



D2.1 OPEN DATA TOPOLOGIES, CATALOGUES AND METADATA HARMONISATION

PROJECT

Acronym: **OpenDataMonitor**

Title: Monitoring, Analysis and Visualisation of Open Data Catalogues, Hubs and Repositories

Coordinator: SYNYO GmbH

Reference: **611988**

Type: Collaborative project

Programme: FP7-ICT

Start: November 2013

Duration: 24 months

Website: <http://project.opendatamonitor.eu>

E-Mail: office@opendatamonitor.eu

Consortium: **SYNYO GmbH**, Research & Development Department, Austria, (SYNYO)

Open Data Institute, Research Department, UK, (ODI)

Athena Research and Innovation Center, IMIS, Greece, (ATHENA)

University of Southampton, Web and Internet Science Group, UK, (SOTON)

Potsdam eGovernment Competence Center, Research Department, Germany, (IFG.CC)

City of Munich, Department of Labor and Economic Development, Germany, (MUNICH)

Entidad Publica Empresarial Red.es, Shared Service Department, Spain, (RED.ES)

DELIVERABLE

Number:	D2.1
Title:	Open data topologies, catalogues and metadata harmonisation
Lead beneficiary:	SOTON
Work package:	WP2: Research studies and stakeholder analysis
Dissemination level:	Public (PU)
Nature:	Report (PU)
Due date:	June 30, 2014
Submission date:	June 30, 2014
Authors:	Elena Simperl, SOTON Yunjia Li, SOTON Tom Heath, ODI Bernhard Krieger, IfG.CC Sirko Hunnius, IfG.CC Dimitris Skoutas, ATHENA Lukas Wenzel, SYNYO
Contributors:	Amanda Smith, ODI Inian Moorthy, SYNYO Bernhard Jager, SYNYO
Reviewers:	Bernhard Krieger, IfG.CC Sirko Hunnius, IfG.CC

Acknowledgement: The OpenDataMonitor project is co-funded by the European Commission under the Seventh Framework Programme (FP7 2007-2013) under grant agreement number 611988.

Disclaimer: The content of this publication is the sole responsibility of the authors, and in no way represents the view of the European Commission or its services.

TABLE OF CONTENTS

1	Introduction (SOTON, ODI)	7
1.1	Open Data Monitor (ODM) contextualisation of this report (SOTON)	8
2	Previous studies on open data (SOTON, ODI)	10
2.1	The potential of open data (SOTON).....	10
2.2	Open data deployment (SOTON)	11
3	The open data landscape: topologies and landmarks (SOTON)	14
3.1	Terminology (SOTON, ODI).....	14
3.2	Open data life cycle (IfG.CC).....	16
3.3	The open data ecosystem and its stakeholders (IfG.CC, SOTON)	18
3.4	Open data topologies analysis (SOTON).....	21
3.4.1	Methodology	21
3.4.2	Open data topology analysis using Google News	22
3.4.3	Limitations	25
4	The open data landscape: A technical view (SOTON)	26
4.1	Open Data Portals (SOTON, SYNYO, ATHENA)	26
4.1.1	Pan-Europe Open Data Portals.....	27
4.1.2	National Level Open Data Portals.....	28
4.1.3	Local Level Open Data Portals	29
4.1.4	Domain specific Open Data Portals.....	29
4.1.5	Open Data Portals in Private Sectors	30
4.2	Open Data Software and APIs (SOTON, ATHENA).....	31
4.2.1	CKAN.....	31
4.2.2	Socrata.....	31
4.2.3	Junar	31
4.2.4	DKAN.....	32

4.2.5	Open Government Platform (OGP)	32
4.2.6	QU.....	32
4.3	Metadata Standards (SOTON).....	32
4.3.1	DCAT	33
4.3.2	Asset Description Metadata Schema (ADMS)	33
4.3.3	DCAT-AP.....	33
4.3.4	CKAN Attributes.....	34
4.3.5	INSPIRE Metadata Schema	34
4.3.6	Common Core Metadata Schema (CCMS) in Project Open Data	34
4.3.7	Data Catalog Interoperability Protocol (DCIP).....	35
4.3.8	Vocabulary of Interlinked Datasets (VOID).....	35
4.3.9	Schema.org.....	35
4.3.10	Google Dataset Publishing Language	35
4.4	Use of metadata standards by portals, software and APIs (SOTON)	35
5	Challenges and Solutions (SOTON).....	36
5.1	Data discovery (SOTON, ODI)	36
5.2	Data awareness and insight (SOTON, ODI).....	37
5.3	Metadata harmonisation (SOTON, ATHENA)	38
5.3.1	Methodology	38
5.3.2	Metadata Status for CKAN Instances	38
5.3.3	Availability of Different Metadata Attributes	39
5.3.4	Proposed Attributes for Harmonisation.....	41
6	Conclusion (SOTON)	43
7	References.....	44
8	Appendix I: Namespace abbreviations in this report	47

LIST OF FIGURES

Figure 1. Heat map of scores according to the Open Data Barometer 2013 (Davies, 2013)	11
Figure 2. Open data readiness in different regions (Davies, 2013).....	12
Figure 3. Snapshot of Open Data Index for different countries.....	13
Figure 4. Snapshot of the scores for Transport Timetables	14
Figure 5. Open data ecosystem from Deloitte (Deloitte, 2012).....	18
Figure 6. Named entity type and number extracted from Google News.....	25
Figure 7. Radar Chart (Ireland, Greece, Europe) of scaled sub-component scores (Davies, 2013)	27
Figure 8. Availability of metadata attributes.....	40

LIST OF TABLES

Table 1. Relationships between D2.1 and other deliverables in WP2	9
Table 2. Terminology Definition	15
Table 3. Named entity type and frequency extracted from Google News	22
Table 4. Metadata adoption for open data software, portals and APIs.....	36
Table 5. Important attributes that are usually missed.....	40
Table 6. Preliminary analysis of thee CKAN attributes.....	40
Table 7. Proposed attributes for harmonisation	42
Table 8. Namespace abbreviations used in this report.....	47

1 INTRODUCTION (SOTON, ODI)

Nations across the European Union have been at the forefront of deployment and exploitation of open data for many years. The diversity of approaches and experiences in European Member States, in the public, but also, though predominantly on a consumer side, in the private sector, provides a rich body of knowledge for those seeking to understand the field in greater detail.

Open data is a broad and multidisciplinary topic, encompassing not only technology, but also a wide range of societal, business, and policy aspects. Consequently, it must be studied from multiple perspectives, and at varying levels of granularity. Notable work has been conducted exploring the economic and social potential of open data (Deloitte, 2012; Manyika, 2013) while others have examined the availability of key data sets in open form, the structural readiness of nations in exploiting the benefits of these, and the impacts such actions are having (Davies, 2013).

This document will review this literature, and build upon the findings and recommendations of previous studies in order to provide a comprehensive technical account of the deployment of open data across Europe. Such an analysis is largely missing, hence making it difficult for open data stakeholders to have a complete picture of the open data landscape and understand the full implications of open data adoption in terms of technology development and use. This report provides technology-centric account. It should be considered and read as a technical companion volume to reports such as the Open Data Barometer (Davies, 2013), and is aimed at those who demand greater insight into implementation specifics and related challenges.

A second notable feature of this report is its broad coverage of open data stakeholders. More specifically, our analysis of the open data landscape includes not only public sector publishers and consumers of open data sets, but also commercial companies that, one way or another, have become part of this ecosystem. Commercial organisations are considered both as data consumers and intermediaries, and as producers and publishers in their own right, based on a recognition of how open data can create commercial benefit, for example through greater efficiency, customer engagement, and open innovation. This analysis complements existing surveys, which, for historical or other reasons, have put more emphasis on public administration data providers.

The result is a comprehensive account of the ‘topology’ of the open data landscape, featuring both technical and non-technical aspects, as well as a richer portfolio of stakeholder groups. This includes a unifying view of the open data life cycle and the related ecosystem, including a glossary of the most common terms, which leverage insights from existing literature and our own research, as well as a survey of the digital artifacts (portals, software, APIs, and metadata standards) that are typically part of an open data deployment, and examples of such deployments.

This analysis allowed us to identify a number of challenges for the development of future studies of such kind, motivating one more time the need for a comprehensive technical solution, including monitoring, analytics, reporting, and visualisation features, to facilitate timely and rich assessments

of the state of the art in the field – which is the core mission of the Open Data Monitor (ODM) project¹ sponsoring this report.

The remainder of this document is structured as follows: we begin with a summary of existing studies of the open data landscape, which informed our research in Section 2. In Section 3 we then give an overview of the open data ecosystem and its stakeholders, including preliminary results of a quantitative analysis of the corporate part of this ecosystem, using automatic information extraction techniques applied on a large news and social media corpus. Section 4 is dedicated to the technical aspects of open data, including a proposal for metadata harmonisation, which we understand as a prerequisite for the operationalisation of open data assessment reports. We conclude with a discussion of the great challenges identified in our work, which will be addressed in the ODM project.

1.1 Open Data Monitor (ODM) contextualisation of this report (SOTON)

This report overviews the landscape of open data, principally in Europe, and is therefore closely related to other deliverables in WP2 (Research studies and stakeholder analysis). In general, this report lists and analyses resources covering different aspects of open data, which kicksstarts the directions that other deliverables will take up for deep analysis in WP2. As has been specified in the Description of Work of this project, ODM will provide:

“sophisticated methods to scan data catalogues, analyse meta-data and provide comprehensive visualisations to compare existent open data resources. Using standardised APIs (e.g. CKAN²) it will be possible to analyse data usage, file formats, updates, licenses and further meta-data to statistically describe and visualise it. This information will be used to identify trends, gaps and potentials of open data resources. “

and

“a scalable open data monitoring concept using metadata, parameters and key-indicators. (MS2)”

To identify the gaps between the current open data monitoring methods and what ODM will achieve, Section 2 goes through the previous studies on open data and reveals that the technical dimensions are seldomly covered in those reports, especially the detailed analysis of different attributes in metadata, which is a gap that ODM project will fulfill. We also introduce the methodologies that applied in those reports and point out in Section 3.4 that those methodologies based on survey and expert reviews are not applicable in ODM project because we are looking for automatic means to analyse open data topologies and those methodologies are not scalable when large amount of open data resources are involved. So we propose a new quantitative methodology in Section 3.4 to automatically identify stakeholders, catalogues and visualise them in different ways. In addition, we define the terminology in Section 3.1 that will be shared across other deliverables to ensure the consistent understanding of the concepts in this project.

¹ <http://project.opendatamonitor.eu/>

² <http://ckan.org> will be discussed further in Section 4.2

In detail, the relationships to other deliverables in WP2 are shown in Table 1.

Table 1. Relationships between D2.1 and other deliverables in WP2

Deliverable	Relationship explained
D2.2 Monitoring methods, architectures and standards analysis report	Section 2.2 discusses two examples related to open data monitoring methods: Open Data Indexer (via expert reviews) and Open Data Barometer (via general survey). Section 4 briefly goes through the technical aspect of open data, such as platforms, APIs and metadata standards, which need to be further analysed in D2.2.
D2.3 Best practice visualisation, dashboard and key figures report	In Section 2.2, 3.4 and 5.3, we quote data visualisations from other reports and propose several ways to visualise open data topology and metadata. They can be further expanded with the metrics defined in D2.3. We also mention in Section 4.2 that many existing open data platforms have built-in visualisations and dashboard features that need to be taken into account in D2.3.
D2.4, D2.6 Open data stakeholder requirement report 1 and 2	Section 3.2 and 3.3 conducted a preliminary discussion of the open data life cycle and stakeholder identifications. Knowing who are the stakeholders and how the data will be passed around and enriched in the life cycle is the first step of delivering comprehensive stakeholder requirements. The topology analysis in Section 3.4 also provides a methodology to discover more instances of the stakeholders in different categories.
D2.5, D2.7 Open data resources, platforms and APIs collection 1 and 2	The Section 4 of this report can contribute to the resources, platforms and APIs collection. The topology analysis methodology in Section 3.4 can help discover more resources.

2 PREVIOUS STUDIES ON OPEN DATA (SOTON, ODI)

We distinguish between two categories of reports in the literature: those discussing and providing evidence for the potential of open data across industrial sectors, and those reviewing its adoption. As noted earlier, most of these studies do not consider the technical dimensions of the open data landscape, or focus on public sector stakeholders.

2.1 The potential of open data (SOTON)

Deloitte's 'Open Data Driving Growth Ingenuity and Innovation' looked at the open data landscape as of 2012 to identify trends for the future development of this area and recommendations for the commercial sector (Deloitte, 2012). They estimate that in the near future business will engage in open data in four aspects: (1) strategically exploit the rapidly growing of their open data assets; (2) opening up their data assets as a revolution way of competing; (3) using open data to improve transparency and engage customers; and (4) work with government and make policies for data responsibility and privacy (Deloitte, 2012).

A report produced by McKinsey around the same time sought to "quantify the potential value of open data by examining applications in seven sectors of the global economy": education, transportation, consumer products, electricity, oil and gas, healthcare, and consumer finance (Manyika, 2013). The report encompasses not only an analysis of the economic value of open data in each of these sectors, but also a discussion of potential barriers to adoption and actions to be taken to ensure that this potential is not lost. For example, the privacy issues are major concerns in nearly all the investigated domains and appropriate legal and regulatory frameworks are urgently needed to ensure that open data is distributed in an anonymous and secure manner. In the 'Researching the emerging impacts of open data' paper (Perini, 2013), it is highlighted that there is a need for methodological and tool support to allow the "various stakeholders to engage in an informed dialogue and to guide the future development of open data".

The 'Open Government Data Stakeholder Survey 2010' (Martin, Kaltenbock, Nagy, & Auer, 2011) led by the LOD2 project³ focused on open government data. They report on the requirements of different stakeholder groups (citizens, public administration agencies, policy makers, industry, media, and science) regarding open data sets and catalogs. From the survey, national and regional data sets are most required by the stakeholders and they demand more data to be published in non-proprietary formats such as CSV and XML.

A position paper by NEF⁴ proposes a research agenda for big and open data. The two are seen as complementary, with open data technologies allowing organisations to easily repurpose their data assets and enrich them with other openly accessible content. Though the focus of the roadmap is on technical aspects, the paper also discusses privacy challenges that need to be addressed when implementing cloud-based data provisioning solutions.

³ <http://lod2.eu/>

⁴ <http://nem-initiative.org/wp-content/uploads/2013/11/NEM-PP-016.pdf>

2.2 Open data deployment (SOTON)

There are already several reports and publications analysing the adoption of open data across the globe. The **Open Data Barometer 2013** (Davies, 2013) focuses on the analysis of open government data (OGD) in 77 countries in terms of context, availability, and emerging impacts. The report is structured in three sections: (1) readiness to develop an open data strategy; (2) the extent to which such as strategy has been implemented; and (3) an update of the state of the art is expected for the second half of 2014.

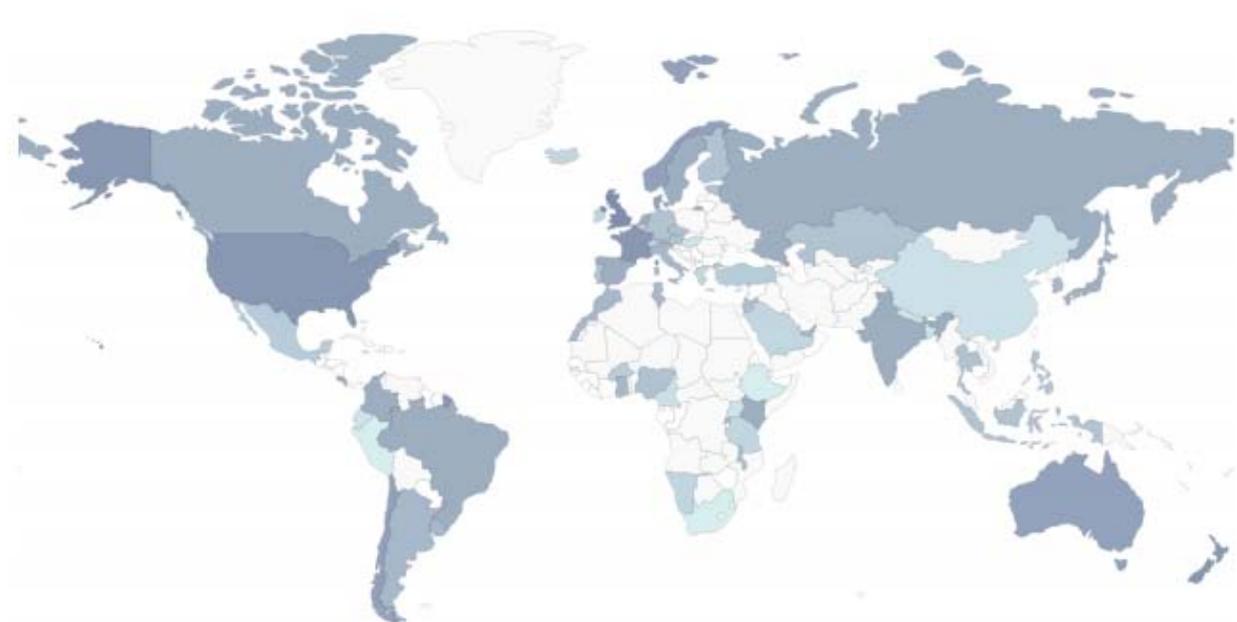


Figure 1. Heat map of scores according to the Open Data Barometer 2013 (Davies, 2013)

Figure 1 illustrates the diversity of the OGD landscape in terms of adoption and readiness of open data and the level of activity of individual governments. Darker colour in the heat map means higher level of readiness of open data and more active involvement of the government. For the open data readiness, the variables are divided into three components: “Government capacity and the presence of government commitments to open data; Citizen and civil society freedoms and engagement with the open data agenda; Resources available to entrepreneurs and businesses to support economic reuse of open data” (Davies, 2013). Figure 2 uses radar charts to illustrate the readiness of OGD in different regions. From the charts, we can see that, Europe is the leading region, while the deployment of OGD in Middle East & Central Asia and Africa is very limited.

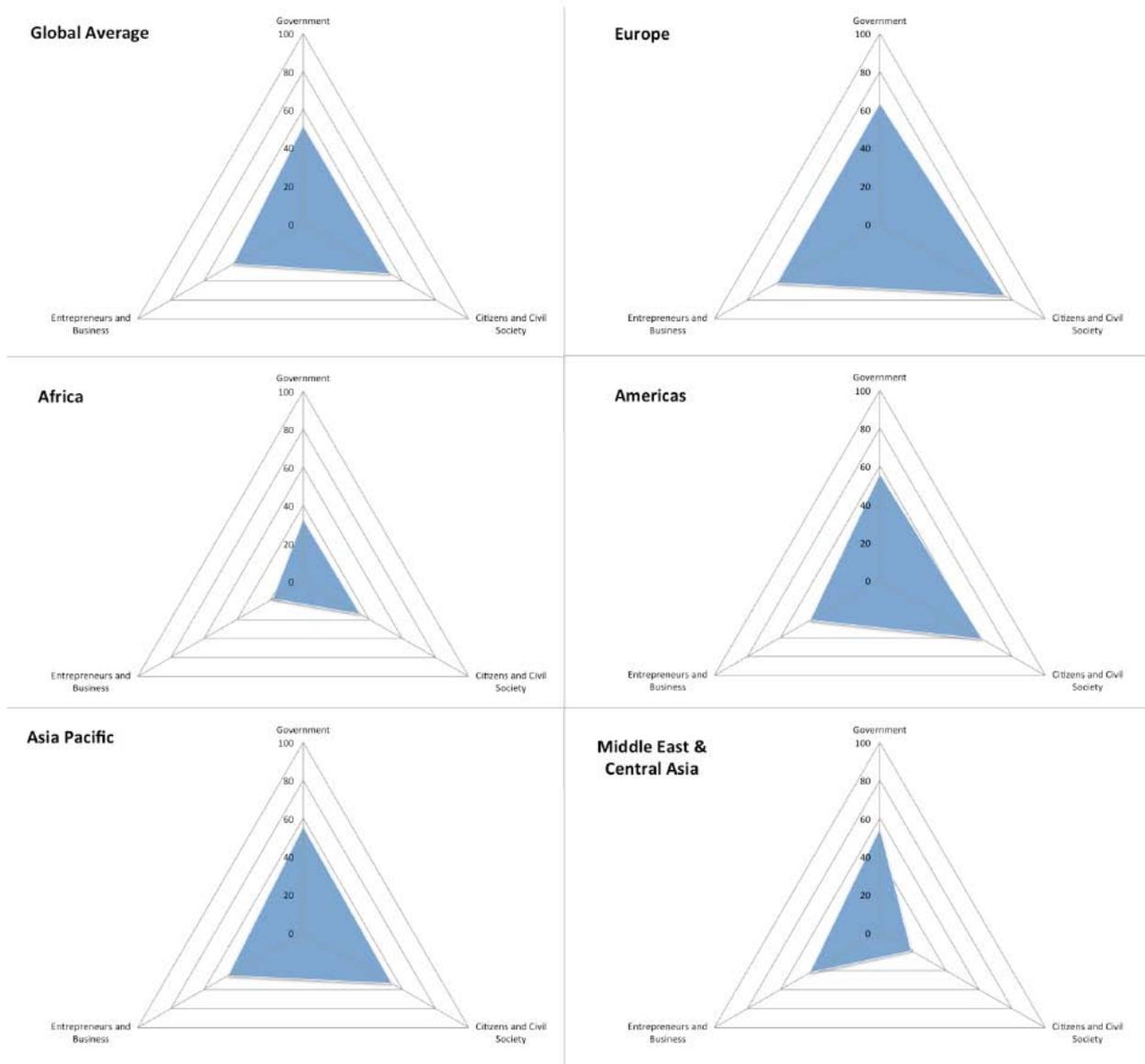


Figure 2. Open data readiness in different regions (Davies, 2013)

The **Open Data Index**⁵ was launched in 2013 by the Open Knowledge Foundation as a mechanism to assess the state of open data around the world. It covers information concerning the data sets published by national governments in over 70 countries. Annual snapshots of the data are presented on the Web site to showcase the results of the project. One of the main goals of this project is to stimulate debate and action between citizens and their governments to lead to the release of further data assets. From the 700 key data sets that have been identified in the current release of the index, only 84 data sets are considered open.

The Open Data Index structures the open data sets into 10 different categories, such as transport timetables, budget, election, national map, etc., and each category uses nine same criteria to

⁵ <https://index.okfn.org/>

measure the availability of the data, such as whether the data is online and free of charge. Based on the submissions provided by the editors, the index gives a total score of openness for each country. Figure 3 demonstrates a snapshot of the countries with top scores in Open Data Index. From the snapshot we can see that, until now May 2014, 6 EU countries are in the top 10 of the index. The Open Data Index also provides a detailed break down score for each country in each category (see Figure 4 for example of Transport Timetables).

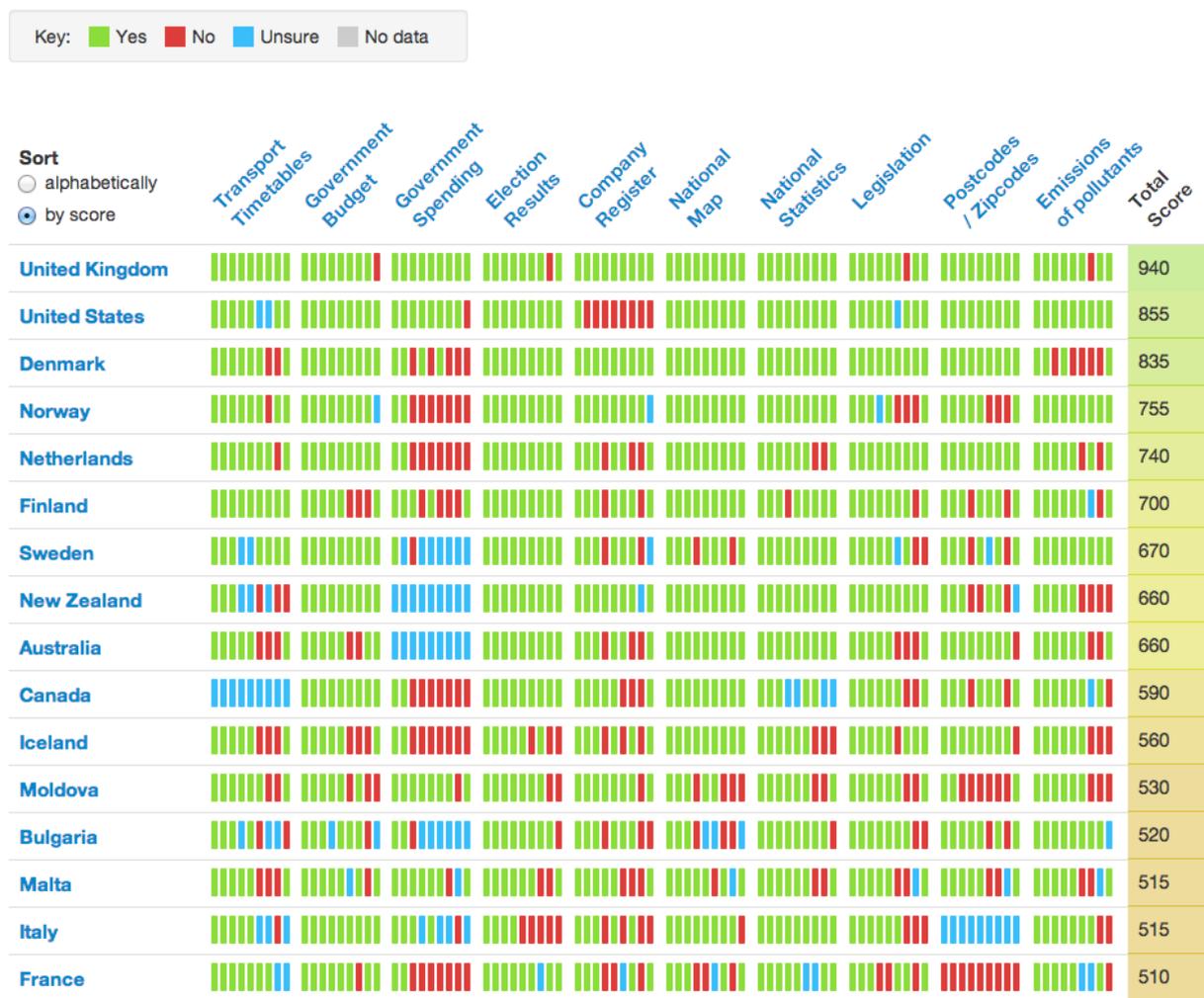


Figure 3. Snapshot of Open Data Index for different countries⁶

⁶ <https://index.okfn.org/country/>

Datasets / Public Transport Timetables

On this page you can see the state of open data for Public Transport Timetables across all countries for which we have information (displayed down the left hand side). Each icon in the data availability column represents important factors indicating data accessibility or availability - mouse over the icons to see what they are and the colours correspond to yes / no / unsure / no data.

Country	Score	Breakdown	Location (URL)	Information
United Kingdom	100%	        	http://data.atoc.org/rail-industry-data	  
Finland	100%	        	http://developer.matka.fi/pages/en/home.php	  
Norway	90%	        	http://labs.trafikanten.no/how-to-use-the-api.aspx	  
Sweden	80%	        	http://www.trafiklab.se/api/gtfs-sverige	  
United States	75%	        		  
Netherlands	70%	        	http://9292opendata.org/datacollecties	  
Israel	70%	        	http://he.mot.gov.il/index.php?option=com_con...	  
Moldova	60%	        	http://www.autogara.md/orar/	  
France	60%	        	http://test.data-sncf.com/index.php/	  
Ireland	60%	        	http://www.transportforireland.ie/	  
Denmark	60%	        	http://labs.rejseplanen.dk/labs/data_brug/rejse...	  
Russian Federation	45%	        	http://pass.rzd.ru	  
Australia	45%	        	http://www.railaustralia.com.au/faresTimetables...	  
Hungary	45%	        	http://menetrendek.hu	  
Bangladesh	45%	        	http://www.railway.gov.bd/all_ic_mail_schedule.asp	  
Brazil	45%	        	https://appweb.antt.gov.br/sgp/src.br.gov.antt/a...	  

Figure 4. Snapshot of the scores for Transport Timetables⁷

3 THE OPEN DATA LANDSCAPE: TOPOLOGIES AND LANDMARKS (SOTON)

In this section we will give an overview of the open data landscape based on the insights gained from the literature surveyed in Section 2 complemented by our own research. In particular, we undertake a more detailed stakeholder analysis as a means to describe and understand the mechanics and evolution of the underlying ecosystem, and compile a first list of corporate organisations that are part of it by mining a large corpus of news and social media.

3.1 Terminology (SOTON, ODI)

In order to describe the concepts involved in open data in a consistent manner, we have defined a series of terms in Table 2, which summarises our understanding of some of the most common terms

⁷ <https://index.okfn.org/country/dataset/timetables>

used to describe the open data landscape. The terminology is aligned with DCAT-AP (DCAT Application profile for European data portals)⁸ and the metadata standards applied in publicdata.eu⁹. When defining those terms, we have also referred to the existing terminologies related to open data, such as data.gov.uk glossary¹⁰, open data handbook¹¹ and W3C Linked Data Glossary¹², and make sure that our definitions do not conflict with theirs. In the Table, we also identify some synonyms and they could be used interchangeably in some circumstances, such as open data catalogue, repository, portal and platform.

Table 2. Terminology Definition

Terms	Definition	Synonym
Open data	“A piece of data is open if anyone is free to use, reuse, and redistribute it - subject only, at most, to the requirement to attribute and/or share-alike” ¹³ .	
Stakeholder	“any group or individual who can affect or is affected by the achievement of the organisation's objectives” (Freeman, 1984)	
Open data repository	An online data storage/hosting service but with no discovery mechanism. This could be as simple as a Web server hosting static files from a single folder, with no additional index or categorisation, except perhaps a 'landing page' for each data set.	Open data catalogue, data hub
Open data catalogue	A curated collection of metadata about data sets. Compared with “open data repository”, “open data catalogue” focuses on the organisation of data sets, while “open data repository” refers to the actual data storage. The catalogue would typically be agnostic regarding where the data itself is located: (1) it may all be published on the same Web server as the catalogue, i.e. the catalogue contains a data repository, or (2) may be distributed across the Web, with the catalogue simply pointing to those remote locations, in which case the catalogue is also referred to as a “data aggregator” or “data indexer”.	Open data portal, data hub, open data repository
Open data portal	Often used synonymously with <i>open data catalogue</i> , but may provide more advanced discovery functionality to complement conventional browse-style catalogue interfaces.	Open data catalogue, data hub, open data platform

⁸ https://joinup.ec.europa.eu/asset/dcat_application_profile/description

⁹ <http://publicdata.eu/>

¹⁰ <http://data.gov.uk/glossary>

¹¹ <http://opendatahandbook.org/>

¹² <http://www.w3.org/TR/ld-glossary/>

¹³ <http://opendatahandbook.org/en/what-is-open-data/>

	For example, there may be text search over the metadata describing the data sets, or the ability to preview/explore the data itself. Distinctions between open data portals and catalogues, and between open data portals and platforms should be considered fuzzy.	
Open data platform	A piece of software that has implemented the core features to manage open data. Those features include, but are not limited to, user management, data publishing, metadata management, data set storage, access control, search and visualisation, etc.	Open data portal, open data portal software
Group	Groups are used to create and manage collections of data sets with some common features.	Collection, category
Data set	A conceptual entity that “represents a collection of data, published or curated by a single agent, and available for access or download in one or more formats” ¹⁴ . A data set is usually hosted in an open data repository and can belong to one or more groups.	Package (in CKAN)
Distribution	A distribution of a certain data set “represents a specific available form of that data set. Each data set might be available in different forms, and these forms might represent different formats of the data set or different endpoints. Examples of distributions include a downloadable CSV file, an API or RSS feed” ¹⁵ .	Resource (in CKAN)
Metadata	The metadata of a data set is a collection of data that describes the data set and provides more information about the data set, such as title, tags, license, maintainer, etc. The metadata can be provided in different format, such as JSON, XML and RDF.	

3.2 Open data life cycle (IfG.CC)

This subsection analyses models that conceptualise the practices around handling data, from its generation to administrative practices involved in the provision of open data by public sector institutions to its use by third-parties. Various models of (linked) open data have been suggested under different terminologies. They have been named the open data life cycle, the open data value chain or plain open data process (Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012). The different terminologies illustrate different purposes – practical guidance (Hyland & Wood, 2011) or analytical understanding – and foci. Whereas value chain models focus more on the creation of value during open data usage (Julien, 2012), the life cycle models aim to structure the handling of the data

¹⁴ <http://www.w3.org/TR/vocab-dcat>

¹⁵ <http://www.w3.org/TR/vocab-dcat>

itself. Existing process models focus on activities within public administration, such as generating, editing and publishing the data without paying too much attention on the outside-use.

Most models contain similar elements and differ only regarding semantics, granularity or the extension of the process. Hyland et al. (2011) provide a six-step guidance model that contains the steps to (1) identify, (2) model, (3) name, (4) describe, (5) convert, (6) publish the data and the reverse activity to maintain it, similar to Villazon-Terrazas et al. (2011). Another model by Hausenblas and Karnstedt (2010) also includes the user perspective, adding the steps “discovery”, “integration” and “use cases”. With the ambition to build tools to support creating linked data, the LOD2 project developed a more fine-grained 8-step life cycle model (Auer et al., 2012). Synthesising various models, van den Broek et al. (2011) derive a life cycle model comprising the steps (1) identification, (2) preparation, (3) publication, (4) re-use and (5) evaluation.

All of these models describe the life cycle as a sequential, one-dimensional process of activities that an unspecified set of actors repeatedly undertake in order to provide a formerly unexposed amount of data to an abstract general public. Furthermore, these models include only one analytical level. They exclusively take the operational processes of open data publication into account (such as extracting, cleaning, publishing and maintaining data), while largely ignoring the strategic processes (such as policy production, decision making and administrative enforcement). Thus, the decisions which data will be published, who extracts data, how are data edited, how data can be accessed, which licenses are available, how data privacy and liability issues are treated, who is involved in these decisions etc. remain underappreciated. These more general strategic processes about open data refer to the governance structure, likely to be connected to an organisation's ICT and data governance.

The issues outlined point to another blind spot of most open data life cycle models that these are actor-blind. If at all, institutional characteristics and actor-interests are considered as “impediments” (Zuiderwijk et al., 2012) or restrictions hindering an inherently good and beneficial idea (Meijer, de Hoog, Van Twist, van der Steen, & Scherpenisse, 2014). This is especially relevant as the different stakeholders involved – which have been outlined below in Section 3.3 – have different understandings of and interests in open data what influences the results (Zuiderwijk & Janssen, 2014). Efforts have thus been undertaken to develop more holistic analytic perspectives on open data e. g. based on complexity theory (Meijer et al., 2014) or the information ecology approach (Harrison, Pardo, & Cook, 2012).

Furthermore, the data itself is often treated as “a commodity rather than an artifact” (Meijer et al., 2014). However, how (open) data is understood and interpreted is shaped by the institutional and legal context, e. g. different perceptions of privacy and personal data. In a similar manner, some data can be considered more politicised than other. Also, different professional perspectives on data that refers to the same material object influence not only the sense-making, but the consideration of what data is actually important, the metrics of measurement etc. Taken together, this might even question the viability of a generic life cycle model.

3.3 The open data ecosystem and its stakeholders (IfG.CC, SOTON)

A digital ecosystem usually contains some features: (1) cyclical; (2) sustainable; (3) demand-driven environments oriented around agents that are (4) mutually interdependent in the delivery of value (Heimstädt, 2014). So if open data is an ecosystem, we should be able to identify where the data comes from, how it feeds back to the provider and who are the interdependent agents that drive the demand of open data.

According to the Deloitte report (Deloitte, 2012), government, business and citizens are being identified as the three key constituencies (stakeholders) in a successful open data ecosystem (see Figure 5). Each actor in this space supplies different classes of data to different types of stakeholders. The government publishes so-called open government data, which are data sets “produced, collected or paid for by public money, subject and restrictions related to national security, commercial sensitivity and privacy”. Businesses publish open business data; this data is still freely opened to public, while being subject to specific restrictions that the businesses might decide to put in place. Finally, individual citizens may release personal and non-personal data to the open domain.

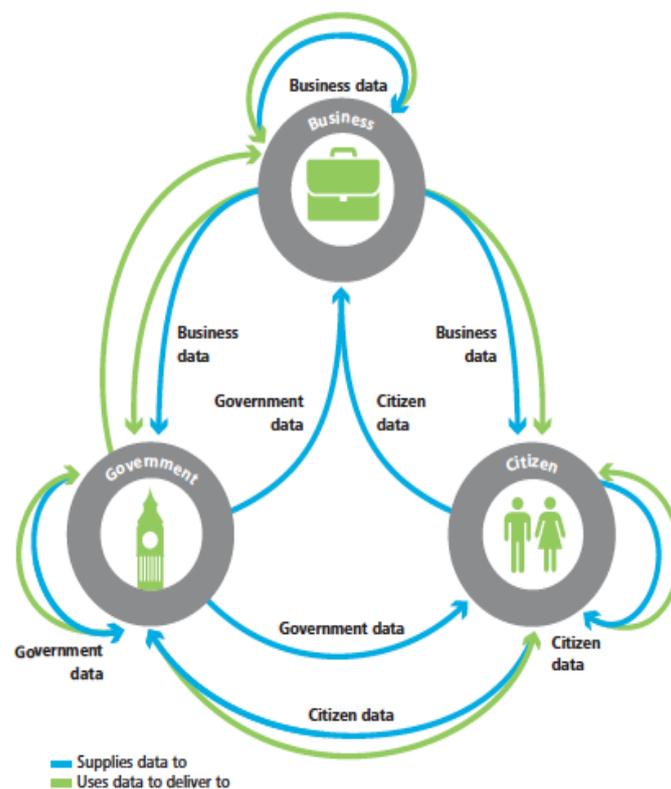


Figure 5. Open data ecosystem from Deloitte (Deloitte, 2012)

To further elaborate on the different types of activities undertaken by these three groups we performed a more detailed stakeholder analysis. The stakeholder approach was developed in business sciences as a means to analyse how groups or individuals with vested interests in a firm are or should be considered by its management. The reasons why a company might be interested in such analysis are very diverse, from moral (normative) to efficiency-oriented (instrumental) (Jones &

Wicks, 1999). It presents an alternative to the shareholder value model, according to which the managers are only accountable to a firm's owners.

Despite its initial focus on private entities, the approach has been applied to public sector settings (Tennert & Schroeder, 1999), in particular to study technology adoption in this space (see Scholl, 2001), mostly with positive results. Notwithstanding criticism of such an expansion (Donaldson & Preston, 1995), it can be argued that due to outward accountabilities (see e.g., Romzek, 2000), largely constitutive externalities (see e.g., Batley, 1994; Haque, 2001) and the network-type interdependencies of public sector management (see e.g., Kickert, Klijn, & Koppenjan, 1997; O'Toole, 1997), the stakeholder approach seems suitable and beneficial for public sector management (Scholl, 2001; Tennert & Schroeder, 1999). We will hence apply it to understand the open data landscape.

Before doing so, we need to consider some of its current limitations. In particular, most applications of the approach to public sector settings largely ignore the question of the reference point. While in an entrepreneurial scenario stakeholders mostly refer to corporations (Donaldson & Preston, 1995), it is harder to delineate to what stakeholders refer to in the public domain. In the main, methods to identify stakeholders start off from an organisation as the reference point (Blair & Whitehead, 1988; Mitchell, Agle, & Wood, 1997). Accordingly, to operationalise the stakeholder approach in the domain of open data, the concept can only be applied to stakeholders to a specific open data project or organisation in charge of open data, not to the abstract concept, or the topic of open data in general.

Different criteria to classify stakeholders have been brought forward in the literature (Tennert & Schroeder, 1999, Blair & Whitehead, 1988; Clarkson, 1995; Mitchell et al., 1997; Tennert & Schroeder, 1999). Following these criteria we distinguish stakeholders in five categories:

1. the stakeholder's power to exert influence
2. the legitimacy of the stakeholder's relationship with the organisation
3. the urgency of the stakeholder's claim on the organisation
4. their potential for collaboration on the one hand
5. for threatening the organisation on the other hand

These criteria can be applied to distinguish stakeholders in the open data domain. Thus, stakeholders have the power to substantially influence an open data endeavor by legitimately furthering or diminishing its effort. Accordingly, there is "no [...] necessity of reciprocal impact" (Mitchell et al., 1997). Based on the direction of impact and the pertinence of criteria, stakeholders can be classified (Clarkson, 1995; Mitchell et al., 1997).

Applying this framework to open data, stakeholders can be identified that are involved along the open data life cycle (see also Section 4).

Open data generators

The concept of open data is about making data openly available and accessible. This presupposes that data which meets the criteria to be considered as open is available and provided. Stakeholders who generate and provide open data thus have substantial power, e.g., commanding various means

to not disclose data. They often claim high levels of legitimacy by reference to data ownership, their mission or scarce resources.

Support units

Along the open data life cycle various stakeholders contribute essential input, e.g., strategic guidance, legal counselling, or platform provision. These support units have the power to raise the stakes or ease the process. Often, open data is none of their prime concern, but they provide specific professional expertise, from which they draw significant legitimacy. Especially in the public sector this expertise is often highly regarded (e.g., licensing, liability, privacy, security).

Open data users

A constitutive attribute of open data is the use of the published data by third-parties. Publishing data on the Web is thus not an end in itself. This creates significant dependency on open data users. Considering large troves of published data, incomprehensible to a single citizen, intermediary users, who provide tools to examine, analyse and understand data (e.g., apps), play a special role.

Politicians

Politicians are not involved in the open data life cycle itself, but play a special role in innovation processes in the public sector, where their support is often credited as crucial (Borins, 2002). Besides the considerable power they play in setting the agenda, this role is legitimate in a hierarchical bureaucratic system (Weber, 1958). Furthermore, if selected, politicians draw a large share of legitimacy by their constituency.

Advocacy groups

Advocacy groups are actively involved in setting the agenda for open data projects. Their power base is largely dependent on intermediary sources, since they cannot exert influence themselves. However, they claim considerable legitimacy, tying the topic of open data to larger democratic values of transparency and accountability. In addition, they in part also provide professional expertise in a topic still new to public administrations, giving them the role of a support unit. This role is somehow diminished though, because they rarely withhold their support and their interests can therefore not be considered urgent.

UK is the most well-developed country in open data, which can be seen from the rankings in Open Data Barometer (Davies, 2013) and Open Data Index. Heimstädt, et al (2014) has conducted a timeline analysis of the open data ecosystem development in UK. The results have concluded that the last 15 years have “shaped the UK’s Open Data environment into an Open Data ecosystem”. Clear signs have been identified as the features in a digital ecosystem: independent actors as data suppliers, intermediaries and consumers have been contributing to the ecosystem driven by demand. However, the ecosystem is still far from mature and sustainable in that the demand in the ecosystem is not yet “fully encouraging supply” and actors “have yet to experience entirely mutual interdependence”.

3.4 Open data topologies analysis (SOTON)

Currently, there are a few deployments about open data topologies analysis as mentioned in Section 2.2. For Open Data Index, 60 editors and nearly 200 people have contributed to the list of open data sets in different countries according to the evaluation metrics defined in Open Data Census¹⁶. Open Data Barometer took the survey methodology to collect data from experts in open data community. Similar approaches are also applied in the research about open data life cycle and stakeholders as mentioned in Section 3.2 and 3.3. Those research methods can provide a snapshot of open data topologies to some degree, however, the nature of the methodologies limit the scalability and coverage of such analysis because the data is updated by a group of people manually following a predefined manner. Therefore, we need to develop a new method to automatically (or semi-automatically) collect the data about open data stakeholders and life cycle from larger archives with better coverage of involved parties (such as companies, organisations, etc) in open data ecosystem. This methodology is especially useful in the context of ODM as we are trying to implement the monitoring functions with minimal manual interference or intervention.

3.4.1 Methodology

Due to the fact that open data has drawn more and more attention from government, business, politics and citizens, it is becoming a hot topic in everyday life. We can imagine that the term of “open data” has been well-mentioned not only in academic domains, but also in newspapers, TV programmes, social networks (such as Twitter and LinkedIn) and other media, where we can find agents playing different roles in the open data ecosystem, such as data publishers, consumers, etc. Rather than applying structured searches on major search engines or conduct expert reviews to analyse open data topologies, we propose a quantitative method of using large digitised corpus, such as news and social media, to reveal the involved parties in open data and, furthermore, their relationships as indicated on social media. There are several important sources we can investigate, such as Twitter, Google News, blog posts and discussion threads in LinkedIn communities related to Open Data.

News corpus and social media usually contains diverse topics, so we need to filter the resources relevant to open data and extract useful information related to open data topologies. To achieve this goal, we will firstly filter the text resources from those social media applications by searching keywords related to open data, such as “open data”, “data sharing”, “open access”, “open knowledge”, etc. Tweets can also be filtered by defined hashtags. Then we will apply named entity extraction on the plain-text to identify the mentioned agents such as companies, organisations, projects, etc. Upon aggregating the named entities from the text, and filtering out artifacts (i.e. noise within data), we will be able to clearly map the major agents involved within the open data ecosystem. Furthermore, we can automatically discover the open data portals or data sets in this way as the input for the data harmonisation and visualisation functions in WP3.

¹⁶ <http://national.census.okfn.org/>

This proposed methodology is an automatic way of content analysis using the data on the web. Named entity recognition to extract key information from news and social media has been applied in various areas. For example, named entity recognition has been used to extract “who, what, where and when” from real-time news articles in different languages in Europe (Atkinson, 2009) and from digitised newspapers in the Europeana project¹⁷. It has also been applied to detect real-world incidents or crises from Twitter (Abel, 2012) and improve health-related information retrieval from medical social media (Denecke, 2009). So we can expect that this method can identify who are involved in open data topologies and what they have published and consumed.

For ODM, another function that can be implemented based on this methodology is monitoring the updating of the data sets, i.e. detecting events of data set updating from social media. We assume that, when new data is available in the data set, the publisher will broadcast the news via online press release and social media channels. By analysing the text, we can get to know which data set has been updated and thus trigger ODM to crawl the new data set. In the next subsections, we will present some preliminary results by applying the above methodology in Google News.

3.4.2 Open data topology analysis using Google News

We use a Google News live crawling service¹⁸ to setup a Google News archive and in order to collect Google News that related to open data, we have defined the following keywords to filter the Google News archive: *open data, open sources, open access, open license, data sharing, open information, open knowledge and open society*.

For this study, we have selected the news between 6th March and 19th May 2014 and filtered them with the keywords defined above. As a result, there are 3,052 news articles collected from the crawling service. Then, we put all the news article bodies through Alchemy API, which is a natural language processing and named entity extraction service. Given the plain-text, Alchemy API will output the entities appeared in the text and specify the type of each entity, such as company, organisation, person, city, country, etc. If possible, each entity, will be provided with disambiguation URIs linking to DBpedia, Freebase, etc. In total, 22,857 named entities are extracted from the news articles with some duplicated entities appearing in more than one news article. Table 3 is an overview of the number of named entities in each type.

Table 3. Named entity type and frequency extracted from Google News

Entity type	Entity frequency with duplication
Anatomy	7
Automobile	4
City	1820

¹⁷ <http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/>

¹⁸ <http://newsfeed.ijs.si/>

Company	3321
Continent	18
Country	348
Crime	41
Degree	166
Drug	51
EntertainmentAward	11
Facility	871
FieldTerminology	1602
FinancialMarketIndex	13
GeographicFeature	374
HealthCondition	135
Holiday	12
JobTitle	1156
Movie	15
MusicGroup	2
NaturalDisaster	5
OperatingSystem	52
Organisation	4823
Person	6890
PrintMedia	505
Product	12
ProfessionalDegree	7
Region	119
Sport	18
SportingEvent	3

StateOrCounty	306
Technology	125
TelevisionShow	4
TelevisionStation	21

The output named entities from Alchemy API are not 100% accurate, but from Table 3, we can at least get a general idea on the major types of entities that are involved in the open data topologies. The named entities could be used to analyse the open data topologies in many ways. Just to give one example for this experiment, we can select the entities of Company, Organisation and PrintMedia as the agents involved in open data topology and get the locations of those agents via the information that can be dereferenced from the disambiguation URIs. Then we can count the agents' location in each country and visualise the results as a heat or choropleth map.

To generate the data for such visualisations, we firstly need to select the named entities of type Company, Organisation and PrintMedia. Then we remove the duplicated entities based on the same disambiguation URI. If Alchemy API does not provide the disambiguation URI for a certain entity, we just simply ignore it as we cannot automatically get the location information from it easily. After the filtering, 2,317 named entities are left with 766 for Company, 1,469 for Organisation and 136 for PrintMedia.

All those 2,317 named entities have disambiguation URIs from DBpedia. So the next step is to query DBpedia and retrieve the location country of each organisation or company if that information is available in DBpedia. Even though all the selected agents have RDF descriptions in DBpedia, but the location information is not available, or at least not explicitly available, for all of them. There are many DBpedia properties indicating the location of an agent. To ease the query process, we only look for the values of five DBpedia properties (headquartercountry, headquarterregion, location, country and locationCountry) in the RDF description of each named entity and construct the query to find location country's name accordingly. As the result, 1,128 out of the 2,317 named entities have successfully retrieved the countries' names.

The last step is to further clean the countries' names and count the agents' number in different countries. There are many countries' names could be combined, such as U.S., USA, United States, and some of them are not country's name that should be removed due to the noisy DBpedia data, such as "47" for example. After the final cleansing, we get a list of countries with the count of companies and organisations. Figure 6 shows the visualised heat map.

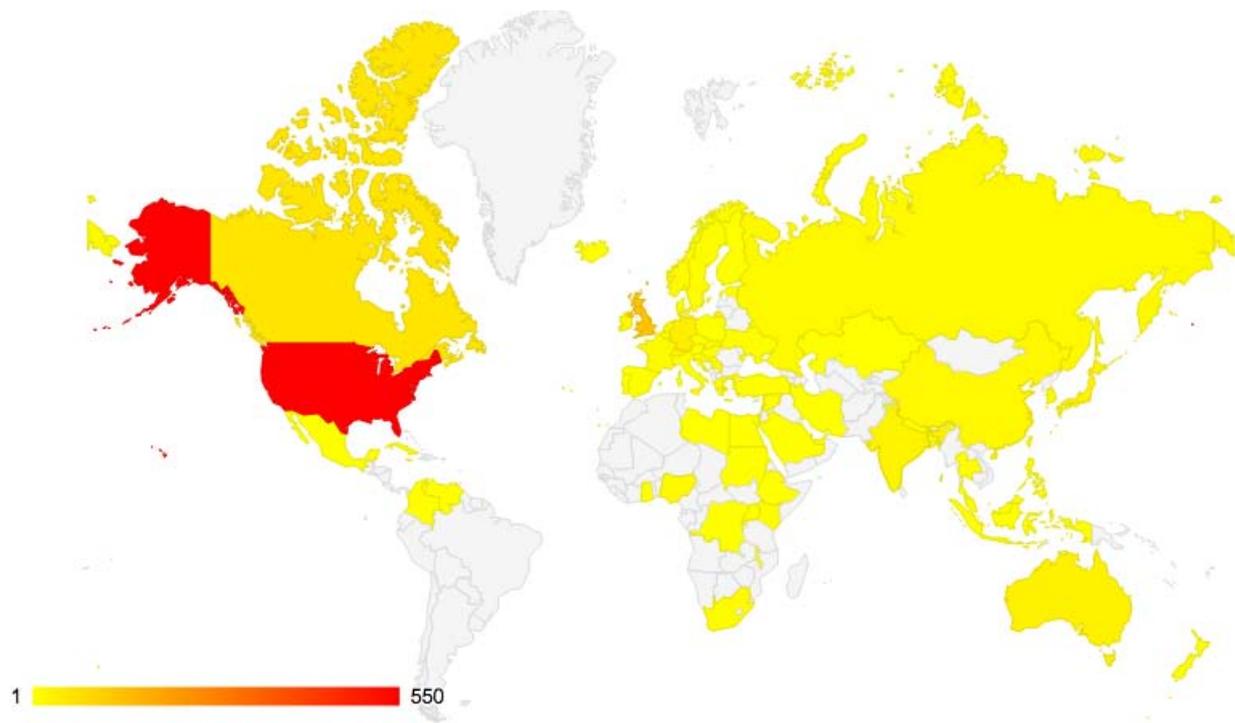


Figure 6. Named entity type and number extracted from Google News

From the results, we can see that United States have the largest number of companies and organisations (550) mentioned in the Google News about open data, followed by United Kingdom (125), Canada (57) and Germany (38). This map shows some interconnections with the heat map in Open Data Barometer 2013. The highest scores in this map also appear in North America and Europe, while Asia and Africa are relatively low. The scores in South America are quite different in the two maps. This might be because this experiment mainly focused on English documents, but as South American countries use Spanish and Portuguese in the press media, the agents in those countries are not available in the news archive we examined.

3.4.3 Limitations

The output of the methodology proposed to analyse open data topology can show some insights into the diversity of the stakeholders and their distribution across countries. Compared with Open Data Index and Open Data Barometer, this quantitative method employs broader raw data, but sacrifices the accuracy and depth of data analysis as each step in this methodology can possibly bring in noisy data. In detail, there are several limitations in the methodology that we can improve in the future.

Firstly, we selected only news written in English as the primary resources for the analysis, which introduced language bias to the results. We are planning to analyse resources in other major languages in EU, such as German, French and Spanish. Secondly, the keywords we have used to filter Google News may introduce irrelevant articles. For example, the keywords “open access” in this article:

<http://www.haverhillecho.co.uk/cmlink/mhep-news-syndication-feed-1-953385>

actually means a clinic is opened and accessible by the patients. So more sophisticated algorithms might be necessary to reject the irrelevant news.

Thirdly, the named entity extraction results from Alchemy API are not accurate, i.e. there are many true-negative and false-positive cases in the extraction results. No named entity service can reach 100% accuracy, so it is inevitable to wrongly recognise an entity or miss an entity, which add noisy data to the final results.

Furthermore, as to the specific example given in this section, the querying of the locations of companies are difficult to process in cases that the location country is not provided in the RDF description. Some locations are presented in other properties, or given as cities or regions instead of countries. Therefore, in this methodology, we have to ignore a large amount of named entities with disambiguation URIs. However, if we go through the companies locations manually, It is possible that the count of companies in each country will be very different.

Even though there are still many limitations to the methodology, this preliminary result have shown some insights into the open data topologies and could guide the development of metrics in D2.3 and the visualisations in WP3. To further extend this methodology, we are also planning to run a similar experiment on Tweets archive. This automatic method will be an on-going effort throughout ODM project to continuously collect and visualise company and organisations' information, and it will be integrated with the ODM framework design and deployment in WP3.

4 THE OPEN DATA LANDSCAPE: A TECHNICAL VIEW (SOTON)

In this section, we will analysis the open data landscape from technological perspective. We will firstly have case studies on the open data portals on different levels and domains. Then we will go through major software and APIs that have been deployed for those portals. Finally, we will review the current metadata standards can how they are related to each other.

4.1 Open Data Portals (SOTON, SYNYO, ATHENA)

Open data portals are the bridge of the data publishers and data consumers, so they are important components in the open data topology. The most well-known and earliest open data portals are deployed by different levels of government. But with the evolution of the open data ecosystem, the data providers become more diverse and the data assets are no longer limited to government related data.

According to the regional analysis on Europe in Open Data Barometer 2013, 4 countries in Europe are listed in the top 5 countries that uptake Open Government Data, and they are United Kingdom, Sweden, Denmark and Norway. Figure 7 shows the average of Europe, Ireland and Greece in each aspects of the metrics used in Open Data Barometer.

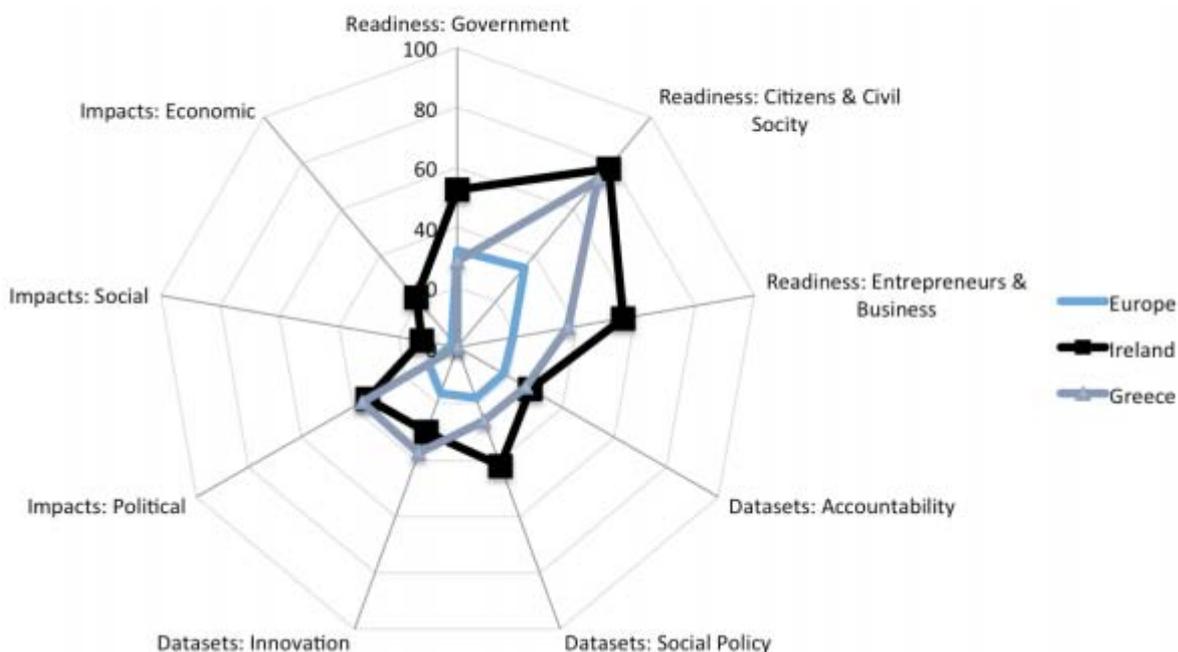


Figure 7. Radar Chart (Ireland, Greece, Europe) of scaled sub-component scores (Davies, 2013)

Currently, there is still no standard rule to categorise the Open Data portals and it is not the main objective of this project to develop such rules in this project. So, in this section we simply categorise the portals in EU by their maintainers and identify the specific domain of the data assets. For this deliverable, we are not trying to list all the open data portals in EU, especially for local government open data portals and the data assets published by NGOs and companies, but rather to offer background knowledge for other deliverables in WP2 on what data sets are available and what should be monitored.

4.1.1 Pan-Europe Open Data Portals

On a European level, there are a few portals officially published as the central hub of EU open data. Examples include the European Union Open Data Portal¹⁹, Europe Public Data Portal²⁰ and EuroStat²¹. Europe Public Data Portal is developed by the LOD2 project²² based on CKAN platform and currently hosts nearly 50,000 data sets harvested from different data catalogues across EU. The European Union Open Data Portal mainly hosts data sets that related to public sectors on EU level. This portal publishes its data set via Linked Data approach and all the data sets' metadata can be accessed via SPARQL endpoints. Compared with the former two instances, EuroStat can be dated

¹⁹ <http://open-data.europa.eu>

²⁰ <http://publicdata.eu>

²¹ <http://epp.eurostat.ec.europa.eu>

²² <http://lod2.eu/Welcome.html>

back to 1953 and its current key role is to supply high quality statistics to the EU Commission and other organisations in Europe.

4.1.2 National Level Open Data Portals

According to the European Union Open Data Portal, 17 members of EU have already developed their own national-level Open Data portals, while many more countries are showing keen interests in setting up their own portals. Data.gov.uk is the very first Open Government Data portal in EU and now it has been expanded to nearly 20,000 different data sets covering various public sectors in UK. Even though the regime and data catalogs are different from one another, major EU countries have published their data via national level data catalogues, including Germany, France, Netherland, Spain, etc. Those portals are usually based on CKAN, but the functionalities of the portal are customised according to the requirements of each country.

Case Study – Open Data Greece

The portal <http://data.gov.gr/> constitutes the central catalog of public data in Greece, providing access to data sets of all public bodies of the Greek government. Its purpose is to increase public access to high value, machine readable data sets by providing uniform services of cataloging, indexing, storage, search and availability of public sector's data and information, as well as, web services to citizens and third information systems. The launch of the portal aims to consolidate all sources of public information available on a single website, which is the focal point of concentration and distribution of public data. Data sets are available with open licenses, allowing further use without restrictions and without cost. Line Ministries are responsible for ensuring the availability of thematic data sets falling under their jurisdiction in the form of open data, giving priority to those with high value and benefit to citizens and businesses. Such data sets may derive from the following thematic areas: financial data, commissions, decisions of the State, taxation and social security contributions, environment, construction activity, investment, culture and tourism, education and research, prices of products and services, efficiency of public services.

Case Study – Open Data Austria

The portal for open data in Austria (<http://www.data.gv.at/>) has the goal of making all open data records in Austria centrally available. Therefore Data.gv.at collects the data in both a manual and an automatic way from the decentralised data catalogues in Austria. The collected data ranges from basic weather information, over to detailed statistical data from various administrative levels. It is possible to browse the data directly on the portal and additionally the data gets submitted to the central European data catalogue, e.g. Public.data.eu. By doing that, Data.gv.at contributes to the open data landscape and supports the idea of linked open data which is a structured way of publishing the data so that it can be interlinked and thus gets even more useful.

4.1.3 Local Level Open Data Portals

Local level portals can host data ranging from region/province, city, district or even smaller area. Compared with national level open data portals, local level portals offer more data on local environment with smaller granularity. Linz open data portal in Austria, Hampshire open data portal in UK, Toscana open data portal in Italy are examples of portals on region level. On city level, major cities in EU, such as London, Copenhagen, Paris, Graz, Rotterdam, etc., have also set up their open data portals. CKAN is also the most popular platform for local level portals. However, considering the technology readiness of different regions and the cost of deployment, many local level portals also choose cloud-based commercial platforms such as Socrata or Junar to host their data

Case Study – Open Data Vienna

In addition to the national open data portal for Austria (<http://www.data.gv.at/>), the City of Vienna maintains an own data portal for its open data. The portal is available through the web site (<https://open.wien.at/>) and it offers data from various sectors specific for the capital of Austria. The availability of the data encourages users to develop innovative applications for the community. At the moment (May 2014) the portal list 143 apps implemented on basis of the provided open data. The data is directly searchable on the website of the portal and furthermore accessible in different file formats.

Case Study – Open Data Tyrol

Tyrol, one of the nine federal states of Austria, offers Open Government Data (OGD) via the portal <https://www.tirol.gv.at/data/>. Due to this approach, the state follows the Open Government Strategy and emphasises the importance of transparency, innovation and participation in public administration. With the focus on the accessibility of the data, the portal provides a barrier-free usage of this data for interested users. Open interfaces and machine readable data formats help to access the data in a cost-free and prompt way. The data catalog of the portal is divided into different categories, e.g. Environment or population, making it easier for users to browse the various data sets hosted on the platform.

4.1.4 Domain specific Open Data Portals

Except for the portals that host data sets in comprehensive domains, there are also portals that focus on data sets in one or more specific domains. For example, the Greece Open Government GeoSpatial Data portal mainly publishes and visualises data with geographical/spatial elements, such as maps, road networks, shoreline and beach boundaries. Transportation for London Open Data hosts real-time traffic data sets (Tube stations, bus stops, real-time bus positions, etc) and they are streamed via Web API. Police forces in England, Wales and Northern Ireland also publish their Crime and

Policing data²³ for users to analyse and build applications. Usually, the domain specific data portals have higher quality data sets in those specific domains compared with comprehensive data portals, their data sets are more structured, which can lead to increased reuse of the data. For example, the Greece geospatial data, data.police.uk crime data and UK land registration data²⁴ are all published as machine-readable Linked Data.

Case Study – Open Geospatial Data Greece

The Open Geospatial Data Greece (<http://geodata.gov.gr>) provides a focal point for the aggregation, search, provision and portrayal of open public geospatial information. It is one of the Greek Government's open government initiatives in the framework of the Open Government Partnership. Further, its operation is included in the Road Map to support the enforcement of Law 3979/2011 for eGovernment, as a best practice example for the application of Information & Communication Technologies (ICT) in the public administration, and as an open data repository for the provision of geospatial information. It also provides technical support to the National Spatial Data Infrastructure, in accordance to the National Strategy for ICT and eGovernment.

4.1.5 Open Data Portals in Private Sectors

Private sectors or businesses are an emerging part in the open data ecosystem, still with a limited interest in comparison with the popularity of the OGD. However, open data in private sectors can lead to new business models and add significant value to businesses. There are some studies about the value of open data in private sectors, such as the Deloitte report (Deloitte, 2012) and McKinsey report (Manyika, 2013) as mentioned previously in Section 2.1.

There is an ongoing project named Open Data 500²⁵, which studies how OGD can generate new business models and products in U.S. Currently, more than 200 companies across U.S. have submitted their surveys and explained how the open data can help improve their businesses. Different from the open data in public sectors, the data publishing platforms in private sectors are highly tailored for the business models and requirements. So each business will usually develop their own platform and API for those purposes.

Case Study – Enel

Enel is the largest power company in Italy and it launched its open data sets, including the sustainability and financial performance data, in August 2011 with three principles objectives: (1) increase transparency and involvement by all stakeholders; (2) to improve the business; (3) to encourage innovation by engaging users and developers to come up with new applications and means of data analysis.

²³ <http://data.police.uk>

²⁴ <http://www.landregistry.gov.uk/>

²⁵ <http://www.opendata500.com/>

4.2 Open Data Software and APIs (SOTON, ATHENA)

There are many existing software or platforms that have been used to deploy open data portals. In this subsection, we will briefly list portal software and their implementations. The detailed analysis about the platforms and APIs will be left to the report for D2.5.

4.2.1 CKAN

CKAN is currently the most widely used open source data management system that helps users from different levels and domains (national and regional governments, companies and organisations) to make their data openly available. CKAN has been adopted by various levels of Open Data portals, and a few popular CKAN instances include publicdata.eu, data.gov.uk and data.gouv.fr.

CKAN provides tools to ease the workflow of data publishing, sharing, searching and management. Each data set is given its own page with a rich collection of metadata. Users can publish their data sets via an import feature or through a web interface, and then the data sets can be searched by keywords or tags with exact or fuzzy-matching queries. CKAN provides a rich set of visualisation tools, such as interactive tables, graphs and maps. Moreover, the dashboard function will help administrators to monitor the statistics and usage metrics for the data sets. Federating networks with other CKAN nodes is also supported, as well as the possibility to build a community with extensions that allow users to comment on and follow data sets. Finally, CKAN provides a rich RESTful JSON API for querying and retrieving data set information.

4.2.2 Socrata

Socrata provides a commercial platform to streamline data publishing, management, analysis and reusing. It integrates many useful features for both portal administrators and end users to manage, access and visualise data sets. For example, The Chicago²⁶ and New York City²⁷ government open data portals are hosted by Socrata.

The platform comprises a series of tools, including an open data portal which stores data in the cloud for users to access, visualise, and share. All the data sets hosted in Socrata can be accessed using RESTful API. This is accompanied by the developer site which documents how to use the Socrata API, including search and filter data sets. In Socrata, users have the ability to customise the data set metadata according to individual's requirements.

4.2.3 Junar

Junar is a cloud-based open data platform with integrated features of data collection, enrichment and analysis. Junar allows the publisher to choose what data to collect and how to present them. It is also possible to determine which data sets are made available to the public and which ones are available only for internal use. The platform also encourages social conversations between open data

²⁶ <https://data.cityofchicago.org/>

²⁷ <https://nycopendata.socrata.com/>

administrators and end user in order to help data publishers to understand what data the end users want and find valuable. For example, the Bahia Blanc City open data portal²⁸ is hosted by Junar.

4.2.4 DKAN

DKAN²⁹ is a Drupal-based open data platform with a full suite of cataloguing, publishing and visualisation features. Compared with CKAN, DKAN is seamlessly integrated with Drupal³⁰ content management system, thus it can be easily deployed with Drupal and customised using different Druapl themes. The actual data sets in DKAN can be stored either within DKAN or on external sites, and it is possible to manage access control and version history with rollback. DKAN provides user analytics and data users can upload, tag, search and group data sets via a web front-end or APIs. In addition, they can also collaborate, comment, and share information via social network integration. The current deployment of DKAN instances include Open Puerto Rico portal³¹ and the city of Cologne portal³² in German.

4.2.5 Open Government Platform (OGP)

The OGP is a set of open source tools that allow any users to “promote government transparency and greater citizen engagement by making more government data, documents, tools and processes publicly available”³³. Its goal is to provide governments to (1) easily publish their data and organise them following the structure of the government; (2) use open source technologies to develop cloud-based structure to share data and encourage the development of applications and services to improve the lives of citizens; (3) create community and interest groups around different topics of open data; (4) engage users with social media, such as Twitter, Facebook and LinkedIn.

4.2.6 QU

QU is an in-progress data platform created to serve public data sets. QU is developed by Consumer Financial Protection Bureau of United States and the goals of this platform are to: Import data in Google-Dataset-inspired format³⁴; Query data using Socrata-Open-Data-API-inspired API (SODA 2.0 API)³⁵; and Export data in JSON or CSV format.

4.3 Metadata Standards (SOTON)

When publishing open data sets, the uploaders or maintainers usually provide complementary metadata. The metadata describes important information about the data set, such as title, license, publication body, update frequency, etc., and the goal of the ODM project is to collect and analyse

²⁸ <http://bahiablanca.opendata.junar.com/home/>

²⁹ <https://www.drupal.org/project/dkan>

³⁰ <https://www.drupal.org>

³¹ <http://abrepr.org/>

³² <http://www.offenedaten-koeln.de/>

³³ <http://ogpl.github.io/index-en.html>

³⁴ <https://github.com/cfpb/qu/wiki/Dataset-publishing-format>

³⁵ <http://dev.socrata.com/consumers/getting-started.html>

such valuable and insightful metadata sources. The current open data metadata standards underpin the metadata harmonisation work in Section 4 and D3.1, which is an important component of the ODM framework.

This subsection will go through the major open data metadata standards published and used by different agents and briefly analyse their relationships and mappings between each other.

4.3.1 DCAT

Data Catalog Vocabulary (DCAT), a W3C recommendation established on 16 January 2014, is designed to “facilitate interoperability between data catalogues published on Web”³⁶. The main goal of DCAT is to improve the data catalogues’ interoperability and make applications easily consume metadata from multiple catalogues.

According to DCAT specification, the main concepts defined are `dcat:Catalog`, `dcat:Dataset` and `dcat:Distribution`, which represents “an accessible form of a data set as for example a downloadable file, an RSS feed or a web service that provides the data”³⁷.

4.3.2 Asset Description Metadata Schema (ADMS)

ADMS³⁸ is a metadata schema created by the EU's Interoperability Solutions for European Public Administrations (ISA) Programme. The goal of ADMS is to help publishers of standards to document the metadata of the standards, such as name, status, theme, version, etc.

ADMS is closely related to DCAT, but the difference in user expectation is the core that distinguishes ADMS from DCAT. ADMS is a profile of DCAT for describing so-called Semantic Assets. DCAT is designed to facilitate interoperability between data catalogs, while ADMS is focused on the assets within a catalog. The core concepts in the vocabulary include: title, alternative title, description, keyword, identifier, document, document/type, document/url, etc.

4.3.3 DCAT-AP

The DCAT Application Profile (DCAT-AP)³⁹ for data portals in Europe is a specification that re-uses terms from DCAT, ADMS, etc., and adds more specificity by identifying mandatory, recommended and optional elements to be used for a particular open data catalogue. Studies conducted by EU commission (Vickery, 2011) have shown that businesses and citizens are facing difficulties in searching and reusing data sets from public sector. Therefore, the availability of a unified method to describe data sets in a machine-readable format with a small number of commonly agreed metadata could largely improve the co-referencing and interoperability among different data catalogues. DCAT-AP is developed under this context and is expected to be applied across Open Data portals in EU countries.

³⁶ <http://www.w3.org/TR/vocab-dcat/>

³⁷ <http://www.w3.org/TR/vocab-dcat/>

³⁸ <http://www.w3.org/TR/vocab-adms/>

³⁹ https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final

4.3.4 CKAN Attributes

CKAN⁴⁰ is the most widely used open data portal software to date, and as such its respective metadata schema is highly relevant to the ODM project. Unlike other W3C standards mentioned above, the CKAN metadata is exposed via RESTful API and data uploaders will need to fill in the metadata with the API request.

CKAN defines three top-level metadata concepts to describe a given data set:

1. **package:** title, notes, tags, revision_timestap, owner_org, maintainer, maintainer_email, etc.
2. **resource:** description format, resource_type, webstore_url, size, etc), group (name, title, type, state, etc.
3. **organisation:** name, id, title ,description, state, etc.

The package, resource and group can be roughly mapped to DCAT as dcat:Dataset, dcat:Distribution, dcat:Catalog and foaf:Agent.

4.3.5 INSPIRE Metadata Schema

INSPIRE is a Directive of the European Parliament and of the Council aiming to establish a “EU-wide spatial data infrastructure to give access to information that can be used to support EU environmental policies across different countries and public sectors”⁴¹. The actual scope of this information corresponds to 34 environmental themes, covering areas having cross-sector relevance, e.g. addresses, buildings, population distribution and demography.

To maximise the interoperability of data infrastructures operated by EU members, INSPIRE proposes a framework using common specifications for metadata, data monitoring, sharing and reporting. INSPIRE consists of a set of implementing rules along with a listing of corresponding technical guidelines. For metadata schema, the INSPIRE Implementing rules include rules for the description of data sets, which could be adopted by open data publishers.

4.3.6 Common Core Metadata Schema (CCMS) in Project Open Data

The Common Core Metadata Schema⁴² is based on DCAT and provides mutual vocabulary that different open data metadata schema can map to. The standard consists of a number of schemas (hierarchical vocabulary terms) that represent things that are most often looked for on the web. CCMS also provide the mappings to their equivalents in other standards⁴³.

The schema is implemented in JSON and CSV format. Similar to DCAT and CKAN, CCMS also defines top-level concepts such as:

1. **dataset:** title, description, keyword, modified, publisher, contactPoint, mbox, identifier, accessLevel, bureauCode, programCode, distribution, etc

⁴⁰ <http://www.ckan.org>

⁴¹ https://joinup.ec.europa.eu/asset/dcat_application_profile/issue/inspire-metadata

⁴² <http://project-open-data.github.io/schema/>

⁴³ http://project-open-data.github.io/metadata-resources/#common_core_required_fields_equivalents

2. data catalog: id, title, description, type, items, etc.

CCMS provides mappings to other major metadata vocabularies, such as DCAT, CKAN and Schema.org. CCMS also develops a Catalog Generator⁴⁴ to help users publish metadata in CCMS format.

4.3.7 Data Catalog Interoperability Protocol (DCIP)

DCAT is the most recent metadata standard that enables the sharing of metadata across different data catalogs. However, the actual implementation of DCAT is still needed to access the metadata and serialize it into different formats. In this context, the DCIP is a specification designed to “facilitate interoperability between data catalogs published on the Web”⁴⁵ and is complementary to DCAT. It provides an “agreed” protocol (REST API) to access the data defined in DCAT. One of DCIP’s main targets is to develop a CKAN plugin to expose CKAN metadata as DCAT, but this work is still in progress.

4.3.8 Vocabulary of Interlinked Datasets (VOID)

VOID is an “RDF Schema vocabulary for describing metadata about RDF data sets”⁴⁶. Its primary purpose is to bridge the gap between data publishers and data consumers using an exclusive vocabulary to describe different data set attributes. The core concepts related to open data sets are: void:Dataset, void:Linkset, void:subset.

4.3.9 Schema.org

Schema.org⁴⁷ is a collection of schemas (in RDF/Microdata format) that webmasters can use to markup HTML pages in ways recognised by major search engines. Schema.org covers many domains and there are classes and properties defined as DataCatalog and Dataset. The metadata harvester within the ODM project can make use of schema.org vocabulary to discover the data sets and data catalogs hosted in a certain website.

4.3.10 Google Dataset Publishing Language

Google Dataset Publishing Language⁴⁸ is a “representation language for the data and metadata of data sets”. Data sets described using this format can be visualised directly from Google Public Data Explorer⁴⁹.

4.4 Use of metadata standards by portals, software and APIs (SOTON)

We have introduced open data metadata, software and APIs in the previous sections. To clarify which metadata standards are applied by which software, portals or APIs, we provide Table 4 to illustrate

⁴⁴ <http://project-open-data.github.io/catalog-generator/>

⁴⁵ <http://spec.datacatalogs.org/>

⁴⁶ <http://www.w3.org/TR/void/>

⁴⁷ <http://schema.org>

⁴⁸ <https://developers.google.com/public-data/>

⁴⁹ <https://www.google.co.uk/publicdata/directory>

the adoption of different metadata standards. Not all the standards listed in Section 4.3 have widespread adoption in the open data domain as some of them are developed for metadata mapping, such as CCMS and DCIP, and some of them are not designed for open data, such as VOID and Schema.org.

Table 4. Metadata adoption for open data software, portals and APIs

Metadata Standards	Adoption
CKAN attributes	CKAN instances, such as data.gov.uk, Open Data Vienna, etc.
DCAT, DCAT-AP, ADMS	publicdata.eu, open data portal of local government of Gijon ⁵⁰ , DCAT is also supported by Socrata
INSPIRE metadata schema	INSPIRE GeoPortal, UK Location Infrastructure ⁵¹ (part of data.gov.uk)
Google Dataset Publishing Language	QU platform

5 CHALLENGES AND SOLUTIONS (SOTON)

This section summarises the outstanding challenges that the open data community faces and the possible solutions to these challenges from technical point of view. The challenges and obstacles will be reflected in system design in D3.1 and will guide the benchmarking of ODM for D3.8.

5.1 Data discovery (SOTON, ODI)

Data discovery is the precondition of effective data analysis. However, with more and more agents involved in the open data ecosystem and large amounts of data sets published in different levels, it is becoming increasingly challenging to discover what data sets are available, which data sets are recently published and what data sets are recently updated. Any open data monitoring framework needs to know where the data sets are located and how to collect them for analysis. So there is a challenge to develop and maintain a data **catalogue registry** with a reasonable coverage of catalogues, data sets and even distributions. The registry will not only record the URL of the catalogues, but also group and categorize them in a logical way.

As a starting point, datacatalogs.org has developed a catalogue registry based on contributions from open data experts. However, the registry can be improved in many ways. Firstly, comprehensive metadata attributes could be included in order to show more information about the catalogues. Currently, there are only five attributes applied in datacatalogs.org, and they are homepage, description, publisher, metadata license and spatial coverage. Secondly, we could add automatic and

⁵⁰ <http://datos.gijon.es/>

⁵¹ <http://data.gov.uk/location>

crowdsourcing mechanisms to help the discovery of data catalogues, while validating the quality of the entries in the registry. Thirdly, awareness of the existence of an open data catalogue does not mean we have fully discovered the data sets hosted by this catalogue. Most of the time, data consumers are looking for a specific data set, or even a specific distribution. So it is more important sometimes to keep the data sets and distributions registered in the catalogue registry for search and analysis.

Another challenge for data discovery is to update the evolution of open data catalogues. When a new data set is published in the catalogue or a data set has been updated, we need to update the corresponding entries in the catalogue registry. To measure the evolution of the data, we need to develop a matrix for such measurements and make it clear on what the matrix means to different stakeholders. Keeping track of the evolution means we must keep the metadata up-to-date and tidy up the historical data. To realise this function, we need to design a **Job Manager**, which triggers the collection of metadata in a periodical manner or from external signals. We also need a **versioning system** to track the changes of the data sets and keep the provenance of the data sets for future validation.

5.2 Data awareness and insight (SOTON, ODI)

With the rapid growth of the open data ecosystem, it is now a challenge to obtain insights of such growth and extract useful information, such as evolution trends and possible gaps, from those data sets.

In detail, we need to define the measures against the data sets and what metrics are useful to show some insights about the data sets. As the requirements of data analytics vary from stakeholder to stakeholder, we need to select the correct attribute(s) from the data sets and provide appropriate visualisation or dashboard functions based on different stakeholders' requirements. So a systematic research on open data metrics and visualisation methods will be necessary to reveal what need to be measured and how. This work will be carried out in D2.3.

To analyse the data sets, we firstly need to harmonise the metadata under one mutual schema. As we discussed in Section 4, there are many existing metadata standards applied in different open data catalogues, so a **metadata harvester and metadata harmonisation engine** is necessary to offer a unique view of the metadata. The open data catalogues may apply different platforms, APIs and access control, so the metadata harvester needs to be flexible and adaptable enough to those implementations. The harvested metadata will follow different schema, and as the metadata is provided manually, it is likely that there will be large volumes of noisy data, so a metadata harmonisation engine will be necessary to select attributes in source schemas and map them to the harmonised target schema. After metadata harvester and harmonisation we need an **analysis engine** to evaluate the quality of the data and restructure the data for data analysis and visualisations.

5.3 Metadata harmonisation (SOTON, ATHENA)

5.3.1 Methodology

In order to better understand what metadata standards and attributes are most widely provided by the publishers in the open data landscape, we develop the following methodology to collect metadata from different open data portals and analyse the possible values for different attributes in the metadata schema. The aim of this methodology is to collect enough metadata and provide an empirical overview on metadata harmonisation and possible ways of visualisation. This methodology offers guidance on WP3, so the accuracy and coverage of the metadata are not the main focus.

As CKAN is the most widely used platform to host and publish open data and all the metadata of data sets are exposed via standard REST API, we treat CKAN instances as the major metadata resources in this methodology. We firstly obtain a list of CKAN instances from <http://ckan.org/instances>. At the time of accessing this list, there were 70 CKAN instances registered from all over the world. As CKAN is only a CMS system, it still depends on the implementers to decide whether or not to publish the data set metadata via API. As a result, not all of them follow the standard CKAN API format⁵². So the second step for us is to manually go through those CKAN instances and find out whether the standard CKAN API is enabled in those open data portals. For those portals, which expose metadata via API, we will develop a programme to automatically crawl the data set metadata and cache them in a local database.

The third step is to go through each attribute in CKAN metadata schema and find out the percentage that this attribute has been used by the publishers, i.e. the attributes that are not left empty by the publishers. By doing so, we will be able to see which attributes are most widely used among open data publishers, and thus we need to consider them as important attributes in the metadata harmonisation. Then we can further analyse the distribution and occurrence of the values appearing in each attribute. For example, we can analyse how many different licenses are applied in all data sets and which one(s) are most popular.

5.3.2 Metadata Status for CKAN Instances

In total, there are 70 CKAN instances from all over the world in the registration list, and they can be divided into three groups: continent, national and local level portals. For all the instances, we have manually examined how the data sets in portal are exposed, so that we can collect the metadata. The results show that:

1. 52 out of 70 instances have followed the standard CKAN API, which means that ODM can programmatically crawl the metadata via the API.
2. 2 out of 70 use other formats of metadata, such as Open Archive Initiatives. In order to use their metadata, we need to further examine the specific metadata format in each particular portal.

⁵² <http://docs.ckan.org/en/latest/api/index.html>

3. 16 out of 70 do not expose any API and CKAN is simply used as a CMS. In this case, ODM can only scrape the metadata from the HTML page.

A detailed analysis of all the CKAN instances is not the main focus of this deliverable, so we only take the 52 instances that expose their metadata via CKAN API for a quick overview on the status of different attributes in open data metadata.

We develop a routine to automatically crawl all the metadata about data sets in a CKAN instance and cache them locally for further analysis. Among the 52 instances, 42 of them are freely accessible, 2 of them need private API keys and 8 of them have some unknown server-side problems that prevent us to crawl the metadata. We automatically crawled the metadata from those 42 instances and as a result, approximately 117,000 packages' metadata were collected. The portal with the largest number of data among those 42 instances is publicdata.eu with 48,000 date sets (packages). The metadata collected from all CKAN instances are all in JSON format, so we choose MongoDB⁵³ to cache the data and the total size of the metadata is around 4GB.

5.3.3 Availability of Different Metadata Attributes

After collecting the all the data sets metadata from the 42 CKAN instances, we firstly count the availability of the metadata attributes in order to see which attribute(s) are mostly provided by the data set maintainers and which attribute(s) are usually missing. There are in total 180 different metadata attributes provided by CKAN that we investigate. The results are shown in Figure 8, where the Y Axis is the name of the attribute and the X Axis is the availability percentage of such attribute in the metadata of the 117,000 data sets.

From the result, we can see that title, maintainer, license (or license id), tags, author and url are commonly available in each data set's metadata. This means that those attributes are most widely used by data set maintainers in, at least, CKAN and should be considered as important attributes in the metadata harmonisation. From another point of view, they are important indicators of the metadata quality that ODM should monitor. One thing we must emphasise is that even though the value of a certain attribute is provided by the data set maintainer, it could still be noisy data. For example, we have found that the titles of many data sets are series of numbers, which are meaningless.

Table 5 shows the attributes that are relevant to the ODM project but are usually missing by the data set maintainers. Less than 40% of the data sets have organisation information, which is a critical attribute to reflect the authority of the data set to end users according to the Open Government Stakeholders Survey (Martin, Kaltenbock, Nagy, & Auer, 2011). The "update frequency" and "last update date" attributes are very useful for ODM to periodically harvest the new data. If those attributes are missing, it will generate significant overload and uncertainty to the ODM system as some data sets will be out-of-date and some of them may not need to be checked frequently.

⁵³ <http://www.mongodb.org/>

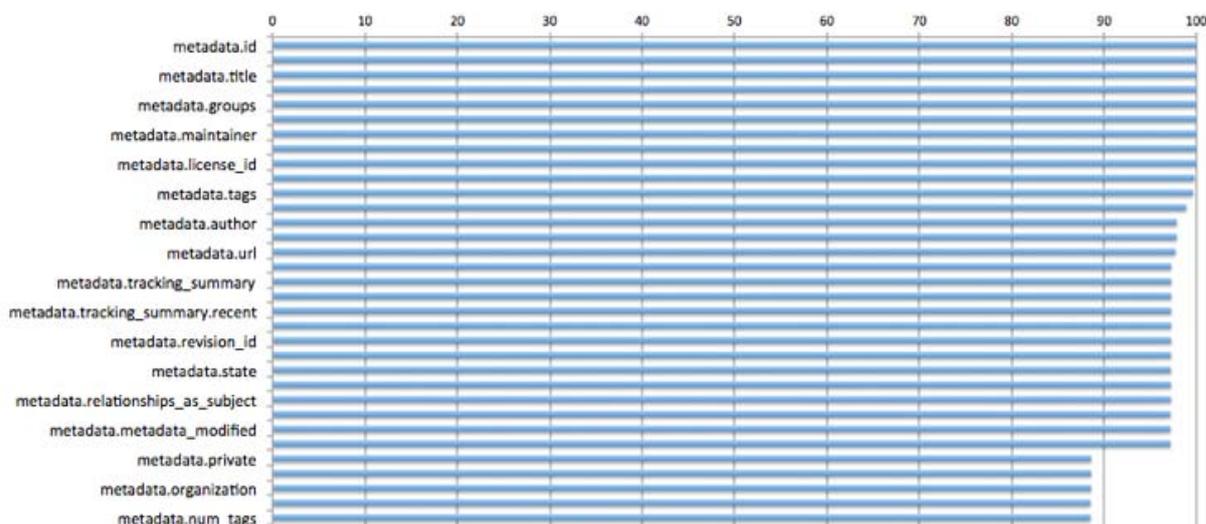


Figure 8. Availability of metadata attributes

Table 5. Important attributes that are usually missed

Attribute Name	Availability Percentage
Organisation information	38.9%
Temporal coverage	2.11%
Spatial coverage	7.35%
Update frequency	8.00%
Last update	7.36%

For ODM, we are going to select several attributes as the main target for analysis and the detailed design of such analysis will be described in D2.3. Here, we have chosen four attributes (see Table 6) to carry out a preliminary analysis in order to guide the design and deployment work in WP3.

Table 6. Preliminary analysis of the CKAN attributes

Attribute Name	Top Values (No. of data sets)	Relationship to ODM
License	<ol style="list-style-type: none"> uk-ogl (20899) cc-by (19327) cc0 (7143) dl-de-by-1.0 (5874) other-nc (5253) 	<p>There are 180 different license values applied to the metadata of the all the data sets. Some of them are identical, but they are represented in different ID or languages. For example, “CC-BY” and “Creative Common By” are the same license with different names in different portals. So ODM could develop a programme to harmonise</p>

		those values and further show to end users on the permissions of different licenses. Furthermore, we can visualise the distribution of licenses over the data sets and show how the usage of the licenses are evolving over time.
Maintainer	<ol style="list-style-type: none"> 1. Ivan Begtin (5793) 2. Statistisches Bundesamt (3170) 3. Statistische Aemter des Bundes und der Laender (1774) 4. Bundesamt für Statistik BFS (1676) 5. Statistisches Landesamt Rheinland-Pfalz (1157) 	The maintainers of the data set could be individual persons or organisations. ODM could identify who they are, where they are and further map them to the stakeholder groups defined in Section 4.1.
Tags	<ol style="list-style-type: none"> 1. Transportation (6670) 2. Water (6651) 3. elevation (6560) 4. hoogte (6351) 5. Landschap (6281) 	There are more than 38,000 different tags that have been used to describe the data sets. ODM can use the tags for data categorisation and index. We can also develop the guidance on how to efficiently assign tags values so that understandable by end users.
Data format	<ol style="list-style-type: none"> 1. CSV (70446) 2. XLS (40637) 3. HTML (19201) 4. PDF (27055) 5. XML (10784) 	CSV still dominates the distribution format of the data set. Quite a lot of data formats are not structured and only a few of them are machine-readable (1,969 data sets are in RDF format). ODM can encourage the publication of more structured data instead of images or PDF so that they can be easily interlinked and reused.

5.3.4 Proposed Attributes for Harmonisation

Based on the experimental results shown in the previous subsection, we proposed several attributes that the ODM should harmonise from different metadata schemas (see Table 7), i.e. the Open Data Monitor Schema (ODMS). The attributes are selected from the DCAT-AP as the standard is developed for EU open data portals and described in a machine-readable format.

Attributes like title, tags, publisher, etc. are properties of a Dataset. For example, ODMS attribute “tag” is equals to `dcat:keyword`⁵⁴ and its domain is a `dcat:Dataset`, which means this attribute uses a keyword to describe a dataset. We also select the temporal (“Temporal Coverage”) and spatial (“Spatial Coverage”) attributes defined in Dublin Core Terms in order to collect relevant data for ODM and visualise them in timeline and geospatial maps⁵⁵. Other important attributes, such as license, format and download url defined in DCAT-AP, are properties about a certain distribution of a data set, so their domains are `dcat:distribution`. In Table 7, we only list the attributes’ names and their corresponding definition in DCAT-AP. The detailed mapping between ODMS and other metadata schema will be presented in D2.2 and how those attributes are used as open data measurements will be discussed in D2.3.

Table 7. Proposed attributes for harmonisation

Attribute Name in ODMS	Class/ Property in DCAT-AP	Domain	Range
Dataset/Title	<code>dcterms:title</code>	Thing	<code>rdfs:Literal</code>
Dataset/Description	<code>dcterms:description</code>	Thing	<code>rdfs:Literal</code>
Dataset/Tag	<code>dcat:keyword</code>	<code>dcat:Dataset</code>	<code>rdfs:Literal</code>
Dataset/Last Updated	<code>dcterms:modified</code>	Thing	<code>rdfs:Literal</code> (ISO 8601 Date and Time String)
Dataset/Publisher	<code>dcterms:publisher</code>	Thing	<code>foaf:Agent</code>
Dataset/Contact Point	<code>adms:contactPoint</code>	<code>dcat:Dataset</code>	<code>vcard:Kind</code>
Dataset/Frequency	<code>dcterms:accrualPeriodicity</code>	<code>dcat:Dataset</code>	Thing
Dataset/Spatial Coverage	<code>dcterms:spatial</code>	Thing	<code>dcterms:Location</code>
Dataset/Temporal Coverage	<code>dcterms:temporal</code>	Thing	<code>dcterms:PeriodOfTime</code>
Dataset/Language	<code>dcterms:language</code>	Thing	<code>dcterms:LinguisticSystem</code>
Dataset/Category	<code>dcat:theme</code>	<code>dcat:Dataset</code>	<code>skos:Concept</code>

⁵⁴ <http://www.w3.org/ns/dcat#keyword>

⁵⁵ <http://dublincore.org/documents/dcmi-terms/>

Dataset/URL	dcat:landingPage	dcat:Dataset	foaf:Document
Dataset/Version	adms:version	dcat:Dataset	rdfs:Literal
Dataset/Version Notes	adms:versionNotes	dcat:Dataset	rdfs:Literal
Dataset/Distribution/Download URL	dcat:accessURL	dcat:Distribution	rdfs:Resource
Dataset/Distribution/Format	dcterms:format	dcat:Distribution	dcterms:MediaTypeOrExtent
Dataset/Distribution/License	dcterms:license	Thing	dcterms:LicenseDocument
Dataset/Distribution/Release Date	dcterms:issued	Thing	rdfs:Literal (ISO 8601 Date and Time String)

6 CONCLUSION (SOTON)

The trend of open data is represented by the idea to freely share, reuse and republish pieces of data without restrictions from copyright and other means of control. In this deliverable, we have reviewed the state of the art of open data landscape, identified several challenges faced by the open data community and proposed solutions for those challenges, especially for open data topology analysis and metadata harmonisation.

Open data is a multidisciplinary concept, so in this report, we looked at an overview of open data from both social/economical and technology aspect. In Section 2, we have gone through major studies on the potential economical and social value of open data, and deployment of open data resources all over the world. From those reports, we have an overview that open data is becoming mainstream and obtaining great attention across the world. Following Section 2, Section 3 has reviewed the literatures from both static side (ecosystem and stakeholders) and dynamic side (life cycle) of open data landscape. Based on the stakeholder analysis approach developed in business science, this report has identified five groups of stakeholders in the open data ecosystem: open data generators, support units, open data users, politicians and advocacy groups. To further identify the individual agents involved in the open data topologies, we have proposed a methodology using named entity recognition to extract agents' names from the text content of social media, such as Twitter and Google News. A preliminary experiment and heat map visualisation has shown that such methodology can successfully collect relevant data and show insights into the open data topologies.

Section 4 focuses more on the technology overview of the open data ecosystem. We categorised the open data catalogues in different levels (pan-Europe, national and local) and domains

(comprehensive, domain specific), and offered case studies for each category. We also briefly analysed the catalogues in private sectors, which are emerging in the open data ecosystem. For open data software and API, there are only a couple of dominators that are widely adopted by open data portals. Meanwhile, there is a diversity of metadata standards applied in different catalogues. DCAT is the most recent standards published by W3C and DCAT-AP has been designed as the future standards to improve the interoperability for open data portals in EU.

Based on the literature review in Section 2, 3 and 4, we generally identified several challenges that open data community is currently facing. The growing number of open data portals makes it difficult to efficiently discover the existing data sets and derive useful information from them. The solutions to these challenges, on one hand, lie in a framework to host a registry of open data catalogues and keep them up to date. On the other hand, we need to harmonise the metadata collected from different catalogues and make them ready for data analysis. In order to reveal which properties are important to the data set and what attributes we will use for data analysis in ODM, we designed an experiment to collect metadata from CKAN instances. The results have shown that many attributes are widely available in each data set, such as title, description, tags and license information. Based on this result, we have proposed several metadata attributes listed in Table 7 that ODM will collect and harmonise from each data set.

The literature review and analysis of the state of the art in this report have revealed that current methodology of monitoring a certain aspect of the open data ecosystem is not automatic. As such, it is challenging to scale up such manual methods at a time when the open data topology is becoming more complex and the volume of open data sets is growing rapidly. The current situation calls for a comprehensive technical solution to monitor, analyse, report, and visualisation useful attributes of open data in a more automatic manner with minimal human interference or intervention. The Open Data Monitor project will provide such a technical solution for the open data community.

7 REFERENCES

Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012, April). Twitcident: fighting fire with information from social web streams. In Proceedings of the 21st international conference companion on World Wide Web (pp. 305-308). ACM.

Atkinson, M., & Van der Goot, E. (2009). Near real time information mining in multilingual news. In Proceedings of the 18th international conference on World wide web (pp. 1153-1154). ACM.

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., ... Williams, H. (2012). Managing the life-cycle of linked data with the LOD2 stack. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, ... E. Blomqvist (Eds.), International Semantic Web Conference 2 (pp. 1–16). Springer.

Batley, R. (1994). The consolidation of adjustment: Implications for public administration. *Public Administration and Development*, 14(5), 489–505. doi:10.1002/pad.4230140505

Blair, D. L., & Whitehead, C. J. (1988). Too many on the seesaw. Stakeholder diagnosis and management for hospitals. *Hospital and Health Administration*, 33(2), 153–166.

Borins, S. (2002). Leadership and innovation in the public sector. *Leadership & Organization Development Journal*, 23(8), 467–476. doi:10.1108/01437730210449357

Clarkson, M. (1995). A stakeholder framework for analyzing and evaluating corporate social performance. *Academy of Management Review*, 20(1), 92–117. Retrieved from <http://amr.aom.org/content/20/1/92.short>

Davies, T. (2013). Open Data Barometer 2013 Global Report. World Wide Web Foundation and Open Data Institute. <http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>.

Deloitte (2012). Open data. Driving growth, ingenuity and innovation. Retrieved May, 2014 from <http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-insights-deloitte-analytics-open-data-june-2012.pdf>.

Denecke, K., & Nejdil, W. (2009). How valuable is medical social media data? Content analysis of the medical web. *Information Sciences*, 179(12), 1870-1880.

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *The Academy of Management Review*, 20(1), 65–91. Retrieved from <http://www.jstor.org/stable/258887>

Freeman, R. E. (1984). *Strategic management: a stakeholder approach*. Boston: Pitman.

Haque, M. (2001). The Diminishing Publicness of Public Service under the Current Mode of Governance. *Public Administration Review*, 61(1), 65–82. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/0033-3352.00006/full>

Harrison, T. M., Pardo, T. a., & Cook, M. (2012). Creating Open Government Ecosystems: A Research and Development Agenda. *Future Internet*, 4(4), 900–928. doi:10.3390/fi4040900

Hausenblas, M., & Karnstedt, M. (2010). Understanding Linked Open Data as a Web-Scale Database. 2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications, 56–61. doi:10.1109/DBKDA.2010.23

Heimstädt, M., Saunderson, F., & Heath, T. (2014). Conceptualizing Open Data ecosystems: A timeline analysis of Open Data development in the UK (No. 2014/12). Discussion Paper, School of Business & Economics: Management.

Hyland, B., & Wood, D. (2011). The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web. In D. Wood (Ed.), *Linking Government Data* (pp. 3–26). New York: Springer.

Jones, T., & Wicks, A. (1999). Convergent Stakeholder Theory. *Academy of Management Review*, 24(2), 206–221. Retrieved from <http://amr.aom.org/content/24/2/206.short>

Julien, N. (2012). *Business Opportunities Arising from Open Data Policies*. Imperial College London.

Kickert, W. J. M., Klijn, E.-H., & Koppenjan, J. F. M. (1997). *Managing complex networks: strategies for the public sector*. Sage.

Manyika, J. (2013). *Open Data: Unlocking Innovation and Performance with Liquid Information*. McKinsey Global Institute.

Martin, M., Kaltenböck, M., Nagy, H., & Auer, S. (2011, June). *The Open Government Data Stakeholder Survey*. In OKCon.

Meijer, A. J., de Hoog, J., Van Twist, M., van der Steen, M., & Scherpenisse, J. (2014). *Understanding the Dynamics of Open Data: From Sweeping Statements to Complex Contextual Interactions*. In M. Gascó-Hernandez (Ed.), *Open Government. Opportunities and Challenges for Public Governance* (pp. 101–114). New York: Springer.

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). *Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts*. *The Academy of Management Review*, 22(4), 853–886. doi:10.2307/259247

O'Toole, L. J. (1997). *Treating Networks Seriously: Practical and Research-Based Agendas in Public Administration*. *Public Administration Review*, 57(1), 45–52. Retrieved from <http://www.jstor.org/stable/10.2307/976691>

Perini, F (2013). *Research the Emerging Impacts of Open Data*
<http://www.opendataresearch.org/sites/default/files/posts/Researching%20the%20emerging%20impacts%20of%20open%20data.pdf>

Romzek, B. S. (2000). *Dynamics of Public Sector Accountability in an Era of Reform*. *International Review of Administrative Sciences*, 66(1), 21–44. doi:10.1177/0020852300661004

Scholl, H. J. (2001). *Applying Stakeholder Theory to E-Government: Benefits and Limits* Center for Technology in Government. In *Proceedings of the IFIP Conference on Towards The E-Society: E-Commerce, E-Business, E-Government* (pp. 735–748).

Tennert, J. R., & Schroeder, A. D. (1999). *Stakeholder analysis*. 60th Annual Meeting of the American Society for Public Administration. Orlando, FL.

Weber, M. (1958). *The Three Types of Legitimate Rule*. *Berkeley Publications in Society and Institutions*, 4(1), 1–11.

Van den Broek, T., van Veenstra, A. F., & Folmer, E. (2011). *Walking the extra byte: A lifecycle model for linked open data*. In E. Folmer, M. Reuvers, & W. Quak (Eds.), *Linked Open Data – Pilot Linked Open Data Nederland* (pp. 95–111). Amersfort: Remwerk.

Vickery, G. (2011). *Review of recent studies on PSI re-use and related market developments*. Information Economics, Paris.

Villazón-Terrazas, B., Vilches-Blazquez, L. M., Corcho, O., & Gomez-Perez, A. (2011). *Methodological guidelines for publishing government linked data*. In *Linking Government Data* (pp. 27–49). New York: Springer.

Zuiderwijk, A., & Janssen, M. (2013). *A Coordination Theory Perspective to Improve the Use of Open Data in Policy-Making*. In *Electronic Government* (pp. 38–49). IFIP.

Zuiderwijk, A., & Janssen, M. (2014). Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View. In M. Gascó-Hernandez (Ed.), Open Government. Opportunities and Challenges for Public Governance (pp. 115–135). New York: Springer.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. Electronic Journal of E-Government, 10(2), 156–172.

8 APPENDIX I: NAMESPACE ABBREVIATIONS IN THIS REPORT

Table 8. Namespace abbreviations used in this report

Abbreviation	Full Name space
dcterms	http://purl.org/dc/terms/
rdfs	http://www.w3.org/2000/01/rdf-schema#
dcat	http://www.w3.org/ns/dcat#
adms	http://www.w3.org/ns/adms#
foaf	http://xmlns.com/foaf/0.1/
skos	http://www.w3.org/2004/02/skos/core#
vcard	http://www.w3.org/2006/vcard/ns#