identify their behaviour. The representative from AWS noted that they were using IP addresses as the unique ID for each session, for instance, but not actually asking users to authenticate or identify themselves. This allows portals to strike a compromise between barriers to logins while also gaining some of the benefits of tracking access.

In addition to recording unique visitors, a portal can record API call logs to gain further insight into dataset access and usage. As it is the case with web analytics tools, API log and analysis suites are readily available, due to their use for improving performance and security in commercial environments. For the purpose of tracking dataset usage, a log needs to be configured to record in a per-dataset way:

1. The method called and the parameters used
2. The return of the call
3. The timestamp of the access

Using this method respondents have tracked the number of queries, the accessing map and usage over time. Each of these gives an overview of access to particular datasets which can be used to create some indication of use. In particular, patterns of usage such as regular spikes in access could indicate particular applications and services that are using the data. Primarily these logs tend to be used to compare between datasets - identifying those that are most popular, several respondents highlighted these comparisons being used to understand the importance of different datasets although they cautioned that they should not be used in isolation as they remain relatively poor in indicating actual use – including representatives from the UK and Russia about publishers on their national platforms, and OpenDataSoft about their city portal clients.

By counting how often a dataset is accessed by a certain unique ID, one can gain greater insight on how the dataset is being used. For example, several different inferences become possible, such as the number of users that access a dataset through the number of unique IDs that access it. Other inferences include:

● A single access through a single method might indicate that the person behind the API key only wanted to check the particular fact answered by the method called. Several other API keys checking the same fact indicate that a particular data point is the most important piece of the dataset. On the other hand, single accesses to the method that returns the whole dataset might indicate that the dataset needs to be analysed as a whole to make sense out of it, or that is part of a complex data analysis process

● Multiple accesses from the same key to the same dataset might indicate that the dataset is being used for an application. Regular patterns are indicative that the application is synchronizing with the dataset at regular intervals, while irregular patterns indicate that the app uses the dataset on demand, perhaps after a user command in the app.

● The same key accessing different datasets might indicate that those datasets are being used in conjunction for an app or for analysis. By analysing the activity of API keys, it is possible to infer how the user behind the key is joining datasets.

This approach is generalizable to any dataset exposed through a query interface for which queries and an identifier like an API key can be logged, like web services, relational and SPARQL endpoints. By providing query interfaces to datasets beyond only downloading, logging and analysing the activity of users, portals can have a more accurate insight on how their datasets are being used.

It is important to note that access through an API is not always useful or possible for users. Some users might not be technically savvy, or they might require bulk access for intensive analysis or integration. However, solutions like OpenDataSoft, which build their platforms around their own API, provide a means to track static datasets using the same metrics and log files. Using this approach, they are able to track how data is manipulated on their own site, as rows and columns of tabular data are indexed individually. However, in some cases this approach may be difficult to implement by requiring a portal

owner to rebuild their infrastructure. This may not be appropriate in many cases or the cost might be too high.

Generating API metrics for non-queryable data

For datasets without a query interface, one option is to create them by:

- Transforming them to RDF using a reference ontology and expose them through a SPARQL endpoint. This also has the advantage of integrating datasets, enabling much richer queries among them. A downside is the transformation cost, that can be run semi-automatically (e.g., with tools like the Silk framework), but still needs human supervision (and thus, effort), and the infrastructure cost of the SPARQL endpoint. For the latter issue, low-cost server-side interfaces[79] that push part of the query processing work to the client have been proven to decrease server load for single very large datasets (DBpedia) and for large collections of small datasets (LOD Laundromat).
- Tools like DreamFactory and Restlet can generate REST APIs for several types of databases and files, including CSV. Similar to a SPARQL endpoint, this requires the addition of a new component to the portal infrastructure. They also include log analysis tools. Some Open Data publishing platforms, such as Socrata and OpenDataSoft, provide a similar functionality as part of their solutions.
- For spreadsheets, Microsoft Excel and LibreOffice provide APIs to manipulate them programmatically. Exposing a subset of basic read-only functions (list headers, get columns, rows and ranges) as a microservice similar to the previous point is relatively easy to implement and maintain. A potential drawback is that if functions depending on the content of the table are desired, extra work designing and implementing the meaningful functions is required on a per-file basis.

The cost of providing query interfaces is proportional to the number, size and complexity of the hosted datasets, and to the expressiveness of the desired query interface (a per-dataset getRow/getColumn API is less expensive than a SPARQL endpoint for all portals). The overhead in logging is low on CPU, but on disk space, it depends on the frequency of access, for example, a dataset popular with applications will receive several hits and will end up with a very large log.

2.2.5.4. Benefits and limitations

Automated access metrics present portals with a significant opportunity to collect data that can inform use and impact measurement. One of the key benefits of an automated access metric approach is the availability of existing commercial, user friendly tools. The ubiquity of these tools and methods across the web means there are existing best practices in place and training available that portals can build on. These metrics also provide opportunity to automatically collect data at relatively low effort and low cost, in contrast to many of the other methods discussed.

While many people are interested in measuring impact of publication and also performance of an initiative, there are a variety of reasons to also measure more direct accesses. For example, this data helps portals identify pain points within their platforms. It can help with understanding the relative demand for different ways of providing data – whether some datasets are primarily downloaded while others have high API usage. Monitoring search and navigation can also provide insight into how people find data. These types of information can help inform portal's technical and design decisions.

---

[79] Linked Data Fragments. Available at: http://linkeddatafragments.org/in-depth/.

However, there are several limitations when it comes to automatically tracking access metrics as a measure of use and impact beyond the issues surrounding the relationship between access and inferred use.

First, several respondents identified the technical difficulties surrounding measuring access in a uniform way. As the survey respondent from Bath:Hacked put it "the many different ways that data is and can be used, e.g. downloaded for offline analysis, dynamically queried via APIs and interactive graphics… make it challenging to implement a single mechanism." Because of this 'heterogeneous environment', portals can find it hard to compare the various automated metrics in a meaningful way. The representative from the UK's national portal highlighted the technical challenge of drawing data from page analytics tools programmatically into their platform, given often changing APIs. This highlights the potential hidden costs of relying on such systems for evaluation.

Second, many respondents highlighted the difficulties in monitoring some of the automated metrics because of their platforms architecture or nature. Some highlighted that the commercial portal solutions they relied on did not allow for some of the metrics to be tracked or that accessing this data to carry out analysis was difficult. However, the primary difficulty came where data is not hosted from portals, namely those which catalogue distributed publishing - either from other portals or individual publishers. In these cases, the portals link to where the data is hosted and as such, they may be able to carry out page analytics but are unable to track direct downloads or API calls – as is the case for Spain and Belgium. Without access to the logs from the backend data hosting, it was increasingly difficult to track access. Even in cases where these platforms provide one way of accessing data directly, either direct download or programmatically through APIs, they are unable to capture all the ways data is accessed which makes comparisons and some metrics less insightful than if all data was accessed directly through the platform.

A third limitation consists in the fact that these metrics remain very partial in helping to understand the impact. As indicated in the opening of this chapter, multiple downloads by a bot may artificially increase the statistics without necessarily translating into re-use nor impact. Alternatively, a download by a bot may also be a means to distribute the data on another – potentially proprietary platform – where the data is then used by 10,000 users.

2.2.5.5. Lessons and best practices

Effectively tracking access through automated data collection clearly provides portals with insight into how their portal is used. This insight can be used for a variety of purposes, including design decisions. While it may not be the most useful means to measure use or impact, it can also be used to infer them to some extent.

Many portals already use page analytics tools, making use of existing commercially available tools and training to lower the cost and expertise required. Efforts should be made to identify measures that within these platforms that indicate intention to use - for instance individual dataset page views or analysis using inbuilt tools. In addition, many portals track the number of downloads of particular datasets. While these do not indicate use, they can be combined with patterns of behaviour when viewing the platform which might indicate use. Portals should continue to track downloads and attempt to link them to page analytics. Many portals already display these numbers on their website to help re-users, and some publish this as Open Data on the platform – for example the UK and Russian portals.

When it comes to tracking APIs, portals providing them should keep logs of the queries and methods being called. Those which do not provide APIs might, where feasible, use the available tools to make their tabular datasets accessible programmatically. This not only provides additional insight into how users are accessing data from their site but might also encourage greater re-use. Where possible, portals should examine their architecture to explore the possibility of driving access to data, whether download, data manipulation tools or APIs, through internal APIs - allowing each kind of interaction to be monitored through a single mechanism. While this might not be currently feasible, it might be considered in the development and procurement of future platforms.

Efforts should be made between portals to identify the most relevant individual metrics which can be gleaned from these automated access metrics. Portals should publish access data under open licences directly to their platform – as is done with the UK portal.

Recommendations for portals

- **Portals should use page analytics and track downloads at the dataset level,** capturing users' journeys through the portal.
- **Where possible, portals should keep APIs logs,** and explore methods for generating similar data from static datasets.
- **Portals should publish their access data under open licences** and work together to identify standardised metrics to track.
  - Best practice example: UK

2.3. Towards better measurement of use and impact

There are a wide variety of methodologies being employed to measure the use and impact of Open Data. Each method has distinct benefits and limitations, and none of them offer a single solution to the challenge of measuring use and impact for portals. However, from our research we have identified several opportunities for portals to improve their approaches to measuring impact. In this section, we examine how portals might make better use of existing data, potential methods they could use to track use automatically and how encouraging a culture of consistent and ubiquitous data citation might enable a much more effective method for measuring use and impact.

2.3.1. Better coordination and data sharing to model use and impact effectively

Across the methods we have examined, no single method offers an approach which provides ongoing assessment, requires minimal resource and still effectively infers use and impact. There are clear benefits to each approach, but these are often counterbalanced by its limitations. In addition, most of the current approaches lack repeatability, replicability and wider applicability between studies, and beyond the studies' scope or subject.

Approach	Scope or subject	Data collection	Effort or resource required	Ability to identify users	Ability to capture impact
Macroeconomic impact studies	Geographic / economic	One-off	High	Low	High
Microeconomic impact studies	Publisher / dataset / sector	One-off	High	Medium	High
Business population studies and user surveys	Geographic	Semi-regular	Medium	High	Medium
Showcases and use cases	Portal	Ongoing	Medium	High	Medium
Automated access metrics	Portal	Ongoing	Low	Very Low	Very low

Figure 5 Comparison of current approaches

If portals are going to improve their measurement of the use and impact of the data they publish then they must overcome these limitations. While we later explore additional approaches and methods that could be put in place, we believe that portals are able to improve their measurement of use and impact using the current measurement approaches.

To do this, portals must take a more holistic approach to measurement, taking advantage of the benefits of different approaches while minimising the limitations. By linking together the different types of assessment, we can begin to build a greater understanding of the ecosystem for the re-use of Open Data, and begin to more effectively model this ecosystem to infer use and impact. These efforts should focus on creating a more joined up approach to data collection and analysis between approaches, and a focus on wider data sharing and publishing by all studies.

2.3.1.1. Joining up data collection and analysis

When it comes to data collection and analysis, the current approaches we identified tend to take varying approaches depending on their specific goals. The aim to take up a more joined up approach is to identify common features between methods that mean they can be used in conjunction to build up a more consistent and fuller picture of use and impact.

To begin to take a more joined up approach, we must first identify a common reference by which all approaches can be usefully linked. The granularity of this reference is determined by the necessity to draw useful insights for portals, publishers and users. With the exception of macroeconomic studies, which analyse entire economies, all the other approaches either can or do capture details at a data publication level, where a data publication is understood as a dataset or datasets focused around a single theme and curated by a single publisher. For example, microeconomic studies tend to focus on known data publications from a single publisher and best practice in showcases is to link use cases to specific named datasets. While business population and user surveys are rarely linked to specific data publications, they often request some information about specific datasets, while most tend to collect information on data source and sector. This current data can be used to infer data publication, while

future studies could be adapted to capture this information in more detail to link the studies to specific data publications. Most obviously, automated access metrics record at a dataset level.

When data publication is taken as the reference point, microeconomic studies may be the best approach. Currently these approaches tend to draw on data about identified users, similar to business population studies, to derive use and impact.

However, one of the key aims of this exercise is to make better use of the methods that require less effort to collect data and do so on an ongoing basis. To do this, we must focus on how automated access metrics, showcases and business population surveys can collect the data required for microeconomic studies on an ongoing basis. For example, use case collection might involve collecting structured reference data on the users in a similar way that business population surveys do, using company numbers or requiring lookups to company registers. Business population surveys might be automatically sent out to registered users and via social media on a regular basis.

On the analysis side, measures might be put in place to calculate impact on an ongoing basis from these additional inputs. For example, TfL could compare data from new use cases or new surveys against existing data on re use to determine an inferred level of use and impact. It could then compare this inferred level of use and impact against the automated access metrics to examine any correlation to infer whether this is a new user or an existing user registering their use. Either way it could then calculate the potential additional use or impact according to its confidence in each of these values. These calculations based on additional data could likely be heavily automated using machine learning techniques, although this would probably require a reasonable amount of initial resource to set up. While this process might not be able to guarantee use or impact measurement with the same level of confidence as traditional microeconomic studies, it would provide more ongoing use and impact inference.

However, this holistic approach presents several key limitations:
- **It requires an initial baseline, microeconomic impact study to infer use and impact,** although data collection could take place before a study was commissioned. These types of studies are often beyond individual data publishers and especially for portals who might have hundreds of different data publications.
- **It does not address the technical issues identified in tracking access automatically** for portals who catalogue data which is published in a distributed fashion.
- **It still relies on self-reporting to automatically capture users and it is unlikely to identify many, if any, downstream users**. One method to counteract this might be to conduct regular manual business population identification methods, however this will likely undercut any of the resource benefits brought about by a more passive data collection approach across the rest of the features.

Recommendations for portals: taking a holistic approach

- Focus on use and impact at a dataset level.
- Connect methods used for macroeconomic and microeconomic studies to data from other collection approaches.
- Examine approaches to automating microeconomic analysis based on the ongoing data collection approaches.

2.3.1.2. Sharing access, use and impact data

The first two limitations of this holistic approach can be counteracted by sharing and publishing the underlying data.

By sharing the data that underlies microeconomic impact studies, other publishers and portals are able to model the impact of their datasets without commissioning their own study. For example, for a specific microeconomic study, the data on automated access (such as API logs, evidence of re-use), number and type of businesses identified, and the analysis for calculating impact are all shared. Another publisher or portal can then compare the type of data publication (e.g. content, quality, etc.) and the access and use data to infer the relative impact of their publication.

Similarly, if portals share their underlying access and use case data, other portals are able to compare their own data publication type, identified users and access metrics in an effort to understand whether they have identified a similar level of use. If the discrepancies are unexplained by differences in data quality or contextual data, such as overall business population, data skills, age of portal etc, it might be used to infer they have other users they have not accounted for. For example, a portal might identify that a dataset experiences a similar level of access to a comparable dataset on another portal. That portal may have several use cases which link to the specific dataset while the original portal may have none. Taking into account the differences between the datasets, portals and context, the investigating portal might infer a proportional level of use. This might be especially useful where one portal has carried out a business population study.

While the above methods might not confer actual use, it could lead to better estimations of use or stimulate future study. For example, if several portals have similar datasets and a number of those have similar use cases registered then it is likely that those patterns of use will be replicated across all the portals, so long as there are similar needs. While there is no certainty that the use case will be replicated, it allows portals to focus their efforts to identify those use cases or it might prompt them to create the conditions for the use to occur.

Portals have already begun to recognise the benefit of sharing some of this data. The representative of Russia's national portal highlighted that they used case studies from other portals, such as the UK's national portal, to demonstrate potential impact. In particular, portals and publishers are keen to compare themselves to others. Several portal owner respondents already published automated access metrics for individual publishers and datasets as Open Data on their portal – including the UK and Russia. The primary motivation is to report these figures to publishers, in some cases to encourage competition. However, when linked to use cases or microeconomic impact studies they could be used to model comparative use and impact. The representative of OpenDataSoft mentioned that portal owners frequently ask to benchmark their activities and access metrics to other comparable portals. By publishing this data, alongside use cases linked to the data, portals can begin to infer use of their data from similar data on similar portals. The more data that is published by portals, the better the models for inferring use and impact get. For example, if one city portal infers use based on just one other portal's dataset it is likely to be subject to significant assumptions, whereas if ten examples exist there is a higher likelihood that the inferred use is accurate.

Data sharing can also help overcome some of the technical challenges to monitoring access. Several respondents highlighted that portals based on catalogues could coordinate with individual publishers to share downloads and API logs – including Spain and Belgium. The core challenge with this is engaging each owner individually. However, once this is done initially, the transfer could be done automatically. Portals could motivate publishers by sending page analytics data or offering to handle

their overall evaluation. The benefits of doing this at a portal level is to coordinate a comparison and to be able to link access to use cases and business population studies.

Data sharing also offers advantages for collecting downstream data on access and use. The representative from Luxembourg spoke about how they had provided funding for an aggregating API service which was developed as part of a hackathon. In return, the aggregator agreed to share the downstream access data with them, enabling them to track downstream use. Similar efforts could also be made with end user services, where data users are incentivised to provide records or summaries of their user community. These in turn could be built into impact measures. One way of incentivising this data sharing may be mandating it for users who receive funding as part of hackathons as done in Luxembourg's gameofcode.eu[80] hackathon. From a few examples, downstream use can be modelled for other identified users, especially if it is combined with other data sources such as app download figures on application stores. A number of respondents highlighted that they are attempting to work with some of their users who are close to them to gather these types of downstream data.

Recommendations for portals: sharing data

- 🌐 **Share, and if possible publish, underlying data from all studies.**
- 🌐 **Use underlying data and methods from other portals to infer use and impact of your data,** based on comparisons in type and access metrics.
- 🌐 **If necessary, collect and share metrics for data published elsewhere than on the platform.**

2.3.2. Exploring technical methods for automatically tracking use

While more joined up approaches might help portals to better infer use and impact of Open Data from their portals, none of the current methods overcomes the challenge of self-reporting and being able to automatically track use. Only 26% of respondents to our survey had attempted to automatically track re-use, and 58% said it was not feasible to do so. However, automatically tracking use could potentially lower the resource burden of identifying users and minimise self-reporting bias. As such there are benefits to exploring the feasibility of automated approaches to tracking use, and during our research we identified three potential methods to automatically detect use: tracking API keys, implementing version control systems and tracking through web search.

2.3.2.1. Tracking users through API keys

The first focused on using API keys, and to a lesser extent logins, to identify who is accessing data. By identifying the person who signed up to the API key, portals are able to understand exactly how certain users are accessing the data and what data they are accessing. If on registration respondents are required to provide data on their organisation, especially using unique identifiers like company numbers, this data can be combined with other sources to determine attributes like economic value. Regular follow-up surveys and contextual data might make it possible to link their specific use data to other quantifiable and qualitative impact outcomes with a lot less effort than existing methods.

There are also potential advantages in federating access to API logs and access to datasets. In practice, that means allowing a user to access and query datasets in different portals with the same API key - each portal would be responsible for storing the log for their datasets, granting read access for this log to other portals. By accessing this log, the second portal would then be able to see the links

---

[80] Game of Code. Available at: http://www.gameofcode.eu.

between this access and accesses registered by their portal, which would be impossible by solely looking at its own log. This information could be very valuable in helping portals to understand their users and how they can collaborate.

There are other advantages to tracking API keys, for example lowering the impact of poor API use. Once the API key corresponding to a frequent usage is identified the portal can contact the person/organisation behind the API key to discuss more efficient ways to access the dataset. This is the approach taken by a number of respondents including TfL. Getting in touch with these "power users" of datasets – those who are frequently using the data – might also uncover further usage of data that is not captured by tracking.

However, as discussed previously, requiring logins and API keys has several limitations; going against the spirit of openness, possibly limiting actual use, lack of incentive to provide correct information. Given these issues, we also explored potential methods to track use which do not involve logins.

2.3.2.2. Tracking use through version control

The second method focused on a technical solution to create communities and track usage of datasets. As explained earlier, a portal's query services provides an opportunity for analysing their activity. Further services that promote collaboration around datasets may uncover hidden use of datasets and support new and innovative ways of using them. The ultimate vision is to create communities around Open Datasets in the same way that social coding tools like GitHub and GitLab have contributed to create and maintain communities around open source projects. By adopting the concept of "repository" from open source development, and borrowing some of their functionalities, data portals and dataset re-users get the following benefits:

- **The concept of a "download" is replaced by "cloning" or "forking" – making copies of a dataset for use by the re-user**. From the point of view of tracking re-use, cloning is similar to download in the sense that the dataset goes directly into the re-used infrastructure without carrying extra information about how a dataset is re-used and by whom. On the other hand, when forking, a copy of the repository under the control of the re-user is stored in the portal infrastructure.
- **Re-users can track changes they do to datasets, for example, cleaning, changing format, normalizing values or adding new data points.** Sets of changes can be proposed to the dataset maintainers to be integrated into the "official" version of the dataset through pull-requests. From the point of view of maintenance, dataset maintainers can crowdsource the task of spotting errors, that often get revealed only when the dataset is being used.
- **From the point of view of tracking re-use, a cloned repository that requests a pull needs to become publicly accessible**, revealing to the dataset maintainer the changes made. A dialog with the requester may follow to gather more insight about how the dataset is being used. In the case of forks, changes can be seen even before a pull is requested. In the event of a pull request being rejected, the modified dataset would still be available on the forked repository as a derivative dataset, automatically preserving the link (thus, a form of data citation) to the original dataset.
- **Users can manage code that is developed with data as starting point, such as visualisation scripts, on the same cloned/forked repository as the data.** As it is the case with change sets, this code and knowledge on how the particular use is visible to others on publicly available clones and forks.

> The availability of issue trackers, forums and wikis integrated with commits can enable further social interaction. This community can also be monitored to understand the types of uses and issues with the datasets.

Besides data itself, a repository can also contain metadata like license, schema, and/or provenance, resembling the concept of a data package[81]. This vision could be realised by reusing elements of established tools for Version Control Systems for software (VCS). It is important to understand the similarities and differences between use cases for software and datasets respectively. These issues are explored in the example below.

Using VCS software tools for data

The differences between use cases for datasets and software can be illustrated through an example. Consider a tabular dataset that describes the location of bike racks within a city. It is comprised of three columns: geocoordinates, type and capacity of the bike rack. A re-user has forked it.

> **Data Cleaning**: When re-user examines the dataset, he notices that all geocoordinates miss the closing bracket, causing an error when he tries to load the dataset in his GIS. Following software development best practices, he creates an issue in the repository of the dataset, add the missing brackets and makes a pull request that solves it.
>
> **Adding data points**: The re-user notices that the dataset is missing the bike rack that is in front of his business. He creates an issue, adds the new row and makes a pull request that closes it. This is roughly similar to when a new functionality is added to a piece of software. Note however that our example dataset is particularly interesting because re-users can easily add new data, and for the data maintainer is desirable to invite re-users to test the quality of datasets.
>
> **Adding properties to data points:** While developing his application, the re-user considers useful to have a 4th column to represent if the bike rack is covered or not. He creates an enhancement proposal and submits a pull request that solves it. Again, similar to the new functionality use case in software development.

Creating a system that brings communities together and tracks dataset usage requires a number of elements. Firstly, a hosting space for the repository is needed (assuming that individuals have decided to re-use established VCS tools). A portal may do it in-house, by integrating GitLab into its own portal, or rely on existing public infrastructure, such as github.com or gitlab.com.

Secondly, each dataset would need a maintainer who is responsible for deciding which pull requests to accept. This system would be limited by size, as larger datasets over 50Mb are not suitable for software version control. For larger datasets, the use of Dat is recommended. Dat is a data sharing tool specifically designed for large datasets that might be updated often. Dat provides a more efficient way of transferring large datasets, at the expense of not having support for pull-requests.

Last, datasets which are updated frequently (e.g., sensor measurements) would generate a large number of commits using a version control system. One way to minimise the impact of this would be to use an API for real-time or near real-time access and upload regular but less frequent snapshots, for example daily snapshots.

---

[81] Frictionless Data. Available at: http://frictionlessdata.io/data-packages/

However, one key limitation of this approach is that it is limited to tracking these actions as they are performed on the platform. While many users may choose to do this for convenience, conferred by the tools provided, eventually most uses will involve drawing the data out of the platform. At this point, it suffers from similar issues as current platforms in tracking the number of re-users – one clone of the dataset may be used by many users downstream which cannot be tracked.

2.3.2.3.   Identifying use through search

The third method proposed during our research was based around using other existing web monitoring tools and techniques to identify use through reference. A number of methods were proposed based on tracking references to the dataset or publication. Several have been tested by Amazon Web Services, primarily because the data they host is stored in s3 buckets which each have a unique name. Using this unique name, efforts were made to use 'news alert' features which notify AWS when this term is used in an indexed web page, however because the names do not have to be universally unique this creates a significant amount of false positives. In addition, AWS searched the open-source code repository, GitHub, for mentions of the s3 bucket name which yielded a number of results. However, this method can only capture use with open source software and does not guarantee that these will be 'live' at the time of search.

Similar approaches have been proposed for other datasets with identifiers, for example one can crawl scientific papers, application descriptions, news sites, blogs and social networks, to collect mentions and citations to the dataset. In the first instance, a weighted average of each type of citation, provides an estimate of the "attention" that the dataset is receiving from these media, in a similar way of how altmetrics do for research outputs. One can also use similar measures of importance for data being used in posts and articles, where datasets are ranked, in a similar way of the transfer of authority of PageRank for web pages or impact factor for research networks. A further improvement is the analysis of how the dataset is being mentioned: is it an article about the dataset? or that simply uses it to support a claim? Is a blog post showing a new way of visualizing the dataset? Or complaining about its poor quality? Sentiment analysis techniques routinely used in social media analysis can be applied to provide further insight.

The benefits of using these techniques are that they do not require portals to track the re-users of their data and they do not require an entire redesign of the portals architecture. If implemented fully they could identify re-users on an ongoing basis, including many downstream users. However, the core issue with using web crawling and search techniques is identifying the dataset or data publication through particular terms. Without a consistent way of identifying whether a dataset has been used, it is almost impossible to use these techniques.

> **Recommendations for portals: Automated approaches to use**
>
> 🌐 Explore how tracking APIs, creating version control hosting and web searching technologies could be used to track use.

2.3.3.   Measuring downstream usage through data citation

One approach to potentially tracking downstream usage of Open Data is the use of data citation. Drawing on the culture of citation in academia, data citation presents the opportunity for increased opportunity to track the usage and impact of Open Data.

By using citation trackers or searching the web for citations using some of the methods outlined above, data citation may enable publishers to more accurately understand how and where data is being used and re-used. Citation counts could be used to measure their contributions to impact, which have previously remained untracked. Publishers may then be able to better support users to encourage greater engagement and use of their data.

2.3.3.1.   What is data citation?

As defined by the Data Citation Synthesis Group, a citation is a 'formal structured reference to another scholarly published or unpublished work'[82]. In the context of data, citation is providing the 'full bibliographic reference information for the object', including a link to its source, that indicates its provenance. This allows anyone using derived content, products or services to verify the accuracy of the content and locate the original data.

Citing Open Data is distinguishable from the dominant practice of attribution. Attribution is a legal requirement of many open licences (including widely used Creative Commons licences[83] such as CC BY and CC BY-SA), which requires the re-user to credit the original creator.

There are, however, a number of limitations with the existing system of attribution:

- **It may be difficult to know how to comply with attribution requirements.** Publishers do not always state their preferred means of attribution, so that users understand how to best comply. Although certain licences like the Open Government Licence provide a default format, this is not true of all licences, which contain attribution requirements.
- **Requirements for attribution can be onerous for some re-users.** Publishers who are unaware of how their data is used may not think about its implication for different formats. For instance, some licences may require the attribution statement to be in a particular size or font that is difficult to comply with across all devices and products using the data.
- **Acknowledging multiple attributions can be difficult for re-users.** An application which uses datasets from many different sources may be required to acknowledge them all. This may lead to a problem of 'attribution stacking', where 'a derivative work must acknowledge all contributors to each work from which it is derived.'[84]. Although Creative Commons licences have tried to address this issue through allowing users to link through to a web page, the variety of attribution requirements may create an unnecessary burden on Open Data re-users that make individuals less likely to use it. The Sunlight Foundation explains: 'the desire to describe data provenance should not involve a legal requirement [to attribute] that will hinder the freest use of the data'.

Data citation presents an alternative solution. Instead of focusing on attributing credit to the original creator, the purpose of data citation is to reference the provenance of data so that end users can 'find out more about that information's context, development and quality'. As in academia, citation is not a legal requirement for using Open Data, but rather a convention that is encouraged for supporting good scholarly practice. In particular, it has helped measure the contributions of researchers who have collected and published data, which was previously untracked.

The use of permanent identifiers (such as Digital Object Identifiers) in data citations would enable portal owners to search for products using the DOI of a specific dataset, using the methods detailed

---

[82] Data Citation Synthesis Group
[83] Creative Commons. Available at: https://creativecommons.org/licenses/.
[84] Digital Curation Center. Available at: http://www.dcc.ac.uk/resources/how-guides/license-research-data.

in the 'identifying through search' section above. In the UK, the UK Data Service has been using this technique through its UKDS.Stat service to understand the re-use of its data in research[85].

Some of the Open Data portal owners that we surveyed for this report praised the success of data citation in academia. The representative from Bath:Hacked highlighted the success in using standardised citation and identifiers to measure use and impact around scientific research, particularly the impact of clearly identifying datasets using the Digital Object Identifier (DOI) system. The possibility of replicating these benefits by implementing a similar system for Open Data should be seriously considered.

Literature around data citation suggests there is more agreement on what a data citation should contain, as opposed to an attribution statement. W3C's Data Usage Vocabulary[86] states that all data citations should include the following:

- Author
- Year
- Publisher
- Distributor (organisation who makes the dataset available)
- Edition or version
- Access information - URL or persistent identifier (most commonly a **Digital Object Identifier**)
- Location

Data citation provides a standardised alternative that may help overcome the friction in current attribution requirements. Encouraging one primary (but flexible) way in which re-users can cite data will ease the process of citation and is more likely to encourage the practice to become more widely adopted.

2.3.3.2. Why should we cite?

Data citation may provide a number of benefits for both Open Data producers and users.
For portals, it may enable owners to demonstrate reach more effectively and secure ongoing funding for Open Data publication. Portal owners can use the number of citations across the web to evidence their Open Data impact and demonstrate their reach. In academia, Thomson Reuters have suggested that tracking citations via their Data Citation Index[87] can help 'validate return on funding investment'[88]. Initiatives that can demonstrate impact with past funding and resources are more likely to secure these in the future.

Additional benefits of data citation may include:
- **Increasing the discoverability of Open Data, enabling it to be more easily located and re-used.** By enabling users to locate the source of data, data citation allows increased discovery of data released by governments and businesses[89]. Increased discoverability may encourage increased use in products, services and journalism, leading to increased impact. It will also help end consumers verify that the accuracy of data-enabled services, building trust between publishing organisations, service providers and consumers.

---

[85] UK Data Service. Available at: http://blog.ukdataservice.ac.uk/tracking-data-citation-in-international-data-use/.
[86] W3C (2016). Data on the Web Best Practices. Available at: https://www.w3.org/TR/vocab-duv/.
[87] Data Citation Index. Available at: http://wokinfo.com/products_tools/multidisciplinary/dci/.
[88] Thomson Reuters (2015). Recommended Practices to Promote Scholarly Data Citation and Tracking. Available at: http://images.info.science.thomsonreuters.biz/Web/ThomsonReutersScience/%7Becce05a1-2d36-4596-81ca-374bcb04ef59%7D_Data_Citation_Index_whitepaper.pdf.
[89] Data Science Journal. Available at: https://datascience.codata.org/articles/abstract/10.2481/dsj.OSOM13-043/.

- **Enabling portal owners and data publishers to understand and more effectively cater to Open Data re-use.** Creating a function that searches the internet for citations of Open Data, as trialled by the UK Data Service for their census data[90], would enable a deeper understanding of the variety of Open Data re-use. Understanding how (and whether) data is being re-used can help portals address problems with its quality or availability: academic Stacy Konkiel states that 'determining the scope and quality of [data's] impact could speak volumes about the quality and utility of the data itself'[91].
- **Encouraging common norms to be adopted across the Open Data community that work with multiple licensing regimes, easing friction in Open Data use.** Data citation represents a standardised, formal process that substitutes the more 'free-form' attribution requirements that currently exist. Encouraging these norms replaces the confusion that may result with creating new legal requirements, and is likely to be an addition that will work across licensing regimes. This will encourage common norms that can be flexibly adopted by users across the spectrum of Open Data re-use.

These advantages have led a number of organisations in the Open Data space to favour citation. The Sunlight Foundation, a transparency and accountability-focused organisation, have encouraged publishers of Open Data[92] in government to publish Open Data in the public domain (using a licence such as CC0) and request citation, rather than attribution. Similarly, the Science Commons Open Access Data Protocol[93] recommends that academic institutions publishing open access data 'must not apply any obligations on the user...even the legal requirement to provide attribution', favouring a 'non-legally binding set of citation norms'.

2.3.3.3. Challenges of data citation

The adoption of data citation may present a number of cultural and practical challenges for the Open Data community.

Overcoming barriers of disagreement about open culture

There is disagreement in the Open Data community about the legitimate expectations that can be placed on a re-user of Open Data. Whilst some argue that attribution and citation are best practice for demonstrating Open Data impact and growing the community, others have viewed them as a barrier that limits what should be freely available data. This is a debate which must be addressed if portals wish to encourage data citation practices.

Some organisations have argued that legitimate Open Data requires as few restrictions as possible on its downstream usage, creating 'licence-free' data where copyright is not a legal issue. In its review of the Ordnance Survey's licensing restrictions on its data, the Open Data User Group maintains that 'permitting the unrestricted downstream use of the data is in line with the fundamental principles of Open Data', and therefore should be lifted in order to 'maximise the economic and social benefits which can be realised'[94].

---

[90] UK Data Service (2016). International and census Open Data use at the UK Data Service and the fruits of data citation. Available at: http://blog.ukdataservice.ac.uk/international-data-use-at-the-uk-data-service-and-the-fruits-of-data-citation/.
[91] Online Library (2013) Tracking citations and altmetrics for research data: Challenges and opportunities. Available at: http://onlinelibrary.wiley.com/doi/10.1002/bult.2013.1720390610/full.
[92] Sunlight Foundation (2014). Don't attribute Open Data — cite it!. Available at: https://sunlightfoundation.com/2014/03/26/dont-attribute-open-data-cite-it/.
[93] Science Commons. Available at: http://sciencecommons.org/projects/publishing/open-access-data-protocol/.
[94] Review of the Ordnance Survey's licensing restrictions. Available at: https://data.gov.uk/sites/default/files/20130717%20OS%20Open%20Data%20Licensing_10.pdf.

Author Joshua Tauberer takes this point a step further[95], stating that requirements to attribute or cite data (terms used interchangeably) 'create a lever - a civil penalty arising out of violation of a contract - by which the government can control speech'[96]. This is especially problematic where citations could make it easier for government to track the re-use of data. Anonymity may be an essential condition for ensuring the safety of activists using Open Data in closed or semi-closed societies, and the requirement to cite data where it is being used could open up individuals to restrictions on particular use by the government.

It is important that the Open Data community condemns any inappropriate policing of Open Data use, regardless of whether attribution or citation is required. Discouraging network effects around data is not in the best interests of data publishers. Governments who wish to track re-use through citations should release guidance that reassures users that they are not being policed, and that citation tracking is used to illuminate the network of re-use by connecting datasets to their use in products and services in wider society.

Practical barriers

Open Data can be used in a huge variety of ways, from reports to applications to journalism. These different communities will have different practices when it comes to citing or attributing data, presenting a problem for consistency in presenting data citations. Practically, encouraging widespread use of standardised data citation will require overcoming this barrier. The academia community has seen similar issues (see the below case study).

Citing Open Data consistently will require a general consensus around when and where data citation is necessary. Publishers and portals should work together to facilitate ways for users to cite date in a consistent format. Publishers who wish to encourage a culture of data citation should issue guidance around the granularity of data that needs to be cited, in order to avoid barriers and encourage a culture of citation to develop.

Case study: data citation in academia

The increasing visibility of data citation in academia provides a case study for addressing these challenges.

Historically, academia has a prominent culture of formally referencing sources of knowledge to support repeatability, verifiability and academic rigour. In an increasingly data-intensive research environment, a number of academic disciplines have been moving towards implementing data citation over the last decade. As described by citation proponent the Future of Research Communications and e-Scholarship 2011[97] (FORCE 11)'s Data Citation Principles, data citation is an essential 'part of the scholarly ecosystem supporting data re-use'.

---

[95] Open Government Data: the Book (2014). Available at:
https://opengovdata.io/2014/no-discrimination-license-free/
[96] Open Government Data: the Book (2014). Available at:
https://opengovdata.io/2014/no-discrimination-license-free/
[97] Force 11(2014). Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Available at:
https://www.force11.org/datacitationprinciples.

A notable case study is the SageCite[98] project, which explored frameworks for implementing citation norms using bioinformatics data. SageCite was a joint initiative between UKOLN, the University of Manchester and the British Library, funded by JISC and with support from Nature Genetics and PLOS. The initiative developed and tested a citation framework that linked data, methods and publications to ease friction and, encourage citation and dissemination of data across research communities. Despite the pre-existing culture of referencing sources, moves to adopt data citation in academia have been stunted in some areas. In 2014, impact-focused Impact Story estimated that only 25% of journals signed up to Force 11's Data Citation Principles were implementing them in practice[99].

Academia provides a number of lessons that the Open Data community could learn from regarding citation. For instance, academic data citation principles provide a framework for overcoming the challenge of enforcing one form of standardised data citation across diverse data use. Academic research data may be used for a wide variety of outcomes. FORCE 11 have encouraged journals and other data publishers to be 'sufficiently flexible to accommodate the variant practices among communities, but [citations] should not differ so much they compromise interoperability...across communities'[100].

Best practice in enforcing a common standard requires that a number of elements are present across every data citation, such as author, permanent identifier and location. A formalised and standard practice is what makes data citation attractive, consistent and standardised. However, by encouraging flexibility, publishers can also support wider adoption of citation norms, allowing them to better track data re-use.

Drawing from this example, Open Data publishers should consider reducing the specificity of guidelines around citation, and instead encouraging flexibility in the form of citation so that different products and services may cite in ways best suited to their use. This will enable the benefits of data citation to be felt by Open Data publishers, intermediaries and end users.

2.3.3.4. Encouraging a culture of data citation

A number of prominent Open Data publishers and organisations have consciously chosen to encourage citation of Open Data over the last few years. After their prominent #CitetheData campaign, the UK Data Service now encourages citation of its census data[101]; likewise, the US National Oceanic and Atmospheric Administration (NOAA) has issued a Data Citation Procedural Directive that sets out the procedures around citing their Open Data. Advocacy organisations such as Creative Commons and the Sunlight Foundation have both issued content which suggests a move away from mandatory attributions towards alternatives (see a CC blog here[102] and Sunlights Open Data Policy Guidelines 3.0[103]). The move towards data citation from Open Data publishers provides encouragement that a culture of citation is growing in the Open Data space.

---

[98] SageCite. Available at: http://blogs.ukoln.ac.uk/sagecite/.

[99] ImpactStory (2014). Tracking the impacts of data – beyond citations. Available at: http://blog.impactstory.org/data-impact-metrics/.

[100] Force 11(2014). Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Available at: https://www.force11.org/datacitationprinciples.

[101] UK Data Service Census. Available at: https://census.ukdataservice.ac.uk/use-data/citing-data.

[102] Creative Commons (2012). Library catalog metadata: Open licensing or public domain?. Available at: https://creativecommons.org/2012/08/14/library-catalog-metadata-open-licensing-or-public-domain/.

[103] Sunlight Foundation. Available at: https://sunlightfoundation.com/opendataguidelines/#license-free.

Encouraging a culture of data citation in academia has required providing incentives to cite data for data sharers and users. Scholars from the London School of Economics, Hyoungjoo Park and Dietmar Wolfram, have suggested[104] that standard formalised citation practices could be an incentive for data authors to make data available for re-use. For Open Data, portals could motivate users by advertising derivative work on the platform or building a community around those who cite a particular dataset.

To further encourage these developments, those who manage and publish data should begin to recommend citation of their Open Data, including developing 'model citations[105]' that illustrate best practice. Moves should be taken to lower the barriers for publishing data with citation, a move encouraged by the Sunlight Foundation[106] and for Open Data released by the US government[107].

Implementing a culture of data citation in a way that encourages common but flexible standards and supports the principles of the Open Data movement more generally will ease the friction in Open Data re-use and enable governments to better track re-use. This will encourage benefits of citation, and Open Data more generally, to be shared among governments, intermediaries and re-users.

2.3.3.5. Lessons and best practice

Portal owners should look into adapting and implementing practical tools that could allow Open Data users to easily generate a data citation for the data that they wish to cite. These could be developed in a way which allows people to generate a data citation that conforms to standard practice, such as including a digital object identifier, and can be formed by simply entering the dataset's URL. The individual creating the citation would then receive a link which they can use for accessing the metadata in one location. This would reduce the barriers to data citation, especially when datasets are sourced from more than one portal. Citations could also be generated in different formats depending on user needs, such as academic citation and in-app referencing.

Recommendations for portals

- Develop 'model citations'[108] that illustrate best practice for re-users to cite Open Data
- Publishers should issue guidance around the granularity of data that needs to be cited
- Public sector bodies should release guidance explaining the use of citation tracking to reassure users that their activities are not being policed, and are used for informational purposes in connecting datasets to their use in products and services in wider society

---

[104] London School of Economics (2017). Formalised data citation practices would encourage more authors to make their data available for re-use. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2017/07/17/formalised-data-citation-practices-would-encourage-more-authors-to-make-their-data-available-for-re-use/.
[105] Sunlight Foundation. Available at: https://sunlightfoundation.com/opendataguidelines/#license-free.
[106] Sunlight Foundation. Available at: https://sunlightfoundation.com/opendataguidelines/#license-free.
[107] The United States Projects (2013). Open Government Data. Available at: https://theunitedstates.io/licensing/.
[108] Sunlight Foundation. Available at: https://sunlightfoundation.com/opendataguidelines/#license-free.

3. Part 2: Creating the funding environment to sustain portals

As portals in Europe become more widespread, mature and advanced, the question of financial sustainability is becoming increasingly important. In more mature countries, Open Data is no longer the highest priority on the agenda, and even with political and legislative commitments in place, portal teams are increasingly being asked to justify their expenditure. As Open Data moves from emerging technology to part of the day to day function of government, falling budgets across government could result in a slow-down or even loss of some public funding.

At the same time as potential slow-down or loss of funding, increased portal maturity leads to greater ambition. Portals have wide ranging responsibilities and objectives, including but not limited to; opening up more data, improving the provision of existing Open Data and encouraging innovative re-use. Each of these responsibilities requires sustainable ongoing funding to deliver against their ambitions. Yet in 2017, 71% of EU28 countries cited a lack of financial resources as a barrier to Open Data publishing. Besides, this lack of financial support is preventing many portals from conducting proper impact evaluations.[109]

With the potential for funding restrictions and the need to support a wide range of activities, it is important for portals to understand how to use the resources that they have effectively and secure funding which can sustain their activities going forward. In addition to being able to understand their impact, portals also need to understand their costs. They also need to look at potential strategies for reducing or redistributing these costs over the long term by partnering with others to share the costs of implementing Open Data initiatives.

Some portals are exploring methods to diversify their revenue streams, looking to supplement their public funding from other sources. Each of strategies for dealing with funding has implications for the organisation and activities of the portal, meaning that portals might need to adapt to new economic realities.

In this chapter, we explore how portals are currently funded and the sustainability of this funding. We then examine the activities of portals broken down across three key dimensions: maintaining portal infrastructure; engaging and encouraging publishers; and building awareness, engagement and innovation. Within each of these dimensions we will examine; the associated costs, their impact on portals and the ways in which portals might reduce these costs. Finally, we investigate the potential alternative funding structures portals might consider to ensure financial sustainability into the future.

---

[109] European Data Portal (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

Recommendations for portals: creating a sustainable environment

Current approaches to funding portals
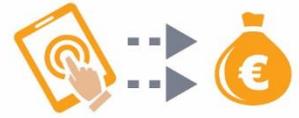
Ensure strong procurement processes are in place if using hosted solutions.

Ensure sufficient levels of technical knowledge through hiring or upskilling staff for open source solutions.

Aim to share the cost burden of technical hosting by partnering with other portals or governments.

Engaging and encouraging publishers

National portals should aim to have individual named publisher contacts.

Aim to reduce costs by engaging multiple publishers through workshops and training.

Explore the use of self-service tools for publishers such as dashboards, usage reports and rankings.

Building awareness, engagement and innovation

Invest in events as a method for encouraging innovation and improve publishing

Create shared innovation strategies and funds across cities and regions

Form partnerships with organisations beyond government, such as universities.

Alternative funding models

Explore how to adopt a freemium model for datasets that are not yet open

3.1. Current approaches to funding portals

The diversity of portals across Europe is informing a wide variety of funding models. Not only are funding models different between Open Data portals in different countries, due to different levels of priority placed on the principle of Open Data, but portals are often funded differently depending on their position at the national, regional, city or local level. This is creating a complex environment that portal teams must navigate when facing the question of financial sustainability.

Despite this diversity, many portals have a common feature: distributing data at a marginal or zero cost. In our survey, 84% of portals explicitly stated that they are distributing Open Data at zero or marginal cost, with the remaining respondents stating that they relied on some form of government funding to distribute data. This supports the EDP's recent findings that, by law, 100% of EU28 countries are using this model[110]. This means that many portals are not receiving any revenue from the sale of data itself.

The rationale behind this funding model is defined partially by the nature and beliefs around Open Data. Many portal teams have consciously adopted a zero-cost model in line with a belief that open government data has already been 'paid for by the taxpayer' and should be free. This is in line with the Open Definition.[111] The principle that Open Data should be provided at no cost may also be stipulated in Open Data strategies or legislation: this is the case in Belgium, where the federal Open Data Strategy promotes minimal restraints and maximum re-use, and in France, where the right of access to administrative documents mandates freely available Public Sector Information.

3.1.1. Funding national Open Data portals

Funding a national Open Data portal is usually the first port of call for a country investing in an Open Data initiative. Consequently, the funding models of national Open Data portals are usually the most mature and developed.

A large majority (91%) of national Open Data portals across Europe are funded as a function of the national or federal government. This support is usually provided as a function of the central government, with portal teams located within central government departments, in reflection of the fact that these departments are often leading the Open Data initiatives. However, the host department varies from country to country. In the UK, the data.gov.uk portal sits within the Cabinet Office, which provides funding, hosting and maintenance costs, with other departments funding their own activities around data collection and publication. On the other hand, the Czech National Open Data Catalog[112] sits within the Ministry of the Interior. Our research showed no significant difference in funding or activities between those portal teams sitting in different government departments.

---

[110] European Commission (2017). Open Data Maturity in Europe 2017. Available at: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.
[111] Open Definition. Available at: http://opendefinition.org/od/2.1/en/.
[112] Czech National Open Data Catalog. Available at: https://opendata.cz.

Size of national Open Data Portals in the EU – best practice Ireland

When zooming in on the size of national Open Data Portal teams in Europe, on average national teams consist of 3.2 people dedicating 2.2 full time equivalents (FTE) to Open Data. In most teams Open Data is only part of their responsibilities, which is mainly true for the smaller sized countries.

Clustering the countries in terms of their maturity level as determined by the EDP Open Data Maturity Report 2017, the size of teams varies substantially. Looking at the trendsetters in Europe, larger countries dedicate between 4.5 FTE to 10 FTE to the national portal consisting of 7 to 10 people. Medium sized trendsetter countries consist of 4-5 people dedicating between 1.6 FTE and 4 FTE to the portal. In the smaller sized trendsetter countries, the team size varies between 1 and 4 people dedicating 0.5 FTE and 3.25 FTE to the portal.

Looking at the large fast-tracker countries, time dedicated to the national portal varies between 1.6 FTE and 2.6 FTE by 2-3 people per team. Medium sized fast-tracker countries have 1-6 people dedicating between 0.29 FTE and 6 FTE to the portal while smaller sized fast-tracker countries have 2-3 people dedicating between 1 and 1.5 FTE to the portal.

Finally, looking at the followers, medium sized national Open Data Portal teams usually consist of 2-3 people dedicating on average 0.7 FTE to 1.1 FTE to the portal. Smaller sized follower countries generally consist of 1 to 3 people dedicating between 0.1 FTE to 2 FTE on the portal.

Looking at this overview, it becomes clear that there is a correlation between the size of a national Open Data Portal team and the maturity level they have. At the same time, relative to population size, smaller countries appear to dedicate more resources to Open Data than larger countries do. In general, most teams do outsource technical support who are not considered to be part of the national Open Data Portal team.

The most mature Open Data country in the EU according to the EDP 2017 measurement is Ireland. How sustainable is the Irish Open Data initiative? The national Open Data Portal team itself consists of 4 people dedicating 3.25 FTE to the portal. However, in addition, the team works closely together with multiple other actors:

1. 4 FTE are contracted out to a company that provides technical support.
2. The Public Bodies Working Group - a network of officials from Public Bodies who they consider to be champions in Open Data - has assisted in a lot of the Open Data publications such as the Open Data Technical Framework which underpins the publication of data on the national portal and had input into the national Open Data Strategy. The group meets quarterly to discuss (mostly technical) Open Data issues.
3. An additional contractor is responsible to roll out Open Data training across public service organisations.
4. A framework for Open Data technical support is in place of around 9 companies who can be contracted by public bodies to assist in data audits, data cleansing, data publication, etc.
5. Recently a network of Liaison Officers had been put in place – some 70 people who are now the point of contact for Open Data in their respective organisations. The national team engages with these contacts to communicate and promote Open Data, to promote the Open Data training etc.
6. Finally, the initiative is overseen by the Open Data Governance Board who comprises of external (non-Public Servant) experts in Open Data. This group is more tasked with strategic issues and meets quarterly.

Where portals are located outside central government, this may create organisational issues for portals in guiding continuity, development and growth. The Russian national portal[113] is funded through a public procurement system where regular contracts for managing and developing the portal are awarded. Organisations who win these tenders have time-limited contracts to carry out activities and must finish these activities within the contract period. Whilst this approach encourages a healthy sense of competition, the bureaucratic procedures inherent to any procurement process may be problematic for supporting long-term projects necessary for a portal to mature.

Whether portals are developed in-house in government or via public procurement processes, there is a clear impetus to ensure that contracts are renewed on time or that knowledge is transferred to ensure continuity. As examined later, technical knowledge regarding portals is often highly concentrated in a few highly skilled individuals, with little resource to train and skill up additional people. Consequently, Open Data portals at every level of government should implement strong knowledge management procedures.

The rationale behind providing national Open Data portals on a zero-cost model, supported by public funding, is largely the result of strategic or policy decisions made at the national government level. On the one hand, zero-cost models can be a strategic decision to reduce barriers to data access and support the growth of an Open Data ecosystem around the portal. This is the rationale behind Luxembourg's cost model, where the government actors responsible for Open Data strategy and portal operation are seen as highly important political and strategic priorities. Alternatively, the French and Belgian survey respondents highlighted that the production of the data that is shared by the government on the portal has already been funded by taxpayers: they have already 'paid' for this data, and therefore should be able to freely access it on the portal.

Where the provision of free Open Data is a strategic policy decision on behalf of the government, funding is often provided through instruments like national budgets. This is the case in Slovenia, where the Slovenian national portal[114] owner commented that their funding directly came from the national budget, meaning the sustainability of the portal was based on national decisions.

A significant minority of portals (40%) expressed that they rely on a legal mandate or framework to provide funding for a zero-cost model, including the Czech Republic, France, Latvia and Spain. This legal mandate is often imposed at the national level, such as the French right of access to administrative documents, but may also be an EU-level order like the Public Service Information (PSI) directive, as in the case of Latvia. Such directives and frameworks are a way of providing a significant guarantee of long-term public funding that provide reassurance for portal teams on their long-term financial sustainability.

3.1.2. Funding local and community-run portals

In comparison to national portals, local and community-run portals may receive a lower level of public funding. In response to these disparities, some portals are sourcing funding from multiple sources.

Public funding is still offered to some extent at a local and community level: 100% of the local and community-run portals in our survey received some level of funding from local government. However, the level of funding and manner of distribution varied considerably between cases, and may be distributed from national, regional or local government bodies. This may be due to funding being

---

[113] National Open Data Portal Russia. Available at: http://data.gov.ru.
[114] National Open Data Portal Slovenia. Available at: https://podatki.gov.si/.

spread across a number of Open Data portals at the regional and local level, compared to the requirements of only funding a singular portal at the national level.

In one case, the capital city's Open Data portal received a higher sum of money for portal development than the federal portal, despite funding for both portals being distributed from the same source. In Italy, the Municipality of Florence Data Portal[115] uses municipal funding to cover a number of its essential activities, such as maintaining the Open Data Portal, as well as supervising data exchange among different bodies within the government.

Other cities are adopting more innovative models that share the financial cost for portals among several public bodies in order to provide free Open Data. This is true of Helsinki's Open Data portal, the Helsinki Region Infoshare[116], which is jointly owned by the four municipalities of Helsinki, Vantaa, Espoo and Kauniainen. These bodies fund and provide data to the Helsinki portal, and govern the portal through a joint steering group which meets regularly to make strategic decisions. Each city is represented in this governing body, as well as a representative from Citra, a Finnish innovation fund. The interviewee shared that this model was extremely successful and facilitated cooperation in the region. Given the success of this portal, portals may consider replicating this model of shared portal costs and governance in other contexts.

Community-led portals often have to combine public funding from local governments with funding from other sources in order to provide data on a zero-cost model. For instance, Bath:Hacked[117] in the UK has a small number of organisations who provide direct financial support to the portal. The respondent noted that this was preferential to adopting a model that monitored and charged for individual usage as it lowered administrative costs for the portal team.

Similar to national portals, local and community-run portals are largely supported by government funding in order to achieve policy, strategic and legal goals, although to a lesser degree than their national counterparts.

In terms of policy and strategic goals, using government funding to sustain local Open Data portals may be seen as a necessary investment that provides long-term social benefit. This may include encouraging Open Data use and impact to lower costs for the government in the long-term. This is the case in Vienna[118], where an interviewee disclosed that the value of Open Data is exceeding the public funding provided to sustain a zero-cost model. In Trentino[119] and Barcelona[120], this model is proving beneficial to both local governments and citizens: in both areas, the low barrier of access to data is stimulating external innovation and impact that could not be carried out by the local governments itself. It is expected that these innovations will lower costs for local governments in these areas, and therefore may decrease the need for public funding to sustain the portal in the future.

National legal mandates may also extend to mandating zero-cost models and funding of regional, city and local portals. However, the character of many of these funding relationships is unclear and may lead to a lower level of sustained funding for this level of portals in the long term. One survey

---

[115] Municipality of Florence Open Data Portal. Available at: http://en.comune.fi.it.

[116] Helsinki Region Infoshare. Available at: http://www.hri.fi/fi/.

[117] Open Data Portal Bath: Hacked. Available at: https://www.bathhacked.org.

[118] City Open Data Portal Vienna. Available at: https://www.wien.gv.at.

[119] Open Data Portal Trentino. Available at: http://dati.trentino.it.

[120] Open Data Barcelona. Available at: http://lameva.barcelona.cat/ca/.

respondent claimed that the level of funding received was not sustainable in the long-term as 'to 'boost a data-driven economy' would require more resources'.

3.1.2.1.   Lessons and best practice

When it comes to the question of financial sustainability, the large majority of portals currently view their funding as sustainable. Our survey indicated that 94% of respondents felt confident in the sustainability of their portal, with most justifying that Open Data was beneficial to society (21%) and there were clear political interests for maintaining the portal (21%).

The financial sustainability of portals appears to be most apparent where it is supported by a legal mandate. A legal incentive to Open Data safeguards the portal from short-term changes in funding, in comparison to yearly government budgets where the financial commitment to Open Data may fluctuate. Consequently, where possible, countries should consider expanding the legal incentives for the provision of portal funding and enshrine commitment to Open Data within law.

At a national level, alternative funding approaches such as public procurement may result in bureaucratic procedures that result in discontinued or short-term funding for portal teams. This can endanger the funding of long-term projects that are necessary for building a portal's maturity. Where the ongoing management or funding of a portal may switch hands due to a public procurement approach or other switch in management, portals should invest time in creating strong knowledge management procedures around portal maintenance.

At the regional, local and community level, there are clear benefits for local government bodies who have a shared interest or overlapping data about an area to jointly fund and maintain a portal. This is likely to lower operational costs for the portal and enable more investment in innovation, strengthening sub-national portals which may otherwise receive little or inconsistent funding from the national level.

Recommendations for portals

- Local government bodies should consider jointly funding a portal where there is shared interest or overlapping data about an area, in order to lower operational costs and enable investment in additional portal activities
- Invest in strong knowledge management procedures where the funding or management of a portal is likely to change hands

---

3.2.   Understanding the costs of Open Data portals

For portal owners, finding a sustainable source of financial support is crucial, as funding will determine the scope of costs they can bear. This will shape activities and future development of portals.

Portals must consider costs around keeping up the delivery of their activities, such as harvesting public sector data, motivating publishers to provide more datasets and building understanding behind these actions. Thus, portals must fund costs in three key areas:

- Maintenance
- Encouraging publishing
- Encouraging awareness and innovation

3.2.1. Maintaining portal infrastructure

The cost of maintaining the technical infrastructure of a portal is a fundamental cost for the survival of every portal. As the main purpose of Open Data portals is to serve as a web-based platform for the discovery of Open Data, portals that are unable to meet this initial cost of setting up and maintaining the portal infrastructure are unlikely to be sustainable in the long-term[121]. Additionally, if a portal is going to engage in any significant growth in providing additional services, it is critical that the costs of maintaining a portal are not too high, in order to allow financial resources to be spent in other areas.

Portals may consider two main options when it comes to the technical maintenance of portals: hosting by a third party, or the use of open source software. Each has associated benefits and costs.

3.2.1.1. Third party hosted solutions

Portals may opt to use a hosted solution for their portal, where the technical infrastructure of a portal is maintained by a third party for a fee. Third parties may offer hosted services using their own specialist software, such as Socrata and OpenDataSoft, or based around open source software, such as Datapress and Vederum (which offer services around CKAN).

The third party hosted solution model is used by several European portals, mainly on the regional and local level. For instance, the Bath:Hacked community portal in the UK uses Socrata. Another solution is the French OpenDataSoft platform, used by portals in Toulouse, Paris and Brussels and recently launched a free version of their software to support 500 city portals in the United States[122].

The costs associated with hosted solutions vary by provider. For example, Open Data Soft provides free unlimited use for both data publishers and users believing that allowing more people to use a dataset improves its quality. The provider placed importance on the volume of datasets available through the portal and usage of the portal. Instead, they price their solutions at a flat rate, with additional cost for use of additional features. One such feature is the sub-domain feature, which allows the creation of links between county and city level portals. Portals are also charged if they opt for an alternative backend provider to Amazon Web Services (to which portals are automatically linked). Portals who are required to comply with data localisation legislation may therefore be required to pay additional costs to use these solutions.

Alternatively, the interviewee at platform provider Urban Tide said they used a freemium model for publishers to encourage portals to explore using their platform while focusing primarily on real-time. They allow users to publish unlimited CVSs for free, however they limit the platform to only 3 publishers and do not provide for APIs. Portal owners then have to pay to add additional publishers and for the use of APIs. There are additional fees to support real-time data and for additional features such as analysis tools.

Third party hosted solutions provide portals with a number of benefits, particularly around lowering costs. As these services are designed to be user-friendly, they do not require a high level of specialist skills and are relatively easy to set up. Consequently, portals are likely to have lower costs associated with the setup and development of a portal. Portal teams also benefit from a single known cost in

---

[121] European Commission (2017). Open Data Portals. Available at: https://ec.europa.eu/digital-single-market/en/policies/open-data.

[122] Government Technology (2017). OpenDataSoft Targets Mid-Sized Cities with 500 New City Portals. Available at: http://www.govtech.com/data/OpenDataSoft-Targets-Mid-Sized-Cities-with-500-New-City-Portals.html.

being able to effectively forecast a budget for the cost of technical hosting for set periods of time. Lastly, these services often offer a large number of features for portals as part of the core offer: for instance, OpenDataSoft provides APIs over tabular datasets, whilst Socrata offer visualization tools for their users.

However, portals using third party hosted solutions may also be limited in their flexibility and independence. In particular, there may be a limited number of solutions or features that a portal can access, or portals may require the cooperation of the third party in order to implement any technical changes. For instance, the use of a hosted solution has prevented one of the survey respondents from implementing necessary changes, as the portal provider owns the platform and user experience. Furthermore, in the long-term, this limited independence may result in a dependency on the tools and APIs of external suppliers. In terms of financial sustainability, this reliance on one external company in a limited marketplace of suppliers, may risk portals becoming somewhat dependent on the goodwill of hosting companies not to raise their prices.

Consequently, use of third party solutions offers a number of benefits for portals in lowering costs, but must be balanced with the negative impact of some limited flexibility and independence. Indeed, the representative of Florence's data portal commented that outsourcing is not always the right solution because in order to control the outsourced process you need to understand the technologies you are procuring so to some extent you need to have the skills internally anyway.

3.2.1.2. Open source software

On the other hand, portals may opt to use open source software directly such as CKAN and DKAN[123]. These are Open Data tools with cataloguing, publishing and visualisation features designed for governments to easily publish Open Data. The CKAN software is widely used by different national and regional government portals and private companies, including the portals of France, Amsterdam, Berlin, Estonia, Zagreb, Wallonia (Belgium) and the Helsinki Region Infoshare, whilst DKAN is used in the Czech Republic and by cities and regions in Germany[124]. A European Data Portal report found that all national portals they surveyed adopted CKAN, whereas regional portals tended to use a mix of Socrata (third party host) and CKAN[125].

The setup costs associated with using open source software for portals are relatively low. By its nature, this software is available for free, meaning the initial cost of launch may be limited. The cost of the actual technical infrastructure for hosting an instance of an open source Open Data Portal is relatively low due to the falling costs of commercial cloud storage from providers such as Amazon Web Services. Variously, costs were estimated between EUR 60,000 - 100,000 among interviewees. The representative from Luxembourg noted that these services were also easily scalable when they need to bring in large amounts of data.

Furthermore, open source software allows Open Data portals to benefit from the development work done by other platforms. For instance, as mentioned in Part 1, the popular 'Showcase' module for CKAN allows owners to register visualizations, blog posts, journal articles and papers that use datasets in the portal.

[123] DKAN. Available at: http://demo.getdkan.com.
[124] CKAN. Available at: https://ckan.org/about/instances/.
[125] European Commission (2017). Recommendation for Open Data Portals: from setup to sustainability. Available at: https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

However, portals using open source software are likely to have more pressing and complex considerations for the costs around training and staffing in the long-term than their hosted counterparts. This is because skilled people are required in order to deploy and maintain this software. These individuals will require higher salaries and it may be difficult for public sector portals to compete with the market. For instance, a respondent shared that Helsinki and other Finnish city portals have seen difficulties in recruiting individuals with the right skill sets. The Helsinki portal are currently working with 8 highly skilled code fellows on the technical side to release data; clearly, recruiting this expertise also comes with associated costs. Likewise, the Florence portal team have favoured spending on internal budget for those with the skills to personalise open source platforms, such as geonetwork and geoserver, instead of using this funding to cover the costs of license agreements or commercial products. Consequently, the choice of open source software may have personnel costs that reduce the ability of portals to run activities or cover costs in other areas.

It should be noted that the trend of open source software leading to higher personnel and time costs is not true across all Open Data portals. For instance, the Paris Open Data portal states that it 'mobilizes 15% of the time 3 people', as well as the support of an apprentice; this appears to be a relatively low investment in terms of time and resources[126]. However, it should be noted that the project also relies on a network of people 'for which Open Data is not the main task', meaning that the costs in time and staffing may in fact be higher. The difficulties of building engagement and effort beyond the central portal team are explored further in the next section.

Secondly, there may be implications beyond just the technical maintenance of portals. The fact that open source platforms can be less user-friendly may mean investing in training for core portal teams and teams more widely in government. This has implications for costs around budgets for engagement and training, which portal teams may struggle to cover with public funding.

Thirdly, the independence from any formal portal provider has some issues around the portal's ability to forecast future costs. Portals using open source software may be subjected to unexpected issues and the costs associated with them; these costs would likely be absorbed by the portal providers of hosted platforms. Consequently, portal teams using open source software may find it more difficult to accurately plan the overall costs of developing and maintaining the portal.

3.2.1.3.   How technical costs differ between national, regional and local portals

The proportion of cost for a portal's infrastructure may vary between national, city and local portals. The difference between funding models for portals at the national level compared to city and local portals means that portal maintenance may form a higher proportion of technical cost for regional and local portals.

Many of the national portals we interviewed viewed their technical costs as relatively low, including Belgium and Luxembourg. Additionally, some regional and city portals, including the Helsinki Region Infoshare and Florence municipal portal, receive significant amounts of funding that cover the costs of the annual portal maintenance. As a result of this low cost, which is not usually subject to change, many portals view their model as financially sustainable.

On the other hand, many regional and local portals are using open source software that, as previously explored, can come with unexpected costs. The lower level of public funding for some of these portals

---

[126] Open Data Paris. Available at: https://opendata.paris.fr/page/faq/.

may leave some portals unable to easily absorb these unexpected costs, and thus, less financially sustainable.

Beyond just the cost of portal maintenance and hosting, portals at both a national and sub-national level have technical costs associated with the upkeep of features on their platforms using open source software. In Spain, although a legal mandate means it is easy to secure money for ongoing maintenance of the portal, the national portal team still have a limited amount of money for portal features that must be prioritised based on demand. At present, activities are prioritised by demand received from users and publishers: this has led to the creation of dashboards for publishers to track access and use metrics for the datasets they publish, as well as user-facing features such as improving search functionality based on user feedback. This approach of prioritising features based on limited funding is also adopted by the Municipality of Florence Data Portal, whose representative expressed that more funding would be needed to develop more operations on the platform. Hosted solutions, as previously mentioned, gain additional features as they are developed by the hosting companies. While portal providers might offer to build bespoke features for portals, these are likely to be expensive and the portals will not retain the features if they shift supplier.

While prioritising services based on the needs and demands of users is a good way to develop platforms, this approach does not allow portals to strategically plan and spend in ways that may better improve the portal's long-term sustainability. For instance, whilst providing visualisation tools may be a way to respond to current user demand, it may prevent investing in improving search architecture, which would allow more people to access the data in the long-term.

However, it is important to note that financial concerns are not always a limiting factor for portal maintenance and development. 15% of portal owners said that although they were able to fund additional features, such as FAQ sections or newsletters, these were not seen as necessary or desirable. This is the case with Belgium's Federal Open Data portal: a representative expressed that whilst the team had sufficient funding to add some features, such as a newsletter and contact details, they felt it was unlikely to produce a sufficient return on investment.

3.2.1.4. Lessons and best practice

The experiences of portals indicate a number of lessons for financial sustainability.

Firstly, both open source and third party hosted portals have advantages and disadvantages when it comes to cost. While the distribution of primary cost varies - with hosted solutions, the main cost is the fee paid to the third party, whilst open source solutions see a primary cost of technical staff - it is important to note that neither option appears to be significantly more expensive in terms of overall cost.

In light of this, the priority for hosted portals should be enacting strong procurement processes that look at all the available options and avoid vendor lock-in. For portal teams wishing to host their own CKAN or DKAN instance, the priority cost should be investing in skilled technical staff. Although these individuals can be recruited, it is often difficult to attract people with the right skills and compete with the market rate salary for the skills. Portals can seek to mitigate these issues by upskilling existing staff through training to administer and maintain these platforms.

Secondly, some cities and regions are adopting a model where the costs of hosting and developing platforms are shared by several different bodies. One method to deliver this is to partner with organisations which can provide the technical expertise necessary to deliver new features and

functions for platforms; a prominent example is the Code Fellowships offered by the Helsinki Regional Infoshare.

Another method, also adopted by the Helsinki portal, is the practice of partnering to create platforms. Helsinki's platform is jointly owned and funded by the four municipalities of Helsinki, Vantaa, Espoo and Kauniainen. As discussed earlier, this joint portal ownership has implications for the way the platform is governed. Partnerships around portal costs can be agreed at the point of platform set up, or at some point during the lifespan of the portal. For example, the Leeds Data Mill, a municipal Open Data platform for the city of Leeds in the UK, recently expanded to cover the city of Bradford. Following this shift, it changed its name to Data Mill North and now receives funding from both municipal governments.

Finally, portals can look to share the costs of launching and developing open source platforms by approaching existing platforms to help them implement them. For instance, Luxembourg drew on the technical architecture of the French national Open Data Portal when setting up their own portal, which lowered the barriers they faced in setup.

Recommendations for portals

- Choose either a third party hosted solution or build an instance of a popular open source platform such as CKAN or DKAN.
  - If choosing a hosted portal, ensure strong procurement processes are in place to identify the best suppliers.
  - If choosing to host an open source platform, ensure there are sufficient levels of technical knowledge on the team, either through hiring or upskilling existing staff.
- Look at ways to share the cost burden of technical hosting by partnering with other portals or governments to develop features, partnering with other governments to share a platform or bringing in technical expertise through fellowship programmes and other partnerships.
  - Best practice examples: Helsinki, Luxembourg / France

3.2.2. Engaging and encouraging publishers

Apart from building and maintaining an Open Data platform, portal teams also play a key role in driving forward the Open Data agenda within government. This involves encouraging and coordinating the publication of data across a wide variety of publishers within and beyond government. However, there are costs associated with building these relationships, which can be expensive and must be factored into a portal's budget.

Responses from our survey indicated that there are clear differences in the nature of relationships built by national portals and those built by regional and local portals.

At a national level, 82% of survey respondents explicitly mentioned that they built relationships across all national and federal organisations, and 27% that they built relationships with local level bodies. No national portals explicitly mentioned building relationships with international organisations.

Comparatively, the large majority (88%) of regional and local portal survey respondents explicitly mentioned building relationships with local and regional government, with only half of local portals building relationships with national bodies. Interesting, 20% of these portals mentioned building international relationships.

The disparity in these figures may represent the difference in the number of stakeholders to engage at each level. National portals generally face a larger task compared to local portals, as the number of data publishers in central government are much higher. For example, France's national portal has 1258 listed data publishers,[127] whereas Paris's Open Data portal has only 31.[128] Indeed, the representative from community portal Bath:Hacked commented that they experience lower transactional costs by only needing to engage with a smaller number of organisations.

However, sub-national portals still need to build other types of relationships, especially with their data providers and other local or regional portals. The Municipality of Florence portal has built a number of formal partnerships: the team are currently collaborating with Digital Italy Agency (AgID) and the team for digital transformation of Italy in Rome to design a data education framework, which illustrates the diversity of city Open Data portals in metropolitan areas across Italy. They are also collaborating with these bodies to produce common standards for some datasets around air quality and museums. The portal also has partnerships with the Tuscany regional portal, the University of Florence, the city's Chamber of Commerce, and the public utilities associations.

For sub-national portals in particular, building these relationships at all levels of government is important for improving data quality and accessibility. As highlighted in the EDP's Open Data Cities 2 report, many of the highest priority datasets that have not yet been opened in Florence, such as pharmacy and street cleaning data, do not originate from the municipality of Florence itself but from external authorities. Overcoming these organisational barriers requires significant investment from local portals in building productive local relationships[129].

3.2.2.1. Directly engaging publishers

The main way that portals currently interact with publishers is through establishing durable relationships. Portal teams form and maintain these relationships through providing guidance and training on how to release Open Data, as well as continuous contact with relevant individuals and teams.

This is the case in the UK, where the data portal owner explained that the portal team maintained ongoing and regular contact with individual publishers who used the national portal. These relationships include providing advice around publishing datasets to EU standards, as well as working out how to harvest data from the publishers with their own portals. In France, Etalab which runs the Open Data Portal is a member of the 'General Secretary for the public action modernization' whose aim is to spread the culture of Open Data, open government and data science throughout government. This means they work in a transversal way with all the institutions.

Another example of successful engagement with publishers is Luxembourg. Since the portal's launch in 2016, the portal team has invested significant time in building relationships with geodata producers and has managed to achieve some quick wins. The team are now working closely with all the national ministries and their dependent services to produce a complete assessment of existing and publishable datasets. Fundamentally, the goal is secure a collaborative relationship with each government function and dependent services within each of the national ministries.

---

[127] Data.gouv.fr. Available at: https://www.data.gouv.fr/en/organizations/
[128] Open Data Paris. Available at: https://opendata.paris.fr/explore/?sort=modified.
[129] European Data Portal (2017). Analytical Report 6. Available at:
.https://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n6_-_open_data_in_cities_2_-_final-clean.pdf.

These types of relationships require different levels of engagement depending on the level of Open Data maturity within different departments. The representative of Spain highlighted the challenge of convincing high level civil servants about the best way to publish data. This is why Spain has adopted a very intensive approach to engaging publishers which is explored in the box below.

> **Spain's approach to engaging publishers**
>
> The Spanish national portal, the Spanish National Open Data Catalogue, has adopted a particular model to engage publishers from across government and create partnerships at all levels.
>
> In the Spanish model, every ministry and local administration is assigned a coordinator by the portal. This individual is provided with training and support and helps coordinate various public bodies within regular inter-institutional meetings and multi-stakeholder working groups.
>
> A representative from the portal stressed the importance of this method in building close relationships with different bodies. This was especially beneficial for connecting with bodies who had been working on the Open Data initiative for a longer period of time, and could provide the portal team with opportunities to broaden their network.
>
> The portal team also run focus groups and events with regional and local public servants at least three times a year. The interviewee shared that this helps understand the data needs of these actors, as well as providing an opportunity to present the benefits of Open Data and share new developments regarding Open Data policies.

In total, 30% of portals interviewed, including Spain and Luxembourg, have made efforts to secure a named contact in each government department. This approach has proved very effective according to these participants in encouraging publishers, contributing to the fact that Spain and Luxembourg achieve very high rankings in EDP's 2017 maturity report.[130] However, maintaining one to one relationships with contacts across government can be very resource intensive for portal teams, requiring them to reach out to time limited civil servants and provide bespoke advice.

3.2.2.2. Lowering the cost of engagement

Because of the high cost of maintaining these relationships, portals have made efforts to provide support for publishers in a less bespoke way. One of the key approaches has been to bring together publishers to help them publish their Open Data to a high standard. For example, the UK portal representative mentioned they hold workshops, research sessions and training for publishers. The Czech portal also organises workshops on Open Data and Linked Data. The representative of the Helsinki portal similarly emphasised the importance of their training for publishers.

However, hosting these kinds of events also comes at a cost – in particular through hiring or procuring Open Data trainers. In the EDP 2017 Open Data Maturity in Europe index, many Member States highlighted that financial concerns were preventing the provision of training to increase the number of qualified Open Data personnel. Many felt unable to meet the costs of training, which is 'still needed

---

[130] European Commission (2017). Open Data Maturity in Europe 2017. Available at: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

in many administrations where qualified Open Data personnel is still scarce'.[131] One lower cost method of mass engagement can be the use of newsletters, which Spain uses to communicate with publishers and solicit feedback from.

Other methods employed to encourage publishers include the development of specific features for publishers to help incentivise publication. In particular these efforts provide tools to publishers to engage them particularly through tracking publication as discussed in Part 1. One example of this is in Spain, where the portal has developed dashboards for publishers to track access and use metrics for the datasets they publish, as mentioned earlier in this report. They did this on the basis of feedback they received from publishers.

Similarly, the representative of the UK portal said they are carrying out research into the needs of publishers and how to best meet those needs. This approach led them to create a feature to develop reports for publishers based on the case studies that have been published. The UK's national portal uses these links to generate usage reports for publishers about how their data is being used. This approach has led some publishers to use the platform's reports about their data as their own internal metrics and evidence for publishing Open Data

Likewise, the Russian national portal publishes rankings of all the publishers on a quarterly basis on how well they are publishing[132]. The goal of this is to get publishers to compete for better rankings incentivising them to be more engaged and conform to best practices. All of these approaches attempt to shift the costs of direct engagement to a more self-service model, with a variety of achievements.

3.2.2.3. Lessons and best practice

Engaging publishers and coordinating the publication of data is a key function of portal teams beyond just providing a platform to access the data. All portals have this responsibility however it is a much greater challenge for national portals than local portals. The primary method for doing this is through cultivating one to one relationships with responsible people in each government department. This can be a very costly activity in terms of resource however the impact on sustainability of Open Data initiatives is evidenced in the maturity of the initiatives that have achieved it - including Spain and Luxembourg.

Because these activities are very costly, portals should explore methods for engaging publishers in a more standardised and coordinated fashion, for example by organising training events for publishers. They can also look to build functions into the portal that meet the needs of publishers and keep them engaged. These methods include things like dashboards, usage reports and rankings that allow publishers to understand how well they are doing and identify best practices without direct contact with the portal team.

> **Recommendations for portals**
>
> 🌐 National portals should aim to have individual named contacts within each government department, if resource permits.
>   - Best practice example(s): Spain, Luxembourg
> 🌐 Portals should aim to reduce costs by engaging multiple publishers through the provision of

[131] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.
[132] Analytical Center for the the Government of the Russian Federation. Available at: http://ac.gov.ru/en/publications/.

> workshops and training.
>   o Best practice example(s): the UK, Czechia, Helsinki
> 🌐 Portals should explore the use of self-service tools that help engage publishers such as dashboards, usage reports and rankings.
>   o Best practice example(s): Spain, the UK, Russia

3.2.3. Building awareness, engagement and innovation

Portals must also consider the significant cost of building awareness, engagement and innovation, both within and beyond government. These costs go further than publishing data to encouraging and demonstrating the long-term impact that many are tracking using the methods outlined in Part 1. Awareness itself is the highest barrier to Open Data re-use, affecting 64% of cities in the EU28 countries[133]. Without awareness among prospective Open Data users of the benefits of Open Data, and how to publish, discover and use Open Data, an Open Data initiative is unlikely to survive in the long-term.

Building awareness, engagement and innovation is an important aspect for many portals as it is linked to their rationale for providing Open Data on a marginal or zero-cost model. As explored earlier, cities/regions like Trentino and Barcelona are investing in public data in order to produce external innovation and impact that could not be carried out by government itself. However, due to budget limitations, many portals are not able to ringfence funds for a specific engagement budget, meaning that these activities may be slightly sporadic in nature.

A notable approach to building awareness beyond government is that adopted by the French national portal. Run by French organisation Etalab, the portal's budget was limited to funding for the website and not to redesign French information systems. In response to this, the Etalab team engaged Open Data activists to 'organize brainstorming sessions [and] hackathons...to come up with what data.gouv.fr should be'[134]. The team also consciously chose to allow anyone to post data, as a way to encourage data publication from actors beyond government in civil society. With this flexible approach engaging actors beyond government, the French portal has managed to grow the data.gouv.fr portal whilst operating on a limited budget.

A more common approach to encouraging innovation, hackathons, has been adopted by portals in Luxembourg and Russia. In Luxembourg, the interviewee shared that hackathons are a useful way of building partnerships with universities and encouraging these bodies to think strategically about the data they hold and how it may be beneficial to users. Events like the gameofcode.eu hackathon, which has been held twice, also provided the opportunity to apply some pressure on public bodies to release additional Open Data for the event. Similarly, in Russia, hackathons have been used to engage and build bridges between Open Data users and publishers; the portal even owns a space which is used by ministries and agencies to host these events.

---

[133] European Data Portal (2017). Analytical Report 6. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n6_-_open_data_in_cities_2_-_final-clean.pdf.
[134] TechCrunch (2014). How France's Open Data Team Is Modernizing The French Government Through Data. Available at: https://techcrunch.com/2014/02/12/how-frances-open-data-team-is-modernizing-the-french-government-through-data/.

Portals may also use funding to engage the community beyond government in the form of training and public awareness campaigns. The most developed of these is Finland's 6aika.fi initiative, which spans six Finnish cities and encourages cooperation around issues concerning Open Data, open APIs and innovation[135] . Each city receives extra funding and coding expertise as a result of this project - in Helsinki, the portal has used this funding to provide training to companies and enhance their awareness of Open Data and the portal. Furthermore, in Portugal, low public awareness has prompted the portal to prepare a series of awareness campaigns around the benefits of Open Data and where it can be found[136]. On a lower level, a number of portals also issue regular newsletters and features to engage the community beyond government. These methods are relatively low cost; however, it is likely that the primary audience for these communications is already highly engaged in the Open Data initiative, and so the rewards from this form of engagement are not quite as wide-ranging as other more expensive forms.

An approach for reducing the costs associated with innovation is forming partnerships with other organisations, such as universities. 43% of interviewees mentioned that they were partnering with universities to raise awareness around Open Data. This includes Luxembourg, where the national Open Data Portal representative commented that they were very keen on building relationships with universities or institutions, and were constantly thinking about hackathons or events they could organise with students, teachers and professors around their data. These findings echo those of the EDP Open Data and Cities report, published in June 2017, which found that 5 out of 8 cities analysed in the report, namely Dublin, Ghent, Florence, Lisbon and Helsinki, have set up partnerships with universities to reach a wider audience. The representative of Helsinki's portal said they have very good cooperation with the technical university, including on one particular course which focuses on the students using Open Data in their work.

One other way to nurture relationships can also be through investing. The Municipality of the Florence Open Data Portal showed that investing in new features to add on top of their portal was also a form of relationship building with universities. Indeed, its representative disclosed that they invested some new funds in the big data area, and that they started with Open Data and now leveraged skills from local universities in big data tools.

Lastly, the Florence portal is collaborating with a local association in the city to map public spaces for citizens. In the future, the portal plans to launch a public call for associations to continue this work. Indeed, as discussed in Part 1, Florence has also partnered with a university to do research into the usefulness of Open Data and how utilities are using Open Data.

3.2.3.1.   Lessons and best practice

Where budget permits, it is clearly beneficial for portals to invest in engagement, awareness and innovation beyond government. Indeed, this engagement forms an important part of the mission of many portals to encourage innovation and impact with Open Data in wider society.

Hackathons are a useful format for encouraging innovation. As they are a widespread format, portals running hackathons for the first time can draw learnings from the myriad that have taken place across Europe over the last few years, and they are likely to attract a significant level of attendance. Furthermore, hackathons can form a useful way to gently pressure publishers to reconsider the percentage of their datasets which can be made open, and be a way to build bridges between users

---

[135] 6aika.fi. Available at: https://6aika.fi.
[136] European Commission (2017). Open Data Maturity in Europe 2017. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

and publishers that make the Open Data initiative more sustainable in the long-term. This has been seen in places like Luxembourg and Russia.

As with sharing the costs of portal maintenance, countries should also consider how investment in a shared innovation strategy could reduce the costs of engagement for all the portals involved. In Helsinki, this shared city strategy is enabling cities to engage wider communities by reducing the level of cost associated with activities like training.

Finally, portals should also consider forming partnerships with organisations beyond government. Portals who are forming partnerships with universities, and investing in joint ventures like hackathons, events and university courses, is an excellent way of both engaging wider communities whilst upskilling the next generation of Open Data users in coding and other much needed expertise.

Recommendations for portals

- Invest in events like hackathons as a method for both encouraging innovation among re-users and greater publishing of Open Data by government ministries
- Look for opportunities to create shared innovation strategies and funds across cities and regions
- Form partnerships with organisations beyond government, such as universities, and invest in joint ventures that may produce long-term impact

---

3.3. Alternative funding models and approaches

As stated earlier, 94.4% of survey respondents believe their current funding model is sustainable. However, across part 2, we have highlighted that in order for portals to become mature and remain sustainable, there are lots of additional costs beyond just hosting the technical platform.

The potentially hidden costs of sustainable portals in hiring technical staff, engaging publishers and running outreach events are significant. As such, portals might want to look at other alternative approaches to supplementing their existing funding beyond the measures to share costs discussed.

One alternative funding approach is to sell data. Instead of opting for a zero-cost model, this model generates revenue through the sale of access to, or the license to re-use, data. One interviewee mentioned that the drive to open more data in some parts of government was hampered by existing financial incentives for those departments already selling data, particularly in those who possess spatial and weather data. In these cases, the data is used to finance the ongoing collection and management of the data, as well as the research and studies of their staff and often support for universities. Thus, the sale of data provides a source of revenue for these departments, and they are reluctant to adopt a model that risks the loss of this venue.

One potential solution to this approach is directly funding these organisations to open up datasets. This model has been adopted for the data of geospatial organisation Ordnance Survey in the UK. However, these models do not appear to be sustainable in the long term.

Very few portals we spoke to had considered any form of alternative funding model, with two exceptions. In the case of Luxembourg's national Open Data portal, the interviewee disclosed that this was due to satisfaction with the portal's funding. However, importantly, they did add that they were

looking at ways the private sector could host data on the site, which could provide another source of funding.

Likewise, the Russian Open Data portal had theoretically considered two funding models to make their portal more sustainable, as the current method of public procurement was leading to gaps in development. First would be a system of public-private partnerships, where the government delegates all operational obligations to private companies for operating, supporting and developing the portal in exchange for private companies commercialising aspects of the service. Alternatively, they believed the portal could be owned through a government investment fund, and likewise win revenue through commercialising aspects of the service.

3.3.1. Adopting a freemium model

One common approach to providing non paid-for services is the use of freemium models. Freemium models are business models where access to basic functions are provided for free, but access to additional functions are paid for. These models are prevalent among digital 'as-a-service' businesses including portal platform providers.

In particular these models have been adopted by private sector organisations who are using them to derive revenue from publishing Open Data. It is possible to keep data open because everyone has access under an open licence. We did not find any examples of public sector Open Data portals in Europe who had deployed these models yet. However, there are a number of private and third sector Open Data providers who have for example, OpenCorporates,[137] TransportAPI[138] and Traveline[139] all use freemium models over Open Data that they publish. In each of these models Open Data is published for free and paid-tiers are only applied to additional features such as analysis of the data or for higher speeds of API datasets uploads and downloads, or unlimited number of downloads per day. A list of examples of the types of freemium mechanisms that have been proposed for Open Data publishing is provided below.

The benefit of adopting freemium models is that the data itself remains openly available. Users who access the data frequently are charged because they place a greater strain on the technical infrastructure. It is also important to note that these users are likely to only be a small percentage of total users and are often using the data for commercial reasons. Another benefit is that it potentially allows revenue to scale as the cost of providing the infrastructure rises, as only high-volume users are charged.

Freemium models are however limited by the types of data, for example it is very hard to instigate a freemium model for static, historical data. Most data must be provided by API and have a degree of frequency of update - primarily it is applied to real-time data such as transport Open Data. In addition, applying a freemium model will likely stifle some of the benefits of Open Data being provided free of charge. To mitigate against this fact, it would be best to place freemium restrictions on data sets which are currently only available as paid-for data. This would allow more users to experiment and innovate with that data than is currently possible, as they do not need to pay upfront, while still retaining the revenue from users who rely on the data for commercial reasons.

---

[137] OpenCorporates. Available at: https://opencorporates.com/api_accounts/new.
[138] TransportAPI. Available at: https://www.transportapi.com/plans/.
[139] Traveline. Available at: http://www.travelinedata.org.uk/about-traveline/.

Examples of
freemium mechanisms associated
with Open Data publishing

By volume or rate

free tiers may offer access to a set volume of data per day, while paid tiers provide higher rates or no-limits

By selective licence

free tiers might offer data under a sharealike licence, while paid tiers might allow customers to keep their derived data closed

By level of service

free tiers might offer no technical support or guarantee on availability, while paid tiers provide service level agreements and support

By time period covered

free tiers might offer access only to historic data, while paid tiers offer more recent data, or vice versa

By geographic area covered

free tiers might provide access to particular geospatial areas, while paid tiers provide data that covers a wider area

By granularity

free tiers might give access to aggregated data over large areas, time periods or other classifications, while paid tiers provide data at a more granular level

By frequency of update

free tiers might offer delayed access to data, while paid tiers get data in real-time

By access method

free tiers might offer a download of tabular data, while charging for access to an API

By subset

free tiers may include only a subset of the entire dataset, while paid tiers might offer additional fields or API methods

3.3.2. Lessons and best practice

There is clear value to exploring alternative funding models if they provide opportunities to make more Open Data open, and make portals more financially sustainable. In terms of particular alternative models, portals should explore how to adopt a freemium model to datasets that are not yet open.

Recommendation for portals

- Explore how to adopt a freemium model to datasets that are not yet open

4. Conclusions: Embedding sustainability into portal initiatives

Open Data portals are a critical part of our data infrastructure that connects publishers with data users. These data users create services that citizens and businesses benefit from and increasingly rely on. To ensure portals are fit for purpose, now and in future, sustainability must be considered and embedded into the operation and planning of Europe's Open Data portals.

Following the recommendations from our first report on portal sustainability, this paper has examined key issues around demonstrating use and impact and building a sustainable environment, especially in funding. Addressing these fundamental aspects of sustainability is essential for building a portal that is responsive and adaptive to the changing technological and political environment around it.

This report has indicated that portals are still at the beginning of considering questions around impact and financial sustainability. Whilst a significant portion of portals are exploring how to measure and track use and impact – the EDP's Open Data Maturity in Europe index on impact of Open Data shows truly remarkable progress from 31% in 2015 and 44% in 2016 to 54% in 2017 – few clear standards and methodologies have emerged. This is supported by our findings that active tracking of impact is very varied amongst portals. This is likely due to the fact that many countries have still been focused on internal Open Data policies and use, rather than impact. Likewise, financial sustainability is clearly important to portals, and many have reviewed their funding strategy; however, with many portals still dependent on public funding, it is important for portals to consider other options.

We have discovered nascent best practice around impact and sustainability that is becoming more widely adopted. The peer-learning approach of EDP encourages portals to look to one another and adopt techniques that have been successful in other contexts. In this spirit, this conclusion reviews each of the recommendations we issue to portals and points to several examples of European (or worldwide) best practices for inspiration.

With the EDP finding an average of 76% maturity in the EU28 countries, portals across Europe are now increasingly providing a range of high quality data, underpinned by good publishing practices within government. Now, portals at all levels of government must consider how to embed sustainability into their initiatives, so the benefits of portals and Open Data are available for all, now and in the future.

4.1. Responding to rapid progression in portal maturity

The results of our surveys and interviews, as well as informal feedback, demonstrated fast progression in portal maturity since the previous sustainability report. Indeed, these findings are supported by the rapid progression in portal maturity in the Open Data Maturity in Europe index between 2016-7, where portal maturity rose by 10% points (a 44% increase on the 2015 figure).

From a concentration on portal setup and internal government readiness, portals are increasingly now looking outward to the use of Open Data and its impact 'on the ground' in the economy and society, to opportunities to collaborate with sectors of society, and to methods for guaranteeing a financially sustainable portal in the future.

The previous report examined portals' difficulties around 4 key areas: governance, financing, architecture and operations, as well as examining metrics for monitoring progress. We issued recommendations around these areas, including that portals should have a business plan and clear

governance structure in place, be open about their funding strategy, select open source software solutions, capture and share operational lessons learned and choose metrics that benchmark the quality of data publisher performance, among others. These recommendations were designed to support countries in ensuring portals are sustainable across its lifespan, from setup to sustainability.

Evidence presented in this report suggests that these recommendations are still relevant, but that further specific guidance was needed in the complex areas of finance and impact. Whilst our surveys and interviews indicated that many portals are now confident in their funding model, with 94% of survey respondents who said it was sustainable, the breadth and depth of variance in portals training, technical aspects, awareness raising, organisational structure and legal aspects means that more specific guidance is needed.

Consequently, this report's recommendations aim to supplement and deepen the recommendations of the first report, supporting portals' progress in implementing best practice for ensuring financial sustainability and impact.

Many portals across Europe are taking important steps towards becoming sustainable. We hope that the recommendations in this report, as well as training and other support offered by the EDP, will aid countries in deepening their portal's sustainability, now and in the future.

4.2.   Making impact relevant and holistic

Successful Open Data initiatives must effectively understand and communicate the progress and impact of their work. It is critical that portals adopt a joined-up approach to measuring impact, in order to produce figures that are credible, repeatable and accurately show the impressive impact that portals and Open Data initiatives are having across Europe.

Recommendations for portals: ensuring impact

Measuring impact through existing approaches

- Adopt and adapt the **Common Assessment Framework** to measure portal performance, identifying and using relevant existing metrics around Data and Context/Environment.
- Ensure **macroeconomic and microeconomic impact studies** provide **clear, detailed and repeatable methodologies** and **publish underlying data and tools**, which allow these calculations to be repeated
- For **business populations and user surveys,** partner with other organisations, **examine existing studies** and pose consistent questions, **publish the underlying data** as Open Data on the portal and make efforts to **automatically collect** and analyse the data on an ongoing basis.
- Establish **showcases and use cases** that allow users to submit their own re-uses, **encourage reporting** of re-use through community engagement, follow up with showcase re-users on a regular basis, **link use cases to the specific datasets** that are used and **collect more structured data** that could be linked.
- For **automated access metrics**, use page analytics and **track downloads** at the dataset level, keep APIs logs, and **publish access data under open licences**

> **Measuring impact through new approaches**
> - Use **holistic approaches** that focus on **use and impact at a dataset level** and examine approaches to **automating microeconomic analysis** based on the ongoing data collection approaches.
> - **Share data** by **publishing underlying data** from studies, using methods from other portals to **infer and compare use** and impact, and **share metrics** for data not published on the platform.
>
> **Technical methods for tracking re-use**
> - For **automated approaches to re-use**, explore how **tracking APIs**, creating version control hosting and **web searching technologies** could be used to track use.
> - Explore **data citation**, by developing **'model citations'** that indicate best practice, **guidance around granularity** of data for citation and guidance that **explains the purpose** of data citation tracking

4.3.   Creating a sustainable financial environment

We have also examined the various ways portals are creating a financial environment that enables and supports their sustainability. Many portals are facing a potential slow-down or loss of funding at the same time as they are becoming more ambitious. Here, we find that the vast majority of EU portals are offering Open Data on a marginal or zero-cost model, covered by government funding. It is unclear how the sustainability of these portals, and the Open Data initiatives and innovation they support, would be influenced by a sudden loss of financial support.

In this report, we find that although 94.4% view their funding model as sustainable, there are many hidden costs for a successful, sustainable portal. These are concentrated in three main areas: maintaining portal infrastructure, engaging and encouraging publishers and building awareness, engagement and innovation. Within each of these dimensions, portals are making decisions which have significant costs and impacts for their activities. Portals need to examine how they share the costs that they incur in seeking sustainability, these decisions are heavily affected by the type of portal they are.

Finally, portals need to examine how to create a financially sustainable model that is fit for purpose, now and into the future. This model might in some cases examine potential alternative funding sources, in particular, the use of freemium models for currently paid-for datasets, as a way of funding them into the future.

> **Recommendations for portals: creating a sustainable environment**
>
> **Current approaches to funding portals**
> - Choose either a third party hosted solution or build an instance of a popular open source platform such as CKAN or DKAN.
>   - If choosing a hosted portal, ensure strong procurement processes are in place to identify the best suppliers.
>   - If choosing to host an open source platform, ensure there are sufficient levels of technical knowledge on the team, either through hiring or upskilling existing staff.
> - Look at ways to share the cost burden of technical hosting by partnering with other portals or governments to develop features, partnering with other governments to share a platform

or bringing in technical expertise through fellowship programmes and other partnerships.

Engaging and encouraging publishers
- National portals should aim to have individual named contacts within each government department, if resource permits.
- Portals should aim to reduce costs by engaging multiple publishers through the provision of workshops and training.
- Portals should explore the use of self-service tools that help engage publishers such as dashboards, usage reports and rankings.

Building awareness, engagement and innovation
- Invest in events like hackathons as a method for both encouraging innovation among re-users and greater publishing of Open Data by government ministries
- Look for opportunities to create shared innovation strategies and funds across cities and regions
- Form partnerships with organisations beyond government, such as universities, and invest in joint ventures that may produce long-term impact

Alternative funding models
- Explore how to adopt a freemium model to datasets that are not yet open

4.4.   The future of portals

Evidence from this report and others indicates that portals have become an essential part of the Open Data infrastructure for countries across Europe. They are an important resource for both publishers and users, providing a central point of access that is helping people find data and encouraging re-use of Open Data. Consequently, the work of portals in supplying data is underlying products and services that are bringing tangible benefits for citizens across Europe and beyond.

The rapid changes of the last few years are unlikely to stop anytime soon. Consequently, when considering their sustainability, portals should also consider how their role within Open Data initiatives may change. Data is likely to be derived from other sources of publicly funded data, through phenomena like smart cities and the Internet of Things[140]. Likewise, as Open Data re-use increases, the number and character of users is also changing. Portals must consider how meeting and encouraging demand and providing value may impact the way that portals provide Open Data in future.

This may lead to a future where API-based distributed publishing becomes more common, providing more value for publishers and re-users alike. In this space, the technical infrastructure of portals could function as a front-end catalogue for datasets accessible elsewhere, whilst the considerable expertise of portal teams is used to drive publication across government, as well as sustaining the catalogue. The work that some portals are currently engaging with various user groups and providing a variety of services, as showcased in this report, provides a signal of this emerging strand of work.

---

[140] European Data Portal (2017). Analytical Report 8. Available at:
https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf.

From setup to sustainability, portals across Europe are changing rapidly, providing new services and data that is supporting the growing impact of Open Data. However, there is more to be done: both portals and governments more widely must take the steps outlined in this report to track, measure and encourage impact, as well as implementing funding models and relationships that will ensure financial sustainability in the long-term. This will enable the benefits of Open Data portals, and the Open Data they provide, to be shared by all.

Annex I - Survey questions

Questions for the online survey aimed at Open Data portal owners across Europe and managers who are responsible for data portals. It was carried out between May and July 2017, receiving 19 responses in that time.

Demographics

1. Email address
2. What portal do you represent?
3. What type of portal is it?
   a. [national/local/community/other]
4. If you read the first report, 'From Setup to Sustainability: Recommendations for Open Data Portals', which is attached to the email you received, and you would be happy to quickly provide us with some feedback, please answer the following three questions. (If not, please skip to section 2). Were the recommendations useful to you?
   a. [Yes/No]
5. How were they useful?
6. What else should have been included?

Re-use of Open Data

1. Do you measure how data on your portal is re-used? If so, how?
2. What type of usage data would be most useful to you? How would you use it?
3. Do you think it is feasible to track or measure the re-use of Open Data automatically?
   a. [Yes/No]
4. Have you made efforts to do so?
   a. [Yes/No, not automatically but manually/No]
5. If you have, can you briefly describe these?
6. What challenges did you find? (eg. financial/technical/privacy)
7. Can you think of examples of ways to understand re-use from other sectors that Open Data Portals could learn from?

Financial sustainability

1. What is your portal's funding model?
   ○ Profit Maximising (selling data for a high price to increase public sector's profit)
   ○ Average Cost/Cost Recovery Model (selling data to get the costs of releasing data back)
   ○ Marginal Cost (Zero Cost) Model (providing data for the costs of processing the data request only or is free of charge)
   ○ I don't know
2. Could you explain why this funding model was chosen?
3. Do you think it is sustainable?

Annex II – Interview questions

These are the questions for the semi-structured interviews carried out with portal owners, data platform providers and others who had carried out Open Data impact assessments. The 11 interviews took place between July and August 2017.

Portal owners

1. Do you have any additional feedback on the areas and recommendations from the previous sustainability report?

Current approaches to tracking access + use

2. How, if at all, do you measure if data on your portal is access?
3. Have you measured re-use in any other ways - e.g. case studies/research/user engagement?
4. What are the challenges to tracking access + use?
5. What would be the most useful metrics and measures you could put in place to justify funding? Have you tried to get these?

Funding situation

6. Where does the funding come from?
   a. What impact does this model have on the nature of the portal - choices about design, hosting, longevity?
7. Why is/isn't this sustainable?
8. Have other funding models been considered?
   a. If so which ones?
9. What relationships do you have with other government bodies?
   a. What do you do with them?

Portal software providers

Current approaches to tracking access + use

1. How do you currently measure how data is accessed through your platform?
2. What metrics do you give platform owners?
   a. What metrics do they ask for?
3. Do you offer any other mechanisms by which use can be tracked - embedded surveys, 'showcase functionality' etc?
   a. Any automatic options?

Possible approaches to tracking access

4. What are the key challenges in your opinion to tracking access + use?
5. Have you considered other ways of understanding and tracking access and use?
   a. Have you tested any?
   b. What are/would be the barriers to implementing these approaches?

Cost structure

6. What cost structure do you use for hosting Open Data portals - per dataset, per user, by feature etc?
    a. What feedback have you got on this system?

Impact assessors

Current approaches to tracking access + use

1. How do you currently measure how data is accessed through your platform?
2. What metrics do you collect? What metrics do you need?
3. Do you use any other mechanisms by which use can be tracked - embedded surveys, 'showcase functionality' etc?
    a. Any automatic options?

Possible approaches to tracking access

4. What are the key challenges in your opinion to tracking access + use?
5. Have you considered other ways of understanding and tracking access and use?
    a. Have you tested any?
    b. What are/would be the barriers to implementing these approaches?
4. Why?
5. Which Government bodies do you interact with?
6. Can you briefly summarise the interaction you have with each of them?