



## Raising Open and User-friendly Transparency- Enabling Technologies for Public Administrations



Project number 645860

H2020-INSO-2014

### D2.1 State-of-the-art Report and Evaluation of Existing Open Data Platforms

(Version 1.1 – Revised Version – 19/05/2016)



WISE&MUNRO



## Document produced by

Organization: Insight Centre for Data Analytics, National University Ireland Galway (NUIG)

Authors/emails: Edobor Osagie, edobor.osagie@insight-centre.org  
Waqar Mohammad, mohammad.waqar@insight-centre.org  
Arkadiusz Stasiewicz, arkadiusz.stasiewicz@insight-centre.org  
Islam Ahmed Hassan, islam.hassan@insight-centre.org  
Lukasz Porwol, lukasz.porwol@insight-centre.org  
Adegboyega Ojo, adegboyega.ojo@insight-centre.org

Subject: State-of-the-art Report and Evaluation on Open Data Portals

Due date: 19 May 2016

**Dissemination level:** [Select among Public PU **[X]**, Confidential CO, Classified CI]

## Reviewed and approved by

Date	Name	Organization
20.5.2015	Stephan Grimmelikhuijsen	S.G.Grimmelikhuijsen@uu.nl
20.5.2015	Jerry Andriessen	jerryandriessen@gmail.com

## Revision History

Version	Date	Authors	Status	Description of Changes
0.1	02-02-2015	A. Ojo	Outline	Outline of the deliverable
0.2	15-03-2015	E. Osagie	Sections 1, 3 drafted	Drafting of Section 1 - Introduction
0.3	30-03-2015	E. Osagie	Section 3 revised	Revised based on comments from A. Ojo
0.4	10-04-2015	M. Waqar	Section 4 drafted	Drafting of Section 4 – Technical Review of Open Data Platforms
0.5	22-04-2015	M. Waqar	Section 4 revised	Revised with comments from A. Ojo
0.61	26-04-2015	E. Osagie	Sections 1 - 4 and Appendix integrated	Integration of revised drafts of Sections 1 – 4 and the appendix
0.62	02-05-2015	A. Ojo	Executive Summary drafted, and revision of Sections 1 - 4 completed	Drafted Executive Summary and revised fully drafted sections (1 – 4).
0.71	03-05-2015	E. Osagie A. Ojo	Final draft of Section 5 – 8 completed	Drafted and revised section 5 - 8
0.7s	05-05-2015	L. Porwol	Introduction revised Section 3 Revised Conclusions Revised	Few sections rephrased, minor fixes – mainly formatting Conclusions written
0.8	06-05-2015	A.Ojo E. Osagie	Executive Summary revised Section 2,3, 5 Updated	Executive Summary rewritten, Section 2 expanded and rewritten, Section 3 – minor fixes, Section 5 rewritten
0.9	06-05-2015	L. Porwol	Content consolidation and general formatting	All sections checked and consolidated, minor formatting
0.91	22-05-2015	O. Osagie	Content update	Update interviews summary in appendix
0.92	29-05-2015	A. Ojo	Restructure proposal	Restricting proposal contents
0.93	31-05-2015	A. Stasiewicz	Section 3 revision	Detailed editing and restricting of Section 3 – Platform Review
0.94	31-05-2015	A. Ojo	Redrafting of Sections 1, 2, 4, 5 and 6	Writing of new Section 4 – Perceptions of Stakeholders and re-writing of Sections 1, 2, 5 and 6.
0.98	01-06-2015	A. Stasiewicz	Consolidation of changes	Merging of Sections
1.0	01-06-2015	A. Ojo	Revision of report	Review of deliverable to produce final version
1.01	12-05-2016	A. Ojo	Revised Executive Summary	Methodology, link to other deliverables and work packages included
1.03	13-05-2016	A. Ojo	Updated Introduction	Introduction Section was update to show link with the other work packages

1.04	14-05-2016	A. Ojo	Section 2 revamped	Major revision to the methodology section to include definition of concepts, discussion of more than one transparency models, clearer articulation of analytical framework and how the evaluation criteria were mapped to transparency qualities.
1.05	15-05-2016	A. Ojo	Section 5 updated	Section 5.2 was updated to indicate stakeholder-specific perspectives on barriers.
1.06	15-05-2016	A. Ojo	New Section 6 included	A new Section called discussion is to discuss the findings
1.1	18-05-2016	A. Ojo	Additional of Interview manuscripts to appendix	Interview manuscript for all 6 interviewees are include as evidence.

## TABLE OF CONTENTS

1	Introduction.....	13
2	Methodology .....	15
2.1	Research Objectives.....	15
2.2	Concepts .....	16
2.3	Analytical Framework .....	17
2.4	Data Gathering.....	19
3	Review of Open Data Platforms .....	27
3.1	Background .....	27
3.2	Characteristics of Open Data Platforms .....	31
3.3	generic Architecture of Open Data Platforms .....	65
3.4	Platforms Extensibility .....	71
3.5	Summary.....	74
4	Perceptions of Stakeholders on Open Data Platforms.....	76
4.1	Barriers to the Use of State-of-the-art Open Data Platforms.....	76
4.2	Solutions and Desired Features for Future Open Data Platforms .....	78
5	Summary of Findings .....	80
5.1	Transparency-supporting features on open data platforms .....	80
5.2	Perceptions on shortcomings of open data platforms .....	81
5.3	Desired features for future open data platforms .....	84
5.4	Extensibility of open data platforms.....	86
6	Discussion .....	87
7	Conclusion .....	88
	APPENDICES .....	92
	Appendix 1: Reports of Interviews with ODP Stakeholders .....	92
	Appendix 2: General summary of ODP features .....	169

## Executive Summary

Opening up government data to the public has been recognized to have a significant impact on enhancing transparency and accountability of public sector entities while promoting new forms of innovation in government and society<sup>1</sup>. Consequently, driven by the European Public Sector Information (PSI) directive; many European Union (EU) member states have launched their Open Data initiatives<sup>2</sup> at different levels of government. Currently, there are over 8,000 datasets available on the EU Open Data Portal<sup>3</sup>. However, barriers challenges such as limited access and use of open data by citizens and third-parties; limited capacity of government agencies to publish new datasets of high value in a sustainable manner; and weak legislative framework to enable ethical reuse of available datasets<sup>4</sup>, have limited the expected returns from these open data initiatives. In addition, there is paucity of guidelines and best practice guide on how public agencies can effectively publish their open datasets and capture some public value from their investment in open data initiatives. All these challenges and innovation opportunities have led to calls for next generation open data infrastructure. Such open data infrastructure among others is expected to support for social interaction over published datasets as a means to increase data and government transparency<sup>5</sup> through the integration of Web 2.0 with traditional Open Data platforms<sup>6</sup>.

The Route-To-PA Project<sup>7</sup>, which stands for Raising Open and User-friendly Transparency-Enabling Technologies for Public Administration; aims to address some of the above challenges associated with City Government open data initiatives through the conceptualisation and design of next generation open data infrastructure as well as the elaboration of a detailed guideline for provisioning a sustainable open data infrastructure and ecosystem.

Specifically, the Route-To-PA project aims to design and develop models, tools, technology artefacts that will simplify and increase access to datasets published on open data portals and also enable citizens to engage on different societal issues by drawing on insights provided from analysis and exploration of available open datasets in different forms. To achieve these objective, the project will deliver three major outputs in collaboration with its five pilot Public Administration (PA) partners: 1) *SPOD* – A Social Platform for Open Data enabling social interactions among end-users drawing on different visualisations of open data, 2) *TET* – a set Transparency Enhancing Toolset that will be designed to extend existing open data platform by a set of features that simplifies access to and analysis of datasets as well as export of different representations of datasets to external platforms including *SPOD*; and 3) *GUIDE* – a set of recommendations on good practices and strategy for Public Administrations to publish high quality datasets and effectively engage citizens to use available dataset for addressing societal issues of interest.

This deliverable D2.1 - “State-of-the-art Report and Evaluation of Existing Open Data Platforms” is produced as the output from task T2.1 (State-of-the-art Investigation). The report is the first in the series of deliverables for Work package WP2 (User and Systems Requirement) that aims to gather the use cases and systems requirements for the major technology artefacts to be developed in WP4 – “Technological Development and

---

<sup>1</sup> Bonsón, E., Torres, L., Royo, S., & Flores, F. (2012). Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly*, 29(2), 123–132. doi:10.1016/j.giq.2011.10.001

<sup>2</sup> Colpaert, P., Dimou, A., Sande, M. Vander, Breuer, J., Van, M., Mannens, E., ... Dimou, A. (2014). A three-level data publishing portal. Athens: European Data Forum. Retrieved from [http://2014.data-forum.eu/sites/default/files/pdf/edf2014\\_submission\\_43.pdf](http://2014.data-forum.eu/sites/default/files/pdf/edf2014_submission_43.pdf)

<sup>3</sup> European Union Open Data Portal, available at <https://open-data.europa.eu/en/data>

<sup>4</sup> Janssen, M., Charalabidis, Y., Zuiderwijk, A., Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits , Adoption Barriers and Myths of Open Data and Open Government Benefits , Adoption Barriers and Myths of Open Data and Open. *Information Systems Management*, 29(4), 258–268. doi:10.1080/10580530.2012.716740

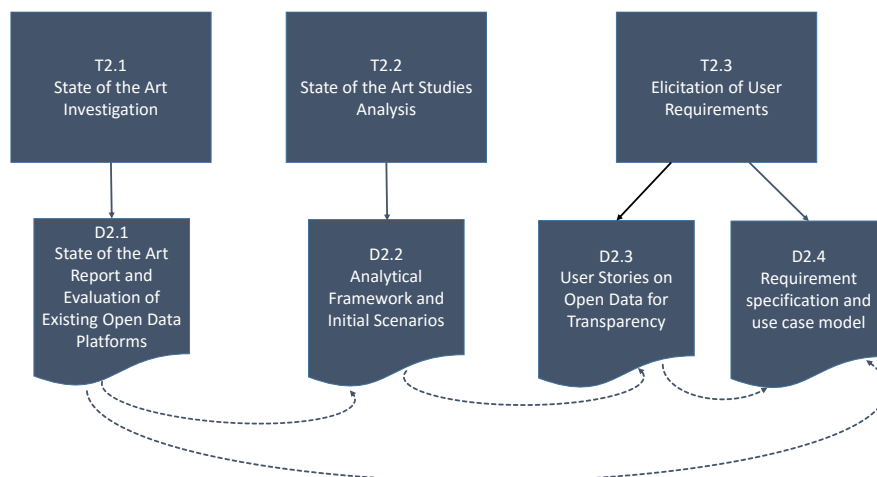
<sup>5</sup> Peled, A., & Science, P. (2012). Effective Openness – The Role of Open Data 2.0 in a Wider Transparency Program. In *3rd Global Conference on Transparency Research, HEC, Paris, France (October 24-26, 2013)* (pp. 44–46).

<sup>6</sup> Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). Designing a Second Generation of Open Data Platforms : Integrating Open Data and Social Media. *E-Gove, LNCS 8653*, 230–241.

<sup>7</sup> <http://www.routetopa.eu>

Integration” which include SPOD and TET. Specifically, results of deliverable D2.1 will serve as input into the choice of base open data infrastructure and platforms to be extended with features described in deliverable D2.4 (Requirements Specification and Use Case Model). In general, this report is useful for open data programme managers and civil society organizations that may have the need to provide a platform for publishing open data.

*Figure 1: Relationship between deliverable and other deliverables in WP2*



Albeit, there are a few existing studies on Open Data Platforms<sup>8</sup>, none of these studies specifically investigates how these platforms support better accessibility and understandability of datasets (i.e. their transparency) published on these platforms. Furthermore, these reports do not also discuss social features nor the extensibility of these platforms.

This report aims to provide some evaluation of existing open data platforms by examining:

- [1] The degree of availability of features that enables Public Authorities and other Open Government Data providers publish high-quality datasets on transparency attributes such as<sup>9</sup>: accessibility, usability, understandability, informativeness and auditability, as well as social interaction and collaboration on datasets;
- [2] The shortcomings of these platforms based on the perceptions of different categories of stakeholders, such as data publishers, data consumers, and mediators
- [3] The platform features, desirable by Open Data stakeholders regarding dataset transparency, social interaction and collaboration on datasets and
- [4] The degree to which these platforms provides extension mechanisms to facilitate the development of additional capabilities.

To answer these questions, the study collected data using four different methods. The first method involved desk research on existing portals and their features and evaluations of these platforms. The desk research was conducted from February 15 to April 30, 2015. The second method involved conducting interviews with six

<sup>8</sup> E.g. the study on Technical Assessment of Open Data Platforms for National Statistical Organisations, 2014, by the World Bank

<sup>9</sup> Cappelli et al, Managing Transparency Guided by a Maturity Model, 3<sup>rd</sup> Conference on Transparency Research HEC PARIS, October 24-26<sup>th</sup>, 2013

experts in the roles of platform developers, open data policy expert, open data publisher, researchers and end users. The interviews were carried out through face-to-face meetings and virtual meetings over Skype from April 27 to 1 May, 2015. The third source of information for the study involved conducting workshops for open data stakeholders in the five pilot locations. These locations include Dublin (Rep. of Ireland) on 17 April 2015, Prato (Italy) on 23 April 2015, Groningen (the Netherlands) on 19 May 2015, Den Haag (the Netherlands) on 11 May 2015 and Issy Les Molineaux (France) on 15 May 2015. In total, 77 stakeholders participated in the workshops across the five locations with 18 in Dublin, 17 in Groningen, 17 in Prato, 17 in Den Haag and 15 in Issy les Molineaux. The stakeholders ranged from platform providers and data publishers (Local Public Admin representative). Technology and open data platform developers, open government researchers, citizen representatives, entrepreneurs, civil society representatives, journalists, Information Manager in City Public Administrations, Census Office representative, open data specialist, software developers, Chief Executives of start-ups. The last source of information is based on results of direct evaluation of instances of selected open data platforms.

The selection of platforms for evaluation in the study is based on two core criteria. The first criterion is that the selected platform must be purpose-built for open data management (not just a web portal framework) with some installed base (information about installed bases of portals is available at [dataportals.org](http://dataportals.org)). The second criterion is the availability of documentation and literature about the platform in the open domain or direct access to the developers (or developer community) of the platforms. A third criterion adopted in the study for selecting platforms to study is related to the availability of advanced features on the platforms. Consequently, the following eleven (11) purpose-built platforms were selected: CKAN, DKAN, Socrata, Junar, DataTank and OpenDatasoft based on the availability of literature and web resources about the platforms; PublishMyData and Information Workbench based on direct access to their developers; while Enigma, Callimachus and Semantic MediaWiki were selected based on their claims of providing advanced features.

To address the first research objective, the platforms were evaluated against a set of 12 criteria that determine the degree to which the platforms support data transparency regarding dataset accessibility and understandability features of the platforms. These criteria include availability of: 1) Metadata, Data and File Format Standards and Schemas, 2) Flexible search facility for datasets, 3) Social Media, Collaboration and Social Sharing tools, 4) Dataset Publishing, 5) Harvesting, Federation and Cataloguing, 6) Data Analysis tools, 7) Visualisation tools, 8) Personalisation tools and 9) Customisation tools, 10) Dataset licensing service, 11) Accessibility and 12) Extensibility mechanisms. These criteria are defined in Section 3. The fourth objective is addressed by considering additional information on whether the platform: 1) is open source, 2) provides concrete extension mechanisms for end-users and developers, 3) provides a guide to support extension activities and 4) allows publishers to customise metadata schemas. Objective 2 is addressed by analysing the barriers contributed by stakeholders that are related data transparency, social and collaboration activities on datasets. Objective 3 is addressed by evaluating the features and solutions to identified barriers and shortcomings of Open Data platforms suggested by stakeholders during interviews and workshop sessions. The findings from the results are as follows:

### **Availability of Features to Support Transparency of Datasets and Social Interaction**

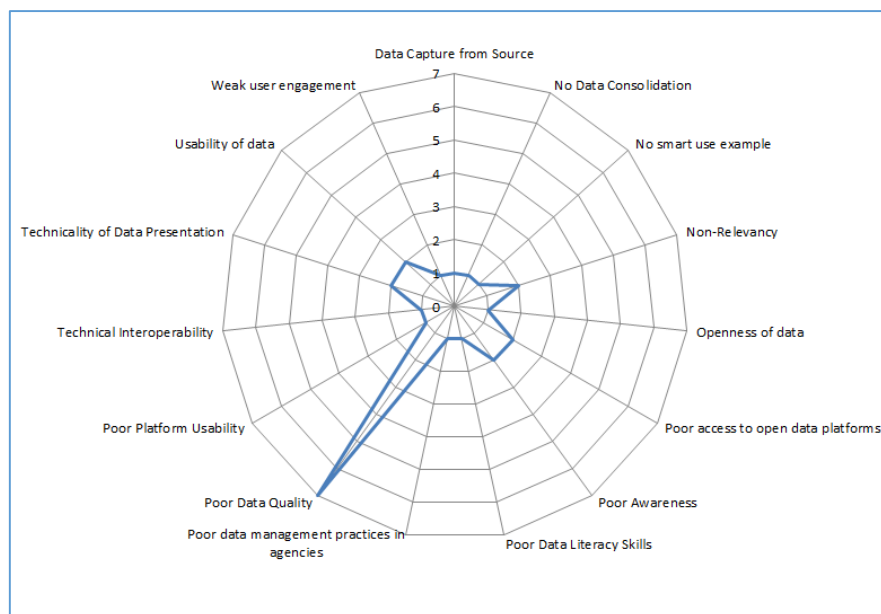
Socrata, CKAN, DKAN and Semantic MediaWiki stand out from other platforms by providing full-fledged features that support at least 9 of the 12 criteria used in the evaluation (see Table 1). Other platforms support between 1 to 7 fully-fledged features. Overall, while the platforms' support for the use of Social Media channels, customisation and personalisation are common features in state-of-the-art platforms, *support for metadata schema adaptation, options for visualisation of datasets and accessibility (including at granular level) to datasets are limited*. However, it must be noted that regarding Social Media integration, these platforms simply allow a link to specific Social Media accounts. Personalisation in the context of this evaluation is only limited to end-user

ability to change the behaviour of the platform based on preferences and does not extend to the aspects like the recommendations of datasets to end-users based on relationships with other users or preferences.

### Shortcomings of State-of-the-art Open Data Platforms based Perceptions of Stakeholders

Our analysis showed that the most common barrier to the use of Open Data platforms and Open Data itself *is the perceived poor quality of datasets* available on the platforms. Poor data quality according to stakeholders is associated with poor metadata, failure to use the right format for different audience and difficulty in locating data of interest. Other barriers identified are related to non-relevancy of available datasets, usability of platforms and data available on the platforms as well as the lack of good examples of prior use of available datasets.

Figure 2: Perceived Barriers to Use and Adoption Open Data Platforms



The figure below presents the associated transparency issues that are related to the above barriers:

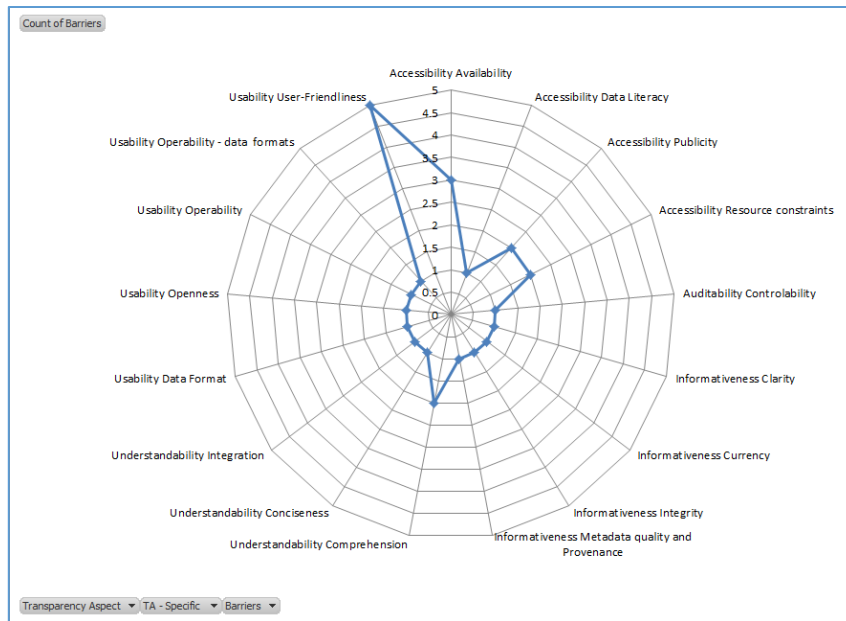


Table 1: Summary of Platform Features

FEATURES	CKAN	DKAN	SOCRATA	PUBLISH MY DATA	INFO WKBENCH	ENIGMA	JUNAR	ODS	CALLIM	DATATK	SMWIKI
DATA, METADATA & FILE FORMAT STANDARDS	●	●	●	●	●	●	●	●	●	●	●
SEARCH & INDEXING	●	●	●	●	X	●	●	●	X	●	●
SOCIAL MEDIA, SHARING & COLLABORATION	●	●	●	●	●	X	●	●	●	X	●
PUBLISHING WORKFLOW	●	●	●	●	●	●	●	●	●	●	●
HARVESTING, FEDERATION & CATALOGUE	●	●	●	●	X	X	●	●	●	X	●
DATA ANALYSIS	●	●	●	X	●	●	●	●	X	●	X
VISUALISATION	●	●	●	X	●	X	●	●	X	X	●
PERSONALISATION	●	●	●	●	●	X	●	●	●	●	●
CUSTOMISATION	●	●	●	●	●	NA	●	●	●	●	●
LICENSING FOR DATASET	●	●	●	●	X	X	X	●	X	X	●
ACCESSIBILITY	●	●	●	●	●	NA	●	●	●	●	●
EXTENSIBILITY	●	●	●	●	●	●	●	●	●	●	●
TECHNICAL ENVIRONMENT	Python	PHP, Drupal CMS	Scala	Ruby on rails	Java & Web apps	NA	Java & Python	NA	Java	PHP	PHP
OTHERS	Good manual Simple to use	Easy to use platform	Tracking & Measure of performance	Flexible, cloud-based, easy to use	R stat, support transparency, linked data	Reliable, scalable, large OD Analyses	Track & measures user impact on OD	Remote web services; easy deployment	Guides, videos, tutorial. Linked data	Deal with fraud, aids transparency	None

● denotes full-fledged solution, ● denotes limited solution, x denotes that solution is not provided, NA denotes information not available

Figure 3: Data Transparency attributes related to the Perceived Barriers



### Desired Features for Open Data Platforms Features by Stakeholders

The desired features contributed by stakeholders for the next generation Open Data platforms were captured under two categories: 1) Social and Collaboration, and 2) Understandability, Usability and Decision making needs. Dataset rating and feedback on datasets, Wall style feedback, collaborative curation of datasets, prioritization and voting on dataset requests, reward system and gamification are some of the features expressed under the social and collaborative needs. To enable better understandability, usability and better decision making with next generation platforms, users requested for customisable dashboards, data mining tools and custom visualization tools, support for Linked Data and map based search as well as question and answering features. Figure 3 was generated from the contributed solutions and features to identified stakeholder needs and barriers.

Figure 4: Keywords generated from desired features for Open Data Platforms



## Extensibility of Open Data Platforms

Based on the four detailed criteria for extensibility of platforms, CKAN, DKAN and Semantic MediaWiki are the most extensible providing free and open source codes, rich set of extension mechanisms and open architecture, guide to support developers in building such extensions and support for additional fields in the metadata schema. However, Callimachus and DataTank being open source could also be modified as desired albeit at a much higher cost compared to the above that provide explicit extension mechanisms. The detailed table of extension features is presented in Table 2 below.

## Conclusion and Recommendations

Guided by the findings we conclude as follows:

- 1) That a few state-of-the-art Open Data platforms such as CKAN, Socrata, DKAN, Semantic MediaWiki provide well-developed features to support good data transparency and quality when publishing datasets. While three of these platforms are open-source and provide extension mechanisms, they arguably stand out as choice base platforms for building next generation open data platforms. CKAN, DKAN and Semantic MediaWiki in particular have a very vibrant developer community that could provide the necessary support in any further development of these platforms.

*Table 2: Availability of Extensibility Mechanism in Open Data Platforms*

Platforms	Extensible	Open Source	Extension Mechanisms	Guide Available	Customisable Metadata
CKAN	●	●	●	●	●
DKAN	●	●	●	●	●
Socrata	•	x	●	●	●
PublishMyData	•	•	•	●	•
Information Workbench	•	•	●	x	●
Enigma	x	x	•	x	x
Junar	•	x	•	x	x
Open Data Soft	•	x	•	●	x
Callimachus	●	●	•	●	●
DataTank	●	●	•	●	x
Semantic MediaWiki	●	●	●	●	●

•denotes extensive solution, ●denotes limited solution, x denotes that solution is not provided

- 2) Despite the features provided by some of these platforms as highlighted above, from the end-user perspective, there are still significant challenges that must be tackled for these platforms to be adopted and used as desired by public administrations and other stakeholders. One of the significant barriers is the perceived poor quality of datasets published on these platforms. Consequently, platforms developers would have to directly address aspects of Open Data quality such as poor context and provenance of published datasets and non-viable data feeds. Feature to explicitly rate datasets in different data quality dimensions could be useful in this regard.
- 3) From the stakeholders' perspectives, social features such as dataset rating, voting and wall-style feedback on datasets and advanced analytics tools such as customisable dashboards, custom visualisation tools should be considered in future enhancement of Open Data portals. This is congruent with findings from technical evaluation of state-of-the-art platform features.
- 4) Open and extensible base technology platforms are available for innovation relating the development of next generation Open Data platforms with features described above. In particular, CKAN, DKAN and Semantic MediaWiki are candidate base platform for such innovation activities.

**Keywords:** Open data platform, Open government data platform, data platform, social media, platform, Social platform on open data, SPOD, Transparency Enhance Toolset, TET, Platform

# 1 INTRODUCTION

---

According to the just published European Union Anti-corruption report<sup>10</sup>, corruption is costing the European economy at least €120 billion annually. With public perception of wide-spread corruption in Europe at about 74%, there is clearly an urgent need to restore public trust and confidence across Europe through greater transparency. Transparency in government decision-making and its use of personal data should in general help to build the citizens' trust and improve accountability of policy makers<sup>11</sup>. Transparency obligations in government are increasingly multi-level. On the one hand, citizens have continued to demand that governments surrender information on their workings. On the other hand governments have are also requiring greater transparency from their dependents such as non-profit organizations, and the entities they regulate in the private sector<sup>12</sup>.

In the past few years, Open data programs have featured prominently as an important instrument or tool for improving transparency. Unfortunately, early and most of the current open data efforts which have largely focused on publishing more data failed to enable the desired transparency in its different aspects. In fact, while opening up datasets, processes and decisions of governments are in general are expected to improve transparency, recent studies have shown high-quality transparency depends not only on how visible information is made but on how well it lends itself to accurate inference<sup>12</sup>. Even more recent studies<sup>13</sup> are showing that understanding transparency as a "purposeful relationship" and architecting this relationship towards greater trust will yield better outcomes from transparency initiatives. *For instance, by understanding open data based transparency as a relationship involving releasing of government data by government agencies to citizens for the purpose of informing and involving citizens in government decision making, it enables focus on needs of citizens in terms of what data is important for them and how best to communicate such data to them.* In our opinion a robust and more holistic understanding of transparency as presented above; must underpin the next generation open-data based transparency initiatives. Thus, future open data based transparency programs and the supporting open data platforms must inter-alia ensure that:

- 1) Published data are those that are of value to citizens and other targeted stakeholders,
- 2) Published data can be presented in different forms to different segments of the citizens and public based on their profiles to facilitate better understanding,
- 3) Published data must have adequate contextual information in the form of detailed metadata and provenance information to enable accurate inference of such data. In general, we expect platforms in general to support the open data best practices<sup>14</sup>
- 4) Citizen-friendly platform (e.g. over existing social networking sites) are provided to enable interactions between public and with government agencies around the published data to better support citizens in the correct interpretation and use of the published data.

In response to the above challenges, the Route-To-PA project (Raising Open and User-friendly Transparency-Enabling Technologies for Public Administration) aims to enable the transition into the next generation open data portal by creating tools that will allow citizens to social engage over open data resources. The project also aims to provide tools that could be integrated into existing open data platforms to deliver greater accessibility to and understandability of available datasets. *However, building such tools and technologies requires good*

---

<sup>10</sup> EU Anti-Corruption Report, Report from the Commission to the Council and the European Parliament, February 3, 2014

<sup>11</sup> European E-Government Action Plan - <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0743:FIN:EN:PDF>

<sup>12</sup> Greg Michener and Katherine Bersch, Conceptualizing the Quality of Transparency, 1st Global Conference on Transparency, 2011

<sup>13</sup> Eliezer N. Mishory, Clarifying Transparency: Transparency Relationships in Government Procurement, Government Procurement Seminar, Chris Yukins & David A. Drabkin, November 4, 2013

<sup>14</sup> Data on the web best practices, <http://www.w3.org/TR/dwbp/>

*understanding and evaluation of state-of-the-art open data platforms to determine their capabilities and how amenable they are to the proposed extensions.* While there are a few existing studies on Open Data Platforms<sup>15</sup>, none of these studies specifically address the affordances of these platforms with respect to the quality and transparency of open data published on these platforms about government agencies and public authorities. For instance, the report by the World Bank on “Technical Assessment of Open Data Platform for National Statistical Organizations (NSOs)”<sup>16</sup> evaluated a selection of platforms in use by NSOs or currently considered for adoption including: CKA, DevInfo, DKAN, Junar, NADA, Nesstar, OpenDataSoft, PC-Axis and PC-Web, Prognoz, Semantic Media Wiki, Socrata and PublishMyData (Swirrl). The evaluated features include support for: descriptive metadata; machine readability; anonymous access; data licensing; data attribution; search; open api; static URI; harvesting; federating; public documentation; standards-implementation; structural metadata, OLAP Hypercubes, data endpoints, visualisation and extensibility. While some of these platform features do impact transparency qualities of data published on them, the analysis carried out in the report is not directly related to transparency qualities.

This report addresses this gap by providing a study on the state-of-the-art of open data platforms from the perspective of how they enable greater organizational transparency. Eleven platforms were reviewed and evaluated in this study including: CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, DataTank, OpenDataSoft, Callimachus, DataTank and Semantic MediaWiki. Specifically, this report D2.1 on “State-of-the-art Report and Evaluation of Existing Open Data Platforms” is produced as the output from task T2.1 - State-of-the-art Investigation. The report is the first in the series of deliverables for Work package WP2 (User and Systems Requirement) aiming to develop the use cases and systems requirements for the major technology artefacts to be developed in WP4 – “Technological Development and Integration” which include SPOD and TET. Results documented in deliverable D2.1 will serve as input into the choice of base open data infrastructure and platforms to be extended with features described in deliverable D2.4 - Requirements Specification and Use Case Model.

The rest of the report is organized as follows: Section 2 presents the methodology for the study while Section 3 describes each of the eleven platform based on evaluation criteria described in Section 2. Section 4 summarises information on perceptions of stakeholders on both barriers and desired features of next generation platforms. Section 5 summarises the findings from the study. Discussion and concluding remarks are presented in Section 6 and 7 respectively.

---

<sup>15</sup> E.g. the study on Technical Assessment of Open Data Platforms for National Statistical Organisations, 2014, by the World Bank

<sup>16</sup> World Bank. 2014. “Technical Assessment of Open Data Platforms for National Statistical Organisations,” *World Bank, Washington DC* (available at <http://documents.worldbank.org/curated/en/2014/10/20451797/technical-assessment-open-data-platforms-national-statistical-organisations>).

## 2 METHODOLOGY

---

This section outlines the overall approach for the study specifically, the questions of interest, the analytical framework underpinning the study and details of the data gathering methods.

### 2.1 RESEARCH OBJECTIVES

The aim of the study is to evaluate existing open data platforms particularly based on the needs of the Route-project, which aims to develop next-generation transparency enhancing open data platform by extending one of the existing open source platforms. The study specifically sets to answer the following questions:

- Q1) The degree of availability of features that enables Public Authorities and other open government data providers publish high quality datasets with respect to transparency attributes such as<sup>17</sup>: accessibility, usability, understandability, informativeness and auditability, as well as social interaction and collaboration on datasets;
- Q2) Their shortcomings based on the perceptions of different categories of stakeholders, such as data publishers, data consumer, mediators etc.;
- Q3) The platform features desirable by open data stakeholders with respect to dataset transparency and social interaction and collaboration on datasets and
- Q4) The degree to which these platforms provide mechanisms to allow modification of their behaviour and to facilitate the development of additional capabilities on the platform.

To answer these questions, we adopted the steps below:

- *Determining degree of availability of data transparency-enhancing features* - to answer this question, the platforms were evaluated based on a set of criteria that enable direct and indirect support for dataset transparency and socialisation on datasets. These criteria include availability of: 1) Metadata, Data and File Format Standards and Schemas, 2) Flexible search facility for datasets, 3) Social Media, Collaboration and Social Sharing tools, 4) Dataset Publishing workshop, 5) Harvesting, Federation and Cataloguing, 6) Data Analysis tools, 7) Visualisation tools, 8) Personalisation tools and 9) Customisation tools, 10) Dataset licensing service, 11) Accessibility and 12) Extensibility mechanisms.
- *Perceived shortcomings of open data platforms* – to answer use of this question, we analysed the barriers contributed by stakeholders that are related data transparency, social and collaboration activities on datasets. These barriers are discussed in more details in Section 4.
- *Platform features suggested by Stakeholders* – to answer this question, we analysed the features and solutions to identified barriers and shortcomings of open data platforms that were suggested by stakeholders during interviews and workshop sessions.
- *Extension mechanisms of open data platforms* - The fourth question was addressed by considering whether the platform: 1) is open source, 2) provides concrete extension mechanisms for end-users and developers, 3) provides a guide to support extension activities and 4) allows publishers to customise metadata schemas.

---

<sup>17</sup> Cappelli et al, Managing Transparency Guided by a Maturity Model, 3<sup>rd</sup> Conference on Transparency Research HEC PARIS, October 24-26<sup>th</sup>, 2013





However, over 40% of the existing open data portals are still based on traditional web portal technologies, content management system (e.g. Joomla, Jadu, eZ Publish, ElementCMS, Drupal and Contao) and frameworks as shown in Figure 5.

*Transparency* - There are several definitions for the concept of transparency. These definitions are as simple as “the ability to look clearly through the windows of an institution”<sup>19</sup>. The definition also includes formal ones like “the measure of the degree to which the existence, content, or meaning of a law, regulation, action, process, or condition is ascertainable or understandable by a party with reason to be interested in that law, regulation, action, process, or condition”<sup>20</sup>. Publishing an appropriate set of open datasets provides “windows of different sizes and clarity” into an organization or public administration. Thus, ease of publishing and accessing published datasets on open data platforms potentially impacts the perceived transparency of government.

## 2.3 ANALYTICAL FRAMEWORK

This section describes how open data platform features may directly impact data transparency concerns and indirectly impact organizational transparency. Following this, we elaborate on a transparency quality framework considered suitable for evaluating the data transparency related features of open data platforms.

There are a number of ways to conceptualise transparency<sup>22</sup>: 1) *an Action* – in which transparency on the part of organizations involves the act of granting access or making information available; 2) a *Communication process* – where transparency is conceived as a communication process which occurs when there is information flow, typically bidirectional information exchange; 3) *an Instrument* – in which transparency is used for financial regulation compliance for creating accountability, for generating trust and for creating competitive advantage through customer relationships and product innovation; 4) *an Outcome* – in which transparency can be viewed as both a ‘means’ and an ‘end’ in organization management; and 5) *a Quality* – here transparency is associated with setting standards to facilitate subsequent evaluation and measurement using qualities such as accessibility, usability, understandability, informativeness and auditability.

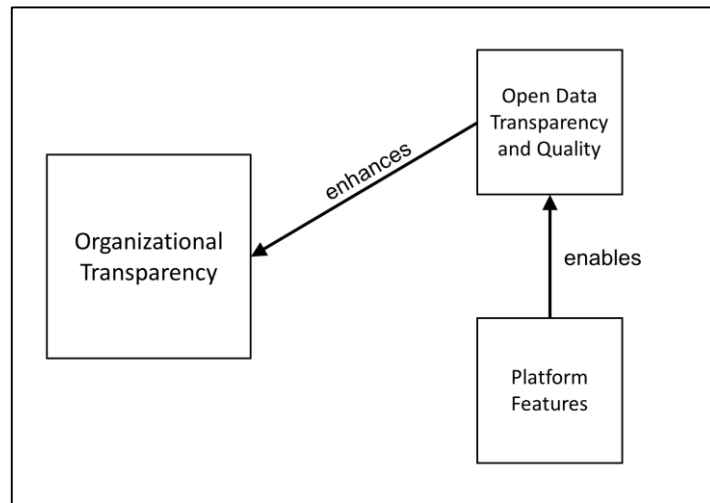
Given our focus on open data platforms, we adopt the notion of transparency as a set of qualities in this study. Thus we are interested in a transparency framework that provides a set of measurable qualities that could be impacted by specific features provided on open data platforms. In our framework, government institutions shares and grants access to data about themselves (open data) which could be evaluated against some set of transparency qualities. In essence, the transparency qualities are measures over datasets on open data platforms. Open data platform features could directly or indirectly (positively) impact the quality of datasets they manage. For instance, a platform providing mediated access to data about an organization could be designed to flag or not allow poor quality datasets to be submitted on the platform. It could also simplify access to the available datasets or make them more understandable for the end-users. Thus, if well designed, open data platforms should enable increased access and understanding of open data describing the state of different aspects of organizations (see Figure 6).

---

<sup>19</sup> Meijer, A. 2009. “Understanding modern transparency,” *International Review of Administrative Sciences* (75:2), pp. 255–269 (doi: 10.1177/0020852309104175).

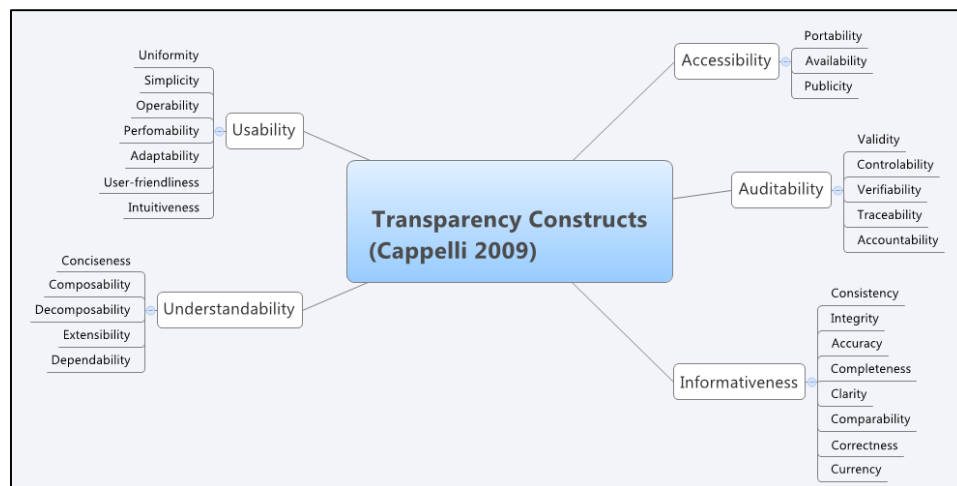
<sup>20</sup> Drabkin, D. A., and Mishory, E. N. 2013. “Government Procurement Seminar Professor Chris Yukins Clarifying Transparency : Transparency Relationships in Government Procurement.”

<sup>22</sup> Mei, C. S., and Dewan, S. M. 2014. “Towards Conceptualizing Information Transparency and its Role in Internet Consumers’ Concerns : A Literature Review.”



**Figure 6: Model for Open-data based Organizational Transparency**

A suitable Transparency model with original in System Sciences is provided by Cappelli et al. (2013)<sup>23</sup>. The model defines a network of 33 qualities that contributes to achieving transparency. The model is expressed using a Non-Functional Requirement (NFR) or Softgoal Interdependence graph (SIG). The assumption in the model is that high-level softgoals can be met by attempting to satisfy lower level ones. According to the model, five major softgoals contribute to the overall transparency quality of information. These are Accessibility, Usability, Informativeness, Understandability and Auditability (see Figure 7). Each of these soft goals is refined lower-level softgoals, for instance accessibility can be enhanced by portability, availability and publicity. Similarly, Informativeness is enhanced through Clarity, Completeness, Correctness, Currency, Comparability, Consistency, Integrity and Accuracy.



**Figure 7: Transparency Construct decomposed into sub-constructs**

Based on these deconstruction we identified features of open data platforms could impact on the above transparency qualities. First we identified a set of relevant features from (World Bank, 2014) – metadata,

<sup>23</sup> Cappelli C et al, Managing Transparency Guided by a Maturity Model, 3rd Global Conference on Transparency Research HEC PARIS, October 24th – 26th, 2013

search, data licensing, harvesting, federation, data analysis, visualisation and extensibility. Other features including Social media, collaboration and social sharing, dataset publishing, personalisation, customisation and accessibility were included based on the goals of the study. These features and the related aspects of transparency they impact are described in Table 3.

*Table 3: Evaluation Criteria and Link to Transparency Aspects*

No.	Features	Description	Related Transparency Aspects
1	Metadata, data and file format standards and schema	Description of the datasets to enable efficient discovery and identification	Accessibility and Contextual understanding
2	Flexible Search Feature	Search function to retrieve datasets of interests based on keywords	Accessibility
3	Social Media, Collaboration and Social Sharing	Function that enables users to share information, discuss and collaborate on datasets	Collective sense making and understandability in addition to increased accessibility through sharing
4	Dataset Publishing	Function to publish a dataset as part of a catalogue and store the datasets if necessary	Accessibility
5	Harvesting, Federation and Cataloguing	Function to load metadata and datasets from external sources into the platform	Accessibility by
6	Data Analysis	Functions to perform analysis on datasets	Understandability through insights from analysis
7	Visualisation	Functions to perform visualise datasets in different forms	Understandability through insights from visualisation
8	Personalisation	Functions that enables users to tailor the behaviour of the platform to meet user-specific contexts such as location, demographic category	Accessibility by reducing irrelevant information
9	Customisation	Function that allows platforms owners to configure features available to end-users by changing styles and including or disabling add-ons	Accessibility by allowing platform providers to change look-and-feel/style
10	Dataset licensing service	Function that allows publishers of datasets to indicate the degree of re-use permitted on datasets	Accessibility though increased reuse
11	Accessibility	Functions that allow end-users with some form of disability to use platforms, for instance in use of colour schemes	Accessibility to end-users with some forms of disability
12	Extensibility mechanisms	Features that enable the development and inclusion of new functions into the platform	Not related to transparency but determines whether platform could be considered as option for base platform for extension to support data transparency features

## 2.4 DATA GATHERING

The study employed four complementary methods for gathering data. The first method involved desk research on existing portals and their features and evaluations of these platforms. The desk research was conducted from February 15 to April 30, 2015. The second method involved conducting interviews with 6 experts in the roles of platform developers, open data policy expert, open data publisher, researchers and end users. The interviews were carried out through face-to-face meetings and virtual meetings over Skype from April 27 to 1 May, 2015. The third source of information for the study are the workshops for open data stakeholders conducted in the five pilot locations including Dublin (Rep. of Ireland) on 17 April 2015, Prato (Italy) on 23 April 2015, Groningen (the Netherlands) on 19 May 2015, Den Haag (the Netherlands) on 11 May 2015 and Issy Les Molineaux (France) on 15 May 2015. In total, 77 stakeholders participated in the workshops across the five locations with 18 in Dublin, 17 in Groningen, 17 in Prato, 17 in Den Haag and 15 in Issy les Molineaux. The stakeholders ranged from platform providers and data publishers (Local Public Admin representative). Technology and open data platform developers, open government researchers, citizen representatives, entrepreneurs, civil society representatives, journalists, Information Manager in City Public Administrations, Census Office representative, open data specialist, software developers, Chief Executives of start-ups. The last source of information is based on results of direct evaluation of instances of selected open data platforms. Detailed information on the data gathering activities are provided below.

*Table 4: Summary of Data Gathering Methods*

Method	Description
Desk research	Employed keywords including: “state-of-the-art”, “evaluation”, “assessment”, “benchmarking” together with terms like “open data portals”, “open data platforms”, “open data infrastructure”; to search for relevant articles and results from the web, Scopus Bibliographic database and Google Scholar. This information was gathered from February 15 to April 30, 2015.
Interviews	Six experts comprising two females and four males were interviewed between April and June, 2015. The interviewee provided insights into challenges associated with the use of existing platforms and desired platform features to address some of these challenges. The interviews were carried out between April 27 and 1 May, 2015.
Pilot Workshops	Five workshops hosted by pilot partners were held in Dublin (April 17 <sup>th</sup> , 2015), Groningen (May 19, 2015), Prato (April 23, 2015), Den Haag (May 11, 2015) and Issy les Molineaux (May 15, 2015). The aim of the interview was to determine stakeholders’ perspectives on barriers to the use of open data and open data platforms. The workshop also aimed at articulating the information, social and collaboration needs and understandability and usability needs of the different categories of stakeholders represented.
Direct Exploration of Platform Instances	Two researchers explored a few instances of selected open data platforms to confirm features specified in the literature about these platforms from March 1 – April 30, 2015. The availability of a set of features were evaluated on instances of 11 open data platforms. Details about the choice of these platforms and selected instances are explained in Section 2.

**Desk Research** – The desk research conducted between February and April 2005 involved systematic literature review of open data literature, review of websites of open data platforms and review of other resources discovered through keyword search on the web. The literature review aimed to analyse past studies on

evaluation or assessment of open data portals and platforms with a view of cataloguing the assessment criteria employed in the respective studies. Scholarly articles were collected from Google Scholar and Scopus bibliographic database using the keywords indicated in Table 4 above. Unfortunately, very few studies (less than 5) were found relevant; thus signalling the lack of scholarly works on the evaluation of open data platforms. However, a similar search on the web produced some notable practitioner-oriented reports like the Work Bank Report on Evaluation of Open Data Platforms for National Statistical Organization<sup>16</sup>, the annual Open Data Barometer report series<sup>24</sup> and the Open Data Toolkit of the World Bank<sup>25</sup>. In addition to obtaining the evaluation criteria, we reviewed platform-specific documentation to have comprehensive information about each of the eleven platform under consideration. The information collected from desk research is used in developing Section 3.

**Expert Stakeholders Interviews** – The expert interview aimed at obtaining the different perspectives of known open data experts on the barriers, solutions, perceived needs and desirable features for next generation open data platforms. Six stakeholders including Open Data Advisor, Data Publisher from Statistics Office, Open and Big Data Researcher, Linked data platform developer, open data consultant and Research institute publishing marine and environment-related datasets were involved in the interview that ran from 12 April to 21 May 2015 (see details in Table 5). All interviews were recorded with the permission of the interviewees received in advance. The transcripts for all interviews are provided in Appendix 1. The analysis of the collected data partly contributed to developing Section 4 of the report.

*Table 5: Interview dates and methods with Expert Stakeholders*

No	Stakeholder	Interview Date	Interview Time	Method
1	Open Data Policy Advisor & Open Data End-user (Ireland)	12/04/2015	10:55am	Face-to-Face
2	Data Publisher from Central Statistics Office, also organizer Annual Competitions for open and public-data based Apps (Ireland)	24/04/2015	11:00am	Face-to-Face
3	Big and Public Data Researcher (Ireland)	27/04/2015	6:00pm	Face-to-Face
4	Open and Linked Data Platform Developer and Entrepreneur (United Kingdom)	14/05/2015	2:15pm	Skype call
5	Open Data Consultant (Belgium)	18/05/2015	1:00pm	Skype Call
6	Marine Public Data Publisher and Platform Provider	21/05/2015	10.15am	Face-to-Face

**Pilot Workshops** – This activity involved conducting workshops hosted by pilot partners in five different locations across four countries in Europe. The workshops were held between 17<sup>th</sup> April and 19 May 2015, with a total of 83 participants involved in the workshops. Table 6 below provides summary of the workshops, while specific organizational information are described under each pilot heading below.

*Table 6: Summary of Pilot Workshops and Stakeholders types*

No	Location	Workshop Date	Number of Participants	Male	Female	Stakeholders Type

<sup>24</sup> <http://opendatabarometer.org/>

<sup>25</sup> <http://opendatatoolkit.worldbank.org/en/odra.html>

1	Dublinked Initiative (Dublin)	17 <sup>th</sup> April 2015 9:30 – 16:30	18	11	7	Platform provider, citizen engagement, technology developer, researcher, data provider.
2	Groningen, Netherlands	19 <sup>th</sup> May 2015	16	11	6	Researcher, PA(policy maker), journalist, PA(Information manager), PA(Open data expert)
3	Prato	23 <sup>rd</sup> April 2015	17			Project contact/facilitator, researcher, open data specialist, representative of local SMEs, census data office, journalist, high school student, SW developer,
4	Den Haag	11 <sup>th</sup> May 2015	17	15	2	PA(project contact), employer, technologist, developed coach-R, researcher, PA(technologist),
5	Issy les Moulineaux	15 <sup>th</sup> May 2015 and 9 <sup>th</sup> July 2015	15			Geographic information system, communication service, social & human resources, association, researcher & CEO start up, CEO-construction industry, CEO-computer graphics, Developers, CEO-social network- community management

*Dublin* - The collective intelligence workshop, held in Dublin on April 17<sup>th</sup> from 9:30 to 16:30, brought together included 18 (11 males, 7 females) expert stakeholders from the fields of public administration, open government, technology, and academia. Table 3 (below) provides a profile of participants. The workshop opened with a presentation which provided details about the Route-To-PA project, as a means of contextualising the day's activities for the participants. Participants were informed that their input, based on their experience, expertise, and needs in relation to open data would be used to inform technology development as part of the Route-To-PA design process. The participants discussed barriers, solution to barriers and developed user stories in three different sessions of the workshop. The profile of the workshop participants are provided in Table 7.

*Table 7: Profile of Participants in Dublin Workshop*

Number	Stakeholder Representation	Organisation
1	Platform Providers /Data experts	Dublinked
2	Citizen engagement	Dublin Community Forum/PPN network
3	Technology Developers	Intel
4	Citizen engagement	Open Government Partnership/ Open Knowledge Foundation
5	Citizen engagement/research	TURAS project UCD
6	Platform Providers /Data experts	All Ireland Research Observatory

7	Technology Developers	IBM
8	Researcher	Trinity College Dublin
9	Data expert/ research	Insight Centre for Data Analytics
10	Researcher	City Share Guide & Global Sustainability Jam
11	Researcher	Callan Institute - National University of Ireland, Maynooth
12	Platform providers	Dublinked
13	Citizen engagement	CiviQ consultation platform
14	Data provider/ Citizen engagement	Your Dublin Your Voice opinion panel
15	Platform providers/data experts	Dublin Dashboard
16	Platform providers/data experts	Fingal Open data and Dublinked
17	Researcher	Trinity College Dublin
18	Technology Developers	ParkYa

*Groningen Workshop* - The collective intelligence workshop held in Groningen, on May 19, 2015, included 16 expert stakeholders including 6 females and 11 males. Eight participants worked for the government as a policymaker, open data, technology or communication expert. All layers of the government were represented: the central government, province and city. In addition, eight citizens participated: (public and private) researchers, a journalist, entrepreneurs and representatives from a citizen movement, social service institute and businesses. Participants were contacted in advance by phone to ask them about their experience with open data. Some participants were experienced open data users, whereas others were experts on population decline, but had not used open data before. Table 8 presents the participant profile of the workshop.

*Table 8: Profile of Participants in Groningen Workshop*

Participant Number	Stakeholder Representation	Type of organization
1	Researcher	University
2	Researcher	Higher education
3	Stakeholder	NGO
4	PA (policy maker)	Province
5	PA (policymaker)	Local government
6	PA (Information manager)	Province
7	Stakeholder	Citizens' initiative
8	Journalist	Newspaper
10	Researcher	Statistical agency
11	PA (policy maker)	Local government
12	PA (communications)	Ministry
13	PA (Open Data Expert)	Ministry
14	PA (policy maker)	Province
15	Stakeholder	Consultancy/research company

16	Stakeholder	Communications company
17	Stakeholder	Consultancy/research company

*Prato* – The Collective Intelligence workshop in Prato took place on April 23<sup>rd</sup>, 2015. The workshop was run with 17 stakeholders including, Open Data specialists, local SME representatives, developers, students, and journalists, among others. At a first stage, all participants (including researchers, open data specialists, journalists, student representatives, and business representatives) worked on their own and reflected upon possible barriers that were reported on a sheet of paper. Then each one illustrated briefly to the audience the two/three barriers considered most important and some debate arose on each topic. In the second session of the workshop participants were involved in identifying Options to overcome the identified barriers for the four most ranked categories. For each category a blank magic board with the title was stuck on the wall. The profile of the participants are given in Table 9.

*Table 9: Profile of Participants in Prato Workshop*

<b>No</b>	<b>Stakeholder Representation</b>	<b>Organisation</b>
1	Project contact/Facilitator	Comune di Prato
2	Researcher / Facilitator	PIN
3	Open Data specialist	Comune di Firenze
4	Representative of local SMEs	Confartigianato (SME organization)
5	Census data Office	Comune di Prato
6	Census data Office	Comune di Prato
7	Representative of local SMEs	Confartigianato (SME organization)
8	Journalist	Press Association
9	High school student	Student Association
10	High school student	Student Association
11	SW developer	Apptec S.r.l. (SW company)
12	SW developer	Mathema S.r.l. (SW company)
13	SW developer	Apptec S.r.l. (SW company)
14	SW developer and Service provider for Pas and business	TT Tecnosistemi ICT company (Representative for business association)
15	Responsible of the City web site editorial staff	Comune di Prato
16	Researcher in ICT application	Mathema S.r.l. (SW company)
17	Researcher in ICT systems for data access and interoperability	C.N.R (National Research Council)

*Den Haag* – This report contains a brief description of the workshop organized in Den Haag, on May 11<sup>th</sup>, in the context of project WP2. 17 participants (2 females, 15 males), including: public administrators, employers, technologists, and researchers were present. At the project level, the workshop was aimed to identify barriers on the access to Open Data and possible options to overcome them, but at the local level, a slightly different goal reformulation was negotiated with the PA representative. Participants profile for the workshop is presented in Table 10 below.



Table 10: Profile of Participants in Den Haag Workshop

#	Name	Organisation	Role
1	Jerry Andriessen	Wise & Munro	Project contact/Facilitator
2	Jan Pieter van de Klashorst	KBM-Alliances	Project contact - PA
3	Louis Wildenberg	Director Wilkohaag	Employer
4	Rob van Leeuwen	Director Van Leeuwen Catering	Employer
5	Heino Walbroek	Director Stichting Marketing Scheveningen	Employer
6	Paul de Jong	Conclusion Digital	Technologist, developed Coach-R
7	Kortekaas	Director Babvios Touringcars	Employer
8	Ben Strijk	The Hague Social Affairs & Employment	PA / Controller
9	Robert Endhoven	The Hague Social Affairs & Employment	PA
10	Martin Wigmans	The Hague Social Affairs & Employment	PA – employer contact
11	Janus	Director LEDconomy	Employer
12	Nathalie Pilk	The Hague Social Affairs & Employment	PA
13	Bob de Jong	Conclusion Digital	Technologist, developed Coach-R
14	Claudio Bolman	Director Bolmancleaning	Employer
15	Mirjam Pardijs	Wise & Munro	Researcher / Facilitator
16	Pim Aerts	The Hague Social Affairs & Employment	Technologist / PA
17	Ron Jansen	Director Baker Tilly Berk	Employer

*Issy les Moulineaux* – Two workshops were planned and carried out, during two hours' duration, for practical reasons of availability of stakeholders. The first one took place in Dijon with young entrepreneurs, France, on the 15<sup>th</sup> of May 2015. The second one took place in Issy-les-Moulineaux, with Public Administrations ("Pas") on the 9<sup>th</sup> of July 2015. These two sessions were exploratory workshops. They allowed us to identify the main expectations of potential open data users and producers in a specific area: business start-ups. The first workshop involved 8 expert stakeholders from the field of information and communication technology. All of them wanted to create a company, or were in the process of doing so. The second workshop involved 7 public administrators of Paris region. They were representative of geographic information systems (they collect, in a database, all cartographic material and manage heritage inventories / compare and disseminate geographic information relating to technical, urban, socio- economic and environmental sectors), representative of associative life (they promote creation and development of local associations) or representative of communication services (they design, in conjunction with other services, communication actions toward general public, media and partners cities). The participants profile for both workshop sessions are listed together in Table 11.

Table 11: Profile of Participants in Issy Les Molineaux Workshop

No	Stakeholder Representation	Organisation
1	geographic information system	Boulogne City
2	Communication service	Paris Region
3	geographic information system	Boulogne City
4	Social & human resources	Boulogne City
5	Association	Issy-les-Moulineaux
6	Communication service	Issy-les-Moulineaux

7	Responsible of Communication service	Issy-les-Moulineaux
8	Researcher & CEO start up	Information and Communication technology (ICT) - Job search platform (Incubator)
9	Researcher & CEO start up	ICT - Job search platform (Incubator)
10	CEO - Construction Industry	"Auto-entreprise"
11	CEO - Computer graphics	"Auto-entreprise"
12	CEO - Social Network - Community management	"Auto-entreprise"
13	Developer	ICSOFIT
14	CEO - Computer graphics	"Auto-entreprise"
15	Researcher & CEO start up	wineConsulting

**Open Data Platform Survey** – This data gathering approach provides a triangulation mechanism for information gathered from platform documentation. Specifically, it enabled the study to have hands-on experience on the use of the purported platform features in literature and product documents. For each surveyed platform, two researchers logged into different instances of each of the 11 platforms and evaluated the platform against the criteria listed in Table 3. These review information for each researcher were recorded in spreadsheet. Once individual reviews were completed by each researcher, the produced evaluations were reconciled to produce a consolidated version. More weight was given to information directly observed than secondary information reported in literature or in product documentations. The information collected was used together with the desk research report to develop Section 3 of this report.

## 3 REVIEW OF OPEN DATA PLATFORMS

---

In this section we present the overview of the open data platforms and findings of the ODP evaluation report. The report has been compiled after survey of 11 major open data platforms being used around the world for publishing of open data, which includes: CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, OpenDataSoft, Callimachus, DataTank and Semantic MediaWiki. The objective of this survey was to explore features provided by existing open data platform, extract common development patterns and to find emerging trends in area of open data solutions. Subsequent sections are the outcome of this technical survey and provides an overview of the features offered by state of art open data platforms, documents the architecture and extensibility of the platforms as well as the design and implementation details of those platforms.

### 3.1 BACKGROUND

The term Open Data Platform (ODP) does not have a universal definition because it is a relatively new concept still under development and not much research and conceptualisation have been done on this field. However, the term “Platform” has a consistent meaning across many different domains where it represents a system defined by three aspects: (1) a stable, low-variety "core", (2) a changeable, high-variety set of "complements", and (3) the interfaces which allow core and complements to operate as a single system (Baldwin, C. Y., & Woodard, 2009). Platform architecture is a related concept defined as "a conceptual blueprint that describes how the ecosystem is partitioned into a relatively stable platform and a complementary set of modules that are encouraged to vary, and the design rules binding on both" (Tiwana, Konsynski, & Bush, 2010).

In the context of ROUTE-TO-PA, information technology platform is a technology infrastructure comprising of the software of a computer ecosystem which determines what kinds of data activities and other possibilities it allows. It encompasses a portal serving as a doorway, a gateway or other entrances such as an internet site providing users the access or link to the resources on the site and/or other sites, and opportunity for users to voice their views or initiate actions (Alexopoulos et al., 2014). The platform in the context of computing typically refers to a computer's operating system; an underlying computer system on which application programs can run (Rouse, n.d.). In relation to this project, open data platforms can be regarded as platforms of standard portals that support the development of applications or systems for the publishing, dissemination, using and reusing as well as sharing the open (government) data by data publishers and consumers alike. ODPs provide spaces for social interactions amount citizens, generation of user metadata and feedback loop for some group of users or stakeholders.

ODPs have made significant contribution in enabling sharing of open data, despite rapid research and development in area; the technology is still in its infancy. Most of the existing open data platforms can be viewed as cataloguing system for open data; they have been extremely useful in kick starting easy publishing of large volumes of open data in diverse data types. But the raw nature of data being shared on these platforms makes it hard for ordinary users to effectively exploit the data shared on these platforms, advanced skills are required to transform the data to appropriate level in which it can easily exploited for analysis and discovery

purposes. Existing open data solutions are missing proper easy to use workflows for extracting and transforming data in machine-readable formats. Existing open data platforms offers search, querying, harvesting, visualizations and limited analysis services but only at dataset level.

Data integration and cross dataset/portal querying and searching is still a challenge. Some platforms have exploited semantic web technology and advance indexing techniques to deal with this challenge to some extent, however more work is required to enable easy integration and exploitation of open data across datasets and portals. Data discovery, fine grain searching, advance analytics and Q&A over open data are essential features required to make open data platforms useable for ordinary users. Existing platform don't allow app development/app marketplace on top open data, API and external tools are normally used to developed applications.

Support of geospatial data standards and tabular formats such as CSV and excel etc. is much better than other formats in most available open data platforms. Basic visualization and analytics being offered by open data platforms is satisfactory. Support for customization, personalization, access control and other configuration features vary across different platforms. DCAT<sup>26</sup> is supported by majority of platforms as format for metadata exchange. Collaboration and sharing is supported widely, either as internal solution or as an extension to platform. Most of the open data solution are either open sources or have community edition with technical support for extensions. The tools and technologies used for the development of open data platforms are quite ubiquitous and easy to learn. In general the documentation provided by most of the platforms is well formed and satisfactory. Majority of the platforms offers the technical support as well as the SaaS features.

The summary below outline the features provided by reviewed open data platforms. Features marked as "Limited" are features that are partially supported by a platform. Below is the definition of the features analysed during the platform review:

**Installed instances:** Indicates the popularity of the platform and the potential community size.

**Metadata, Data and File Format Standards and Schemas:** Data refers to the data that has been stored on the platform or the reference to the external data sources. Usually it is limited to the sequence of numbers, stored somewhere in the memory or in the file system that represents the structure of the data. The most popular formats are XML, CSV, JSON, XLS, PDF, HTML. Term Metadata is the data about the data - about the structure of the data (i.e. keys, indexes, columns), information about the dataset (i.e. title, author, subjects, keywords) and provenance information (publisher, revision history, changes, source of data). The metadata and extend the search capabilities and permits interoperability between different systems. Formats that are machine-readable such as CSV, XML, Geo, XSL etc. can be easily be parsed and interpreted by applications. The data can be stored in structured data store rather than file store for efficient retrieval and querying. Data in RDF format can be easily queried with SPARQL.

**Flexible search facility for datasets:** Search is a powerful and easy to use feature, which lets users to retrieve datasets mentioning the provided keywords. Most of contemporary platforms only provide search capability on metadata associated with the dataset and supports filtering. Emerging platforms such as Enigma<sup>27</sup> is offering more advanced search capabilities such as search at record level granularity and data filtering at multiple levels. Indexing provides more efficient searching and speed up the process.

---

<sup>26</sup> <http://www.w3.org/TR/vocab-dcat/>

<sup>27</sup> <http://enigma.io/>

**Social Media, Collaboration and Social Sharing tools:** This feature is a collection of mechanisms that allow interaction between users. This includes social media tools (i.e. Facebook, Google+, Twitter, etc.) to communicate & collaborate, to comment, review and rate the datasets, share links and so on.

**Dataset Publishing workshop:** Publishing and workflow are all the features and tools offered through the datasets publication process. It may include the data refinement, separation of public / private datasets, files upload via web UI, API or linked to an existing file on the web as well as the access control & addition of metadata to workflow for data upload.

**Harvesting, Federation and Cataloguing:** Federation allows data replication across different instances, and provides seamless integration between different independent portal instances – i.e. by performing a search across multiple instances of the platform. Harvesting feature allows extraction of open data from the open data portals, dumps or other data sources. This feature includes the data conversion to the form required by the platform. Catalogue describes the implemented mechanism for the datasets navigation.

**Extensibility mechanisms:** Extensibility of the platform is expressed by the number of features provided on the platforms to enable adaptation and extension (i.e. provision of APIs and libraries, support for website branding, and connectors, plugins and extensions).

**Data Analysis tools:** Support for data analysis varies between the platforms and basic analysis features are included in majority of platforms. Some platforms offer more advance features such as statistical operations, OLAP, dashboards and analysis widgets etc. In addition Socrata and Information Workbench provide supports for R statistical programming language<sup>28</sup> extensions.

**Visualisation tools:** Basic visualizations such as maps and charts are supported by most of the platforms. Visualizations make use of exiting maps services such as OpenStreetMaps, Google and Bing maps etc; and library such as D3.js<sup>29</sup> and recline.js<sup>30</sup> are commonly used for creating visualizations.

**Personalisation tools:** Personalisation is a set of features that allows: (1) modify the portal look and feel by portal administrators (i.e. branding, logo, colours), (2) customise the portal view to the users (i.e. personalised sorting, auto filtering, proffered view)

**Customisation tools:** Customisation is a set of features allowing the portal administrators to define the metadata standards, portal rules, enable tools and features as well as to configure the data store and limits.

**Others:** All the additional features (i.e. data consumption statistics, overall performance, contextualisation tools and so on).

**Dataset licensing service:** Describes how the licensing information can be added to dataset (i.e. as metadata).

**Accessibility:** It defines how easy it is to access the data. One of option is application program interface (API) - a set of routines, protocols, and tools for building software applications. Platforms export their capabilities by providing APIs to external applications. API provides clear specifications for external applications

---

<sup>28</sup> <http://www.r-project.org/>

<sup>29</sup> <http://d3js.org/>

<sup>30</sup> <http://okfnlabs.org/recline/>

for interaction with the services offered by the platform. Normally API is exposed as REST (Representational State Transfer) or SOAP (Simple Object Access protocol) services.

**Technical Environment:** Describes the working environment and the programming language used while platform development.

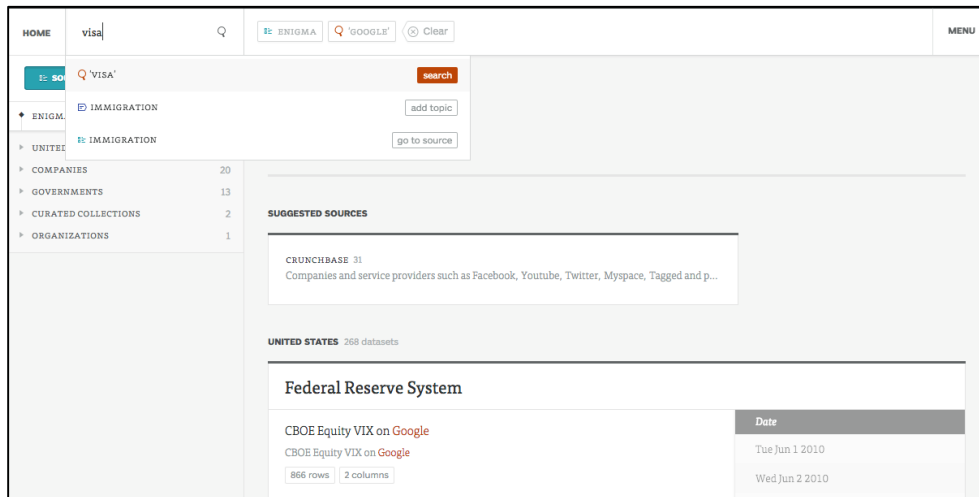


Figure 8: Enigma search user interface

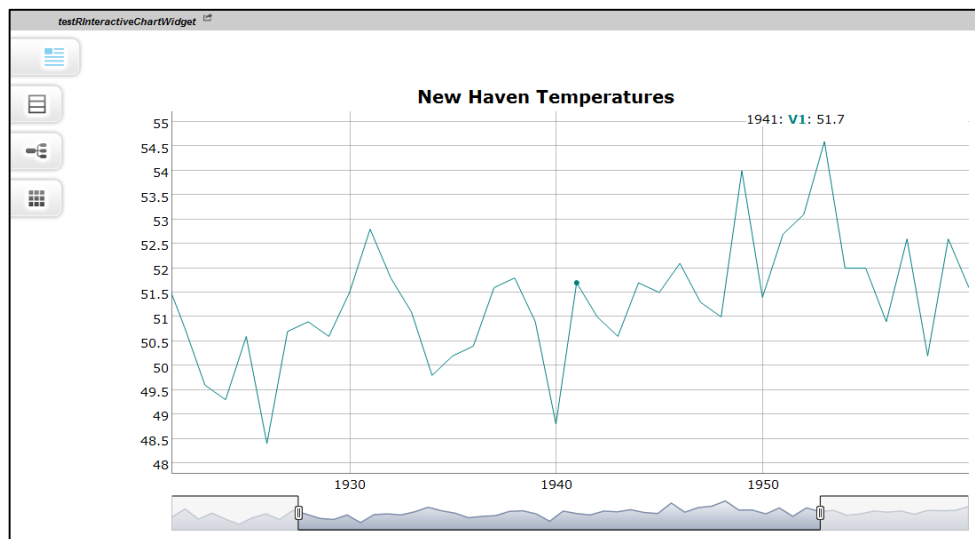


Figure 9: Interactive R chart in Information Workbench.

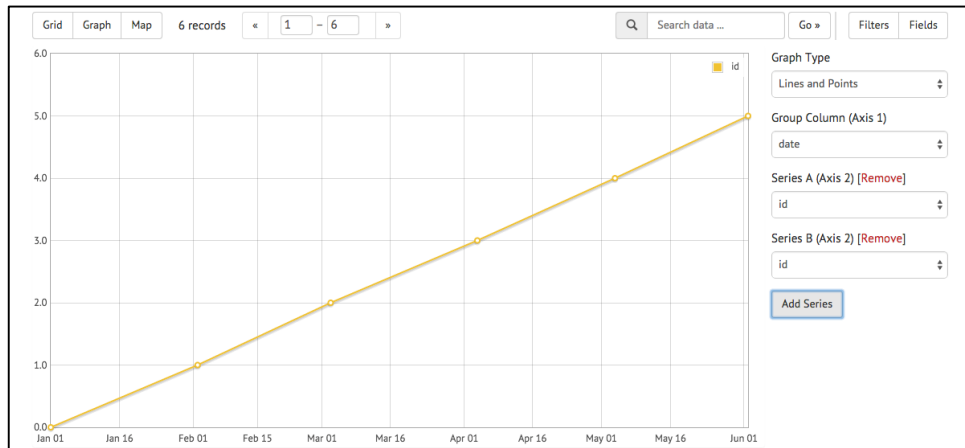


Figure 10: Recline.js visualization library used by CKAN

### 3.2 CHARACTERISTICS OF OPEN DATA PLATFORMS

Desktop research produces a number of existing open data platforms around the world currently offering services to groups of stakeholders. Even though they share a lot in common in terms of aims and objectives, however, there exist a considerable differences in their design, architecture, file formats, features and functions (lemma et al., 2014). The first generation of open data platforms built basically on the paradigm of Web 1.0 have main purpose of making Open Government Data (OGD) available to users rather than offer value-generating functionality on data for them (Alexopoulos et al., 2014). The evolutionary dynamics of platform-based ecosystems and their modules, argues Tiwana et al. (2010) are influenced by the coevolution of the platforms owners' personal perception of the ecosystem; for example, the platform architecture and governance on the one hand and the environmental dynamics exogenous to the ecosystem on the other. In the study of motivation for online community, it was discovered that *"giving back to the community in return for help was by far the most cited reason why people participate"* in online community activities (Antikainen et al., 2010). Furthermore, for open data platform to truly meet its goals, it should be able, by design and architecture, spur usage and enable citizens to engage in discussions and collaborations using the data available on the platform. Citizens should be encouraged to participate, comment and share, not just the data, but the innovative ideas, suggestions, criticisms, grievances and other comments arising from the community as they use the data on the platform and engage with each other in the user community – the public (Antikainen et al., 2010). Unfortunately, due to technical difficulties (requiring a level of expertise) or lack of motivation arising from inadequate supply of user-friendly tools on the platforms, citizens have not been motivated enough to collaborate intensively and extensively on open data platforms (Antikainen et al., 2010). This section evaluates some ODPs such as those in the list below under the various benchmarking features earlier established in the Authors' review section. At the end of this section, a summary of findings is presented in a tabular format. The analysed platforms are as follow:

- Comprehensive Knowledge Archive Network – CKAN<sup>31</sup>
- DKAN<sup>32</sup>

<sup>31</sup> <http://ckan.org/>

<sup>32</sup> <http://nucivic.com/dkan/>

- Socrata<sup>33</sup>
- PublishMyData<sup>34</sup>
- Information Workbench<sup>35</sup>
- Enigma<sup>36</sup>
- Junar<sup>37</sup>
- OpenDataSoft(ODS)<sup>38</sup>
- Callimachus<sup>39</sup>
- Datatank<sup>40</sup>
- Semantic MediaWiki<sup>41</sup>

Open data platform is ICT hub that do not only provides the room for gathering and storing data from the public administration activities and other domains, it also facilitates value improvement of the datasets, use, reuse and sharing of the resources by users. Open data platform is the medium through which open government datasets are made accessible to the public; a platform that assembles the legacy data from various sources and organises them in a manner that supports easy downloading, modification and sharing of the data (Duval & Brasse, 2014).

### 3.2.1 CKAN

Comprehensive Knowledge Archive Network (CKAN) is the largest most well-documented community-based and widely adopted platform in the market (lemma et al., 2014; Lindén & Stråle, 2014). It has one of the best installation procedure manuals with support for any file format. CKAN was developed by the non-profit organisation – Open Knowledge Foundation (OKFN), however, managed by CKAN. In accordance with the above, CKAN claims that it is the world's leading open-source data portal platform delivering a powerful data management system that makes data accessible through the provision of tools to streamline publishing, sharing, finding and using data (CKAN, n.d.). CKAN is aimed at data publishers of any background including national and regional governments, companies and organizations that are interested making their data open and available to the public.

**Features:** As an overview, CKAN's main features include finding and publishing datasets, storing and managing data, engaging with users and other stakeholders, and customisation and extension. Data publishing is done by importing datasets via a web interface, and offers a searching functionality by keyword or filter by tags. This is a rich search experience which allows quick 'Google-style' keyword search and

---

<sup>33</sup> <http://www.socrata.com/>

<sup>34</sup> <http://www.swirrl.com/publishmydata>

<sup>35</sup> [http://www.fluidops.com/en/portfolio/information\\_workbench/](http://www.fluidops.com/en/portfolio/information_workbench/)

<sup>36</sup> <http://enigma.io/>

<sup>37</sup> <http://www.junar.com/>

<sup>38</sup> <http://www.opendatasoft.com/>

<sup>39</sup> <http://www.callimachus.com/>

<sup>40</sup> <http://www.datatank.co.uk/>

<sup>41</sup> [https://semantic-mediawiki.org/wiki/Semantic\\_MediaWiki](https://semantic-mediawiki.org/wiki/Semantic_MediaWiki)



faceting by tags and browsing between related datasets to enable users see available datasets, formats of data and licensing metadata in the search result. Thus it is possible for users to search on all datasets metadata – title, tag and publisher using search options such as:

Fuzzing-matching – allowing searches for closely matching terms instead of exact matches,

Faceted search – allowing a drill-down search via facets (e.g. tags, formats, license and publisher) with the ability to narrow search into specific dataset formats or tags, and

Searching via API – the API search is possible for sort of searching criteria.

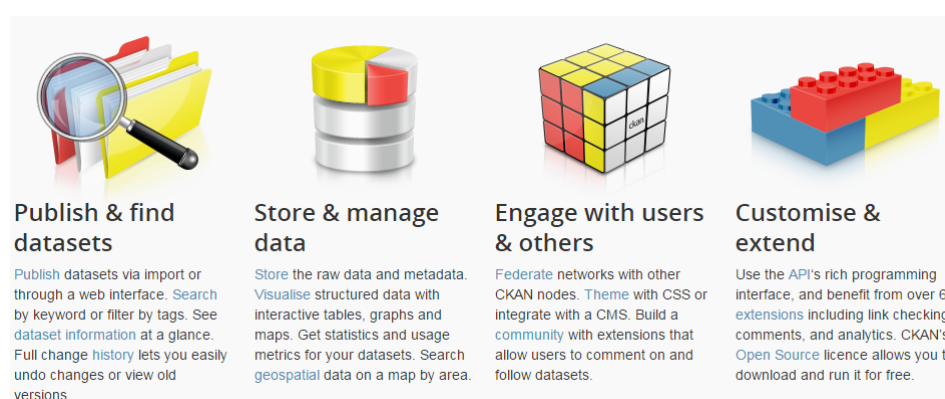


Figure 11: CKAN's main features - extract for ckan.org

**Publishing and managing** data is done on a web interface which allows publishers and curators to register, update and refine datasets in a distributed authorisation model which enable each publisher to maintain their individual data entry and approval. Entry and edit of data can be done in many ways – directly via the web interface, using CKAN's rich JSON API and via custom spreadsheet importers (CKAN, n.d.).

CKAN has a customisable data **harvesting models** which provide the mechanism for importing datasets from users' existing repositories into CKAN's facility. These models, already being used to fetch data from data.gov include: Geospatial CSW Servers, existing web catalogues, simple HTML index pages or Web Accessible Folders, and ArcGIS, Geoportal Servers as well as Z39.50 databases. Other features of the platform available for users are publisher tools, which includes:

**Admin dashboard** for members and data management;

**Workflow** system which separates public from private datasets for controlling visibility of who sees what on the system;

**Geospatial** features that provide data preview, search and discovery;

**Community services** features that offer users the ability to communicate and collaborate with each other on data. These features include *comments* extension, *share* and *RSS feeds* as well as *follow* and *to do* extensions

**Visualisation tools** – data visualisation by table and charting, mapping and image, etc;

**Themable features** – to create a customisable settings according to users preferences

**API** – for the purpose of querying and access to dataset information, CKAN provides a RESTful JASON API which gives access to a number of services such as full querying/searching, data information and download, dataset listing and etc.

**Storage, History, Extension and Federate** are other features of importance which enable users to store data, metadata and links to offsite repositories; provide histories of edits and versions of dataset metadata using Version Domain Model (VDM); up to 60 extension options for user to use for their data and provide the opportunity for users to create federate network of CKAN nodes involving other CKAN facilities.

In terms of popularity and user base, CKAN which is aimed at the government, is being used, so far, by 50 out of 330 data catalogues worldwide (Iemma et al., 2014). CKAN platform has the capability to provide rich service to users based on the possession of the following features (CKAN, n.d.):

- Complete catalogue system with easy to use web interface and a powerful API
- Strong integration with third-party CMS's like Drupal and WordPress
- Data visualization and analytics
- Workflow support lets departments or groups manage their own data publishing
- Fine-grained access control
- Integrated data storage and full data API
- Federated structure: easily set up new instances with common search

Table 12: Summary of CKAN features

1	CKAN
<b>Features</b>	<b><u>Website literature review</u></b>
<b>Installed instances</b>	CKAN has 116 well-known instances on the web and several other instances.
<b>Metadata, Data and File Format Standards and Schemas</b>	Support for any file format including tabulated geospatial data formats e.g. CSV, XLS, ArcGIS, Inspire and GeoJSON. API – for querying and accessing datasets; uses RESTful JASON API for access to services. Any file format can be uploaded. Other files supported. Store metadata of dataset and supports DCAT.
<b>Flexible search facility for datasets</b>	APIs for searching, querying & accessing datasets; RESTful JASON API for querying/searching, data, information & download, dataset listing etc. Searching by keyword or filter by tags; drill-down search via facets. Uses metadata fields to create the index.
<b>Social Media, Collaboration and Social Sharing tools</b>	CKAN has many social media tools: Facebook, Google+, twitter, etc. for user to communicate & collaborate, to comments, share, RSS feeds, follow, & To-do extensions.
<b>Dataset Publishing workshop</b>	Streamline publishing by importing datasets via a web interface which allows update & refine datasets in a distributed authorisation model. Workflow for groups to customised data publishing, separation of public / private datasets. Fine-grained access control & addition of metadata to workflow for data upload. File upload via web UI using API or linked to an existing file on the web; dataset upload by adding metadata on workflow
<b>Harvesting, Federation and Cataloguing</b>	Customisable data harvesting fetches data from sources: Geospatial CSW Servers, existing web catalogues, simple HTML index pages or Web Accessible Folders, ArcGIS, Geoportal Servers & Z39.50 databases. Complete cataloguing, easy interface & API. Strong integration and federate capability. Supports federation & has easy to use cataloguing & search service.
<b>Extensibility mechanisms</b>	Has up to 60 extension options for users; it's Open source, very extensible platform, has JSON API. Allows links to external datasets
<b>Data Analysis tools</b>	Administrative dashboard for members and data management but no special tools for data analysis
<b>Visualisation tools</b>	Basic visualization for tabular data and also by charting, mapping and imagery, etc.

Personalisation tools	Themable features – personalised settings for users' preferences.
<b>Customisation tools</b>	CKAN has customisable data harvesting models which support importing datasets from users' repositories. Customization using extension
<b>Dataset licensing service</b>	Licensing information can be added during the upload process
<b>Accessibility</b>	No special features related to accessibility
<b>Technical Environment</b>	Build using python programming languages with pylon web framework.
<b>Others</b>	Supports all file format; ease of use, detailed documentation, vast user base; self-hosted or accessed as SaaS

### 3.2.2 DKAN

DKAN is an open data platform that is based on Drupal and maintained by NuCivic. It is a tool which provides a full suite of cataloguing, publishing and visualization features that allow governments, non-profit organisations and universities to easily publish data to the public. With supports and inputs from OKF, DKAN is designed after CKAN 2.0 functionality, standards and API configuration; and does, in fact, reuses CKAN components wherever possible (Hoppin, Byrnes, & Couch, 2013; World Bank, 2014). There is however, a point of difference between CKAN and DKAN in that, DKAN is a distribution (pre-configuration) of Drupal and as such is also a complete CMS offering comprehensive tools to manage content, documents, and community, in addition to datasets which is presumably impossible in CKAN (World Bank, 2014).

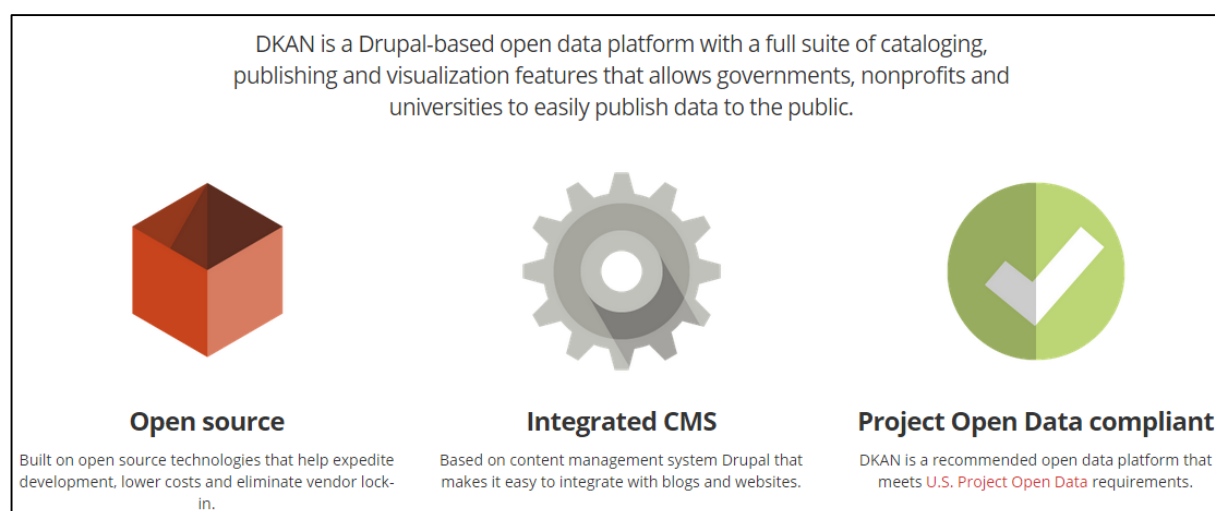


Figure 12: DKAN web Interface. Source: <http://nucivic.com/dkan/>

**Features:** Some of DKAN's features takes advantage of Drupal's well-developed Flexible and Theming system for **Customisation**, and others are derived from CKAN since both platform are intended to be compatible. More key features of DKAN include – ease of data publishing in key machine-readable formats (including JSON, XML, RDF), **share datasets** through an API as well as manage the **upload** of large datasets. DKAN has 18,489 extension modules to customise functionalities and has the capability to **manage dataset** easily (Hoppin et al., 2013). DKAN is built so that it can support **social media** tools such as blog, comment module from Drupal, **Disqus** comments for **collaboration and interaction** purposes amount users. Data **Workflow**, Editable Universal Unique Identifier (UUID) Field, **Google Analytics** Reports, Publishing of maps with CartoDB and DKAN, Visualization Entity and **Datastore API** are other examples of features of the platform. The search facility is clearly presented; permits filtering by metadata to returns results with titles and descriptions (World Bank, 2014)

Features	
Data publishers	Data users
✓ Manage documents, data and content within a single platform	✓ Explore, search, add, describe, tag, group datasets via web front-end or API
✓ Online community features	✓ Collaborate with user profiles, groups, dashboard, social network integration, comments
✓ Publish data through a guided process or import via API/harvesting from other catalogs	✓ Use metadata and data APIs, data previews and visualizations
✓ Customize your own metadata fields, themes and branding	✓ Extend and leverage the full universe of more than 18,000 freely available Drupal modules
✓ Store data within DKAN or on external (e.g. departmental) sites	
✓ Manage access control, version history with rollback, RDF support, user analytics	
✓ Enterprise-quality commercial support and FISMA-certified cloud hosting options available	

Figure 13: DKAN features in brief. Source: <http://nucivic.com/dkan/>

A summary of the features of DKAN is offered by World Bank (2014) is presented below.

DKAN imports and interprets datasets in CSV, XLS, XLSX and PDF file formats and also text files in a machine-readable format. As a current shortcoming, DKAN render data to users in the same format as it obtain datasets from publisher without any data transformation.

DKAN has a clear and thoroughly documented online but complex API which allow data resources to be downloaded via the API with output available as JSON or XML.

DKAN harvests existing data resources and is able to regularly update streaming data, via the API. However, there is currently no user-interface for setting up automated harvesting tasks.

Federating is made possible through DKAN's interconnections with Drupal

As part of standardisation policy, DKAN is aligned with best practice in the open data industry, yet offers no support for metadata and data structure.

DKAN's visualisation tool is described as 'public facing visualisation library', limited in support and does not permit functionalities to *save* or *share* of specific visualisation materials but a new set of tools developed recently supports embedding and saving charts, including geospatial data, as part of data-driven initiative

Integration toolkits were developed to facilitate integration with third-party data visualisation web services such as CartoDB.

Table 13: Summary of DKAN features

2	DKAN
<u>Features</u>	<u>Website literature review</u>
<b>Installed instances</b>	No good estimate available
<b>Metadata, Data and File Format Standards and Schemas</b>	Designed after CKAN 2.0 functionalities, standards and API configuration with supports for standard file formats including DCAT, INSPIRE, CSV, JSON, XML, & RDF. Upload Files in any format

<b>Flexible search facility for datasets</b>	<p>The search facility is clearly presented; permits filtering by metadata to returns results with title and description</p> <p>DKAN provides search UI and allows filtering on metadata fields</p>
<b>Social Media, Collaboration and Social Sharing tools</b>	DKAN is a distribution (pre-configuration) of Drupal with a complete CMS. Offers tools to manage content, documents, & community. Sharing via API; Supports social media, e.g. blog, comment, Drupal, Disqus comments, collaboration and interaction.
<b>Dataset Publishing workshop</b>	Full suite of cataloguing, publishing features. Ease of data publishing in key machine readable formats (e.g. JSON, XML, & RDF). Data Workflow, Editable Universal Unique Identifier (UUID) Field. Upload data using DKAN web front-end and provides web-based workflow attaching metadata to dataset
<b>Harvesting, Federation and Cataloguing</b>	DKAN has complete suite of tools for cataloguing and harvesting dataset.
<b>Extensibility mechanisms</b>	DKAN has 18,489 extension modules to support customizable functionalities with easy dataset management. Open source project; based on popular Drupal CMS which can be easily extended
<b>Data Analysis tools</b>	No special data analysis functions or tools, support Google Analytics, Publishing maps with CartoDB.
<b>Visualisation tools</b>	Visualization features exist for users to display their dataset in reports but limited support
<b>Personalisation tools</b>	Theming is available for personalisation.
<b>Customisation tools</b>	CKAN can be customized with theming, JSON API and Drupal extension
<b>Dataset licensing service</b>	Licensing information can be added during the upload process
<b>Accessibility</b>	No build-in support for accessibility, but accessibility features could be added using Drupal accessibility modules
<b>Technical Environment</b>	Uses PHP based CMS Drupal
<b>Others</b>	Data store API, easy to use extendable platform, Similar to CKAN; Provides complete CMS functionality

### 3.2.3 SOCRATA

Socrata is an open data platform providing Software as a Service (SaaS) with a “range of extension for dashboards, live reports and the ability to manipulate and update existing data live in the portal” (World Bank, 2014). It offers citizens a direct way to access and use public information by by-passing the formal process of requesting information from the government (Russell, Kristin, n.d.). This means citizens are granted access and opportunity to review, compare, visualize, and analyse data as well as share their discoveries in real time. The vision is to transform how citizens and government interact and to enable citizens make their charts, graph and maps about what interest them most.

**Features:** Under the term “Streamline Data Publishing and Management”, Socrata explains the provision of a scalable cloud platform which helps users create a sustainable open data program. As a data publishing platform optimised for business users, Socrata is an easy-to-use set of tools that require no *special skills* to publish data because it permits *automatic* publishing with *API-based* client libraries in a ‘push mode’ (Iemma et al., 2014), and allow **configuration** of publishing and **workflow** organisation. It offers **Flexible metadata management** by means of which users can implement a defined standard of vocabulary for their organisation, and create and maintain an enterprise data inventory via APIs or data.json file type. Network creation with regional hubs, cities and counties is simplified into a one-click process that seamlessly allows users to **Federate** with other Socrata customers. Socrata also offers the users the possibility to **measure their performances** on the platform in real-time consumption and distribution of their data and API. Publishers can track which data is most consumed and how. **Real-time reporting** allows monitoring of poignant (‘hot’) datasets, trending keywords and API usage tracking. Another important feature of Socrata is the freedom of **portal administration** it grants to users which allows them access to tools to secure their sites and manage resources. This privilege also enables granular control over every dataset with publisher’s option to keep private or share with the public; manage their sites with end-to-end datasets, users analytics, licensing and attribution.



Figure 14: Socrata web interface. Source: <http://www.socrata.com/>

Under the term “*Modern, Consumer-friendly Experience for Citizens*”, Socrata provides tools that ensure users can easily discover, explore, visualise and share government data to make it more impactful. **Searching** data on the portal is made possible by a robust weighted search index that combines metadata as well as row-, column- and cell-level to maximise relevance in searches. A special advantage provided by Socrata to users is the fact that non-technical users can easily interact with the data online and make a sense of it using *capabilities such sorting, auto-filtering to create a personalised view in addition to mapping and charting capabilities*. On **social aspect**, Socrata provides a platform that supports **civic engagement and participation**, bringing social experience around data in the form of comments, rating, and even more importantly, a feedback loop that drives further adoption and data consumption culture across social networks. The platform also **support co-creation and crowd-sourcing** functionalities by helping specialised users such as journalists and bloggers to contextualise government data and use it to share their stories. In order to support contextualisation,



more tools to embed datasets, to visualise and propagate data on blogs and media sites are provided on the data portal.

Under the third service category, “Developers, Apps and the Data Economy”, Socrata connects open data initiatives to the broader **app ecosystem** supported by open data API and other developer resources. Thus there exists a robust RESTful open API that reduces implementation costs, reduces the barrier on developer community engagement and hence increases the probability of developers’ further investment. A useful summary of Socrata features can be made out of the report produced by World Bank (2014).

Socrata can deal with dataset in these file formats: CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML, OData, Shapefile, KMZ, and KML; and as a part of standardisation feature, licencing is on each dataset (individual dataset licencing) with clear labelling. However, there is no clear attribution feature. Unfortunately, Socrata does not provide standardised metadata for dataset structure or format, nevertheless, customisation of metadata fields is possible for publishers.

The platform maintains a fast searching on a user-friendly interface and permits data filtering by view types, categories and topics. Additional offer through search is dataset description, abbreviated view of the first three matching rows of data and rapid assessment of results which can be filtered and faceted via the web interface as well as the API.

Socrata produces a wide range of outputs or endpoints via their API, including REST JSON, CSV and RDF-XML. Socrata API allows the development of, and the availability of dashboard supports the management of automated processes for uploading fast-changing datasets or importing existing resources.

In terms of extensibility, Socrata supports The White House’s /data.JSON URL extension specification. Extra extension developed enables importation of metadata from alternative open data portals, such as CKAN. The adoption of a mixed licensing approach permits systems scalability and extending Socrata is straightforward due to a wide range of available API.

Federating feature is enhance because all Socrata sites run on a single server; and this also make sharing resources/datasets between Socrata sites is a straightforward process.

A wide range of visualisations tools is available of Socrata platform for creating charts of various types such as Area, Bar, Column, Donut, Line, Pie, Time Line, Tree Map and Heat Map. There is Geospatial support and visualisations including location data, or GIS files such as Esri shapefiles, KML/KMZ files, using either Google Maps, Bing Maps or ESRI.

The design of Socrata to support easy deployment and management, unfortunately, also limits the degree of websites customisation.

Table 14: Summary of Socrata features

3	SOCRATA
<u>Features</u>	<u>Website literature review</u>
<b>Installed instances</b>	No estimate available
<b>Metadata, Data and File Format Standards and Schemas</b>	There exists a robust RESTful open API, open data API that supports App ecosystem. File formats include JSON, CSV, XLS, and XML. Uses DCAT also Support geospatial formats
<b>Flexible search facility for datasets</b>	Searching data on the portal is by a robust search index & allows filtering.
<b>Social Media, Collaboration and Social Sharing tools</b>	Supports civic engagement, participation & social experience: comments, rating, & feedback these help adoption & data consumption across social networks. Connects OD initiatives to the broader app ecosystem supported by OD API & other developer resources.

<b>Dataset Publishing workshop</b>	Automatic publishing with API-based client libraries in a 'push mode'. Configuration of publishing and workflow. Granular control of dataset with publisher's option to keep private or share. A web-based data upload using API.
<b>Harvesting, Federation and Cataloguing</b>	Network creation with regional hubs, cities and counties is easy & allows users to Federate with other Socrata customers. API allows powerful harvesting features; sharing datasets across multiple Socrata portals; provides catalogue service.
<b>Extensibility mechanisms</b>	Scalable cloud platform. Support co-creation & crowd-sourcing, helps specialised users e.g. journalists and bloggers to contextualise government data and use it to share their stories. API and libraries which allow developers to easily extend capabilities.
<b>Data Analysis tools</b>	Capabilities for mapping and charting are available. Some basic Business Intelligence services. Library for working with statistical package R
<b>Visualisation tools</b>	Provides powerful tools for visualizing machine readable data in various formats. Visualizing geospatial data
<b>Personalisation tools</b>	Personalised sorting, auto-filtering to create a preferred view. The portal administration allow personalization of portals
<b>Customisation tools</b>	Freedom of portal administration & metadata management for users to implement their standards; create and maintain an enterprise data inventory via APIs or data.json file. Verity of tools for customization.
<b>Dataset licensing service</b>	Licensing information can be added to dataset as metadata
<b>Accessibility</b>	Uses common best practices to allow accessibility
<b>Technical Environment</b>	Most of Socrata components are written using scala
<b>Others</b>	Measure user's performance on platform in real-time consumption & distribution of their data & API. Track data consumption. Tool to support contextualisation and embed datasets.

### 3.2.4 PUBLISHMYDATA

PublishMyData was developed by Swirrl to use the standard of Linked Data from W3C to publish data and also Linked Data on the platform in a model that brings remarkable benefits to users. PublishMyData as a platform is focused on technical users of statistical data, and offer RDF Data Cube support which offers a comprehensive data publication and management service to users communities (World Bank, 2014). Some of these benefits – which are essentially the features of the platform, include the following:

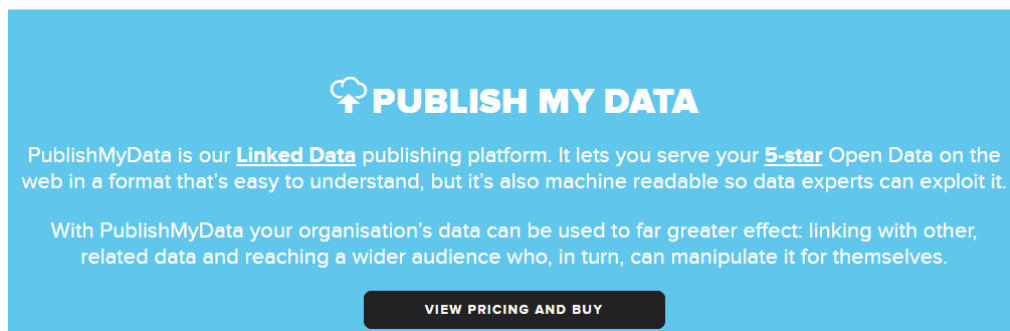


Figure 15: PublishMyData web interface. Source: <http://www.swirrl.com/publishmydata>

**Features: Cloud Operation** – the platform operates SaaS meaning it takes charge of the tasks of maintaining, supporting and improving the system while granting **administrative control** of data publishing and **customisation** of platform (including development of branded data site) to the users. The PublishMyData v2.2 is a powerful software which provides developers with a suit of features to ease the task of developing with Linked Data by making a range of Linked Data API available to developers, file formats such as RESTful JSON, Turtle, RDF/XML, interactive documentation tools and SPARQL and other query tools. The PublishMyData service offers **Browsable Data Website** feature, that is, it is **flexible** with nothing to install as it runs totally on the cloud; and also **compatible** with recent versions of all major browsers. The system is reliable with and fast and according to the platform claims on the service feature webpage, a Service Level Agreement (SLA) targets a 100% system availability and entitle to claim refund by users if availability drops below 99.5% (Swirrl, n.d.). There notable open data platform users in the **user community** of PublishMyData and these include – The Department for Communities and Local Government (UK), Hampshire County Council, Aberdeen County Council, The Department for International Development, The Scottish Government and the Greater Manchester Data Synchronisation Programme. Again, the study by World Bank (2014) has an interesting summary of features (also in ) for this platform:

Standard machine-readable formats such as CSV and XLS/X (Excel) are supported but there is currently limited recognition of geospatial datatypes and similar ones. Data licensing is by individual dataset and attribution for every dataset is implemented.

Data publication is only by simple listing because PublishMyData does not have a user-interface for searching data although there are possibilities for data to be traversed and searched via the API.

Static URIs are used to present all data and endpoints and add the advantage of making sharing and referencing data straightforward.

Dashboard activities are supported by SaaS capabilities while API supports integration with various visualisation systems as the platform does not maintain native visualisation system.

Federating from non-RDF sites is limited whereas API permits importing metadata from other RDF-compatible services; and it is the cleanest implementation of RDF for open data currently in the market whereby it adheres closely to W3C standards.

Full customisation is possible with the community edition of Swirrl's PublishMyData platform at GitHub ([http://github.com/swirrl/publish\\_my\\_data](http://github.com/swirrl/publish_my_data)) while the online platform version (SaaS) can also be customised or integrated into other systems.

*Table 15: Summary of PublishMyData features.*

4 PublishMyData	
<u>Features</u>	<u>Website literature review</u>
Installed instances	6 well-known instances

<b>Metadata, Data and File Format Standards and Schemas</b>	<p>Uses standard of Linked Data with a range of Linked Data APIs for developers e.g. file formats such as RESTful JASON, Turtle, RDF/XML.</p> <p>1) Uses RDF files formats as input 2) Supports SPARQL querying 3) Metadata about the datasets is valuable in DCAT.</p>
<b>Flexible search facility for datasets</b>	Provides SPARQL & other query tools for searching functionality but limited keyword search on catalogue data
<b>Social Media, Collaboration and Social Sharing tools</b>	<p>Interactive documentation tools that support collaboration and sharing</p> <p>Provides no special features for sharing and collaboration</p>
<b>Dataset Publishing workshop</b>	Use the standard of Linked Data from W3C to publish data with RDF. Converts file from CSV to RDF
<b>Harvesting, Federation and Cataloguing</b>	Provides datasets catalogue for users
<b>Extensibility mechanisms</b>	<p>Allows development of branded data site by users. PublishMyData v2.2 provides developers with tools for developing with Linked Data. Flexible browsable data website compatible with all major browsers.</p> <p>Community edition is available as Open Source project</p>
<b>Data Analysis tools</b>	NA
<b>Visualisation tools</b>	NA
<b>Personalisation tools</b>	Grant administrative control of data publishing and customisation of platform
<b>Customisation tools</b>	Granting administrative control of data publishing and customisation of platform but limited
<b>Dataset licensing service</b>	Linking information can be added as metadata
<b>Accessibility</b>	Simple & easy to user interface and simple intuitive navigation on linked data.
<b>Technical Environment</b>	Developed using Ruby on rails
<b>Others</b>	Browsable Data Website, flexible, nothing to install – cloud-based & compatible with recent browsers. Easy navigation on Linked Data

### 3.2.5 INFORMATION WORKBENCH

Information Workbench was built by **fluidOps** as a part of the company to provide a semantic integration platform that offers innovative cloud management tools and links it with best-in-class data centre technologies (Walther, n.d.). The Information Workbench as a Self-Service Platform for

Linked Data Applications, a platform which automatically analyses and uses any data regardless of source and format. It helps clients to find quick solutions to their complex questions, achieve tangible results, identify new opportunities quickly and utilise their competitive advantages.



Figure 16: Information Workbench web interface. Source: <http://www.fluidops.com/en/>

**Features:** **Transparency** is one of the benefit clearly mentioned accruing to the users of the platform because Information Workbench (IW) creates **transparency** in the users' datasets by removing the dead weight from your data, besides enabling users to access, use analyse and visualise and combine data in **flexible** manner. IWB manipulates data systems, workflows and processes as well as the underlying IT and infrastructure to:

- **integrate** datasets using sematic data model and API support, and **link** organisations together,
- process applications and information from **social networks** or other web sources
- develop and deploy custom apps
- support flexible data-driven user interface to unleash authoring, collaboration, visualisation and self-service through the numerous predefined widgets based on the powerful API interface.
- support developers' community, by providing a **comprehensive SDK for building apps** to support clients' individual scenarios and requirements. For the summary of the features, see table below.

Table 16: Summary Information Workbench features

5 Information Workbench	
<u>Features</u>	<u>Website literature review</u>
Installed instances	NA
Metadata, Data and File Format Standards and Schemas	1) Uses RDF format storing data 2) Supports SPARQL queries
Flexible search facility for datasets	Provides no user interface or API for searching
Social Media, Collaboration and Social Sharing tools	Process applications and information from social networks or other web sources. Uses wiki style user interface for collaboration
Dataset Publishing workshop	Data manipulation, workflows & processes to support data integration using semantic data model and API. Has connector to various data formats: CSV, XLS & database connections etc. that allows conversion to RDF.
Harvesting, Federation and Cataloguing	No specialized support for harvesting and federation



<b>Extensibility mechanisms</b>	Able to integrate dataset & link organisation together. Supports developers' community, via SDK for building apps. Allows extension and connectors
<b>Data Analysis tools</b>	Enables users to access, use, analyse and visualise & combine data in flexible manner. Supports R statistical package
<b>Visualisation tools</b>	Visualization of Twitter followers. Provides variety of visualizations widgets
<b>Personalisation tools</b>	Supports flexible data-driven user interface. Self-service through the numerous predefined widgets using powerful API interface. Wiki style interface allows users to organize content according to user preferences
<b>Customisation tools</b>	Develop and deploy custom apps. User interface is easy to customize, allow custom extension
<b>Dataset licensing service</b>	NA
<b>Accessibility</b>	NA
<b>Technical Environment</b>	Written using java and common web technologies
<b>Others</b>	The Information Workbench as a Self-Service Platform for Linked Data Applications. Creates transparency in the users' datasets by removing the dead weight from your data.

### 3.2.6 ENIGMA

Enigma was founded in 2012 and is based in New York as a platform that brings together thousands of various public data sources. The platform assembles rich data resources also include linked data to enables users make a better sense of the huge data by allowing them view and analyse data under various data variables, combining and viewing datasets to provide new insights and analysis (Programmableweb, n.d.).



Figure 17: Enigma Web interface. Source: <http://enigma.io/>

Enigma spreads out its service coverage across three areas:

- Discover Public Data – a repository of data collected from governments, universities, companies, and organizations to provide new insights into economies, companies, places and individuals.
- Build Apps and Services – provision of API tools for developers to power applications and data-rich services for maintaining real-time applications with direct access to billions of records, and
- Enterprise Solutions – providing companies with the opportunity to leverage Enigma’s expertise for their information awareness and decision-making supports.

**Features:** Enigma API contains several tools that enable developers to access and integrate the functionalities of the platform with other applications and to create new applications altogether. For example, Enigma Public Data API provides a **Direct Plug-in into Enigma** infrastructure through RESTful APIs to access the full range of Enigma datasets and analytics, and supports **application development** as well as **augment data understandability** by placing users’ data in context with relevant public datasets. In addition to the above services, Enigma offers enterprise-class performance with reliability and scalability of services including Data as a Service (DaaS) and analysis of massive datasets on demand. Other services available to users are – Entity resolution, data security, geocoding, time series, join analysis, **augmented research tools** (identification and connections in unstructured text) and data cleaning. Summary of features of Enigma platform is presented in the table below **Table 12**.

Table 17: Summary of Enigma features

6	Enigma
<u>Features</u>	<u>Website literature review</u>
Installed instances	One instance
Metadata, Data and File Format Standards and Schemas	Enigma Public Data API provides a Direct Plug-in into Enigma infrastructure through RESTful APIs to access the full range of Enigma datasets and analytics
Flexible search facility for datasets	Provides augmented search tools (identification and connections in unstructured text). Powerful search user interface and API; can search for data at record level
Social Media, Collaboration and Social Sharing tools	NA
Dataset Publishing workshop	Discover Public Data – a repository of data collected from governments, universities, companies, and organizations to provide new insights into economies, companies, places and individuals.  NA
Harvesting, Federation and Cataloguing	NA
Extensibility mechanisms	Build Apps and Services using the provided API tools for developers.
Data Analysis tools	Enables users to view and analyse data under various data variables, combining & viewing datasets e.g. Time series and Join analysis.
Visualisation tools	NA
Personalisation tools	NA
Customisation tools	NA
Dataset licensing service	NA
Accessibility	NA
Technical Environment	NA
Others	Enigma offers enterprise-class performance with reliability and scalability of services including Data as a Service (DaaS).

### 3.2.6 JUNAR

Junar was founded as a cloud-based open data platform operating as SaaS to help organisations open up their data and facilitating end-to-end data projects for businesses, governments, NGOs and academic institutions. Although Junar provides variety of open data publication services (with its proprietary SaaS services focused on ease of deployment and providing visual tools with plenty of hooks for downloading and developing custom applications), unfortunately, it has no support for structural metadata or hypercubes required by NSOs (World Bank, 2014). For businesses, Junar can transform data into interactive resources for internal and public uses; give opportunity for developers to access data from their application through Junar API, etc. For the government open data initiatives, Junar simplifies the data publishing process and transforms data into interactive resources for citizens to use, share and distribute. It offers the government the ability to attain the open data legislative requirements, to deploys and maintains open data programmes. Junar offers the NGOs the opportunity to view government transparency initiatives, adhere to their transparency mandates, develop and maintain their open data initiatives. In general, Junar services are to collect, enhance, publish, share and analyse data for the data hungry society of today and to support open data enhanced economy.

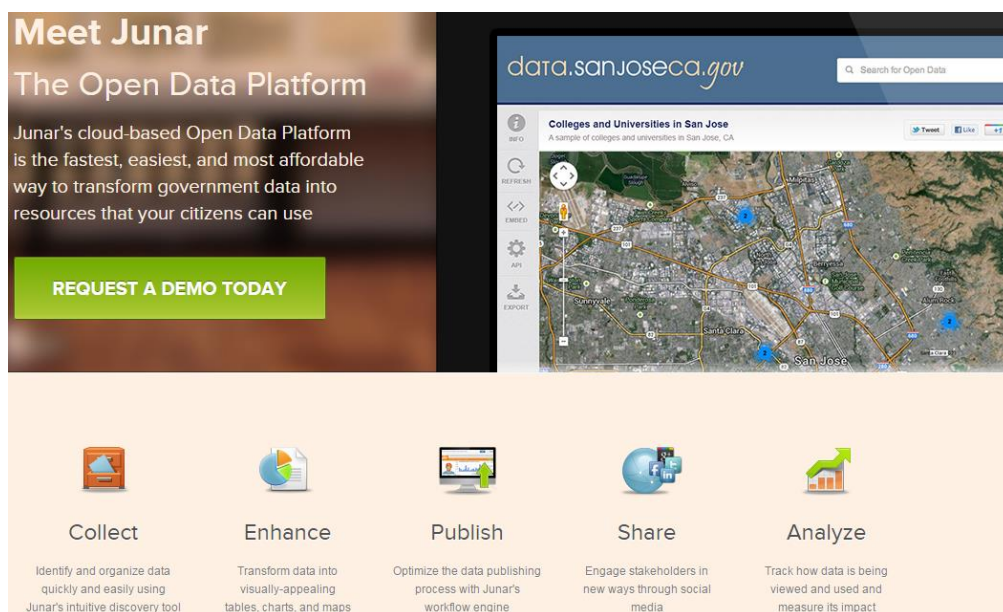


Figure 18: Junar web interface showing features. Source: <http://www.junar.com/>

**Features:** Junar provides an easy-to-use platform to **collect data of any format** from any location without need for conversion; then categorises the data and adds metadata to improve searchability. **Data enhancement** tools turn data into visual charts and tables with many options for dashboards and can **integrate** the data directly into the user's site with the help of a **workflow** mechanism that optimises data publishing. After data publishing, users can follow and share the data resources by means of Junar API that promotes **social networking** systems such as Twitter and Facebook on the data among data users in Junar community. With Junar facilities, it is possible **to measure and report the impact** of users open data in the community or society. These measurements include how the

data is being viewed, used and specific dataset being consumed the most, track popularity and downloads, and present these reports in visualised tables, charts and dashboard formats with option to integrate with Google Analytics. Furthermore, measurement facilities also provide opportunity to analyse and report on **feedbacks** from data audience. The following feature summary is offered by World Bank (2014) for Junar platform:

*Metadata, API & machine-readability: datasets are described using RDF file format and presented in Dublin Core and DCAT. Further support are offered for a wide range of different machine-readable formats, including CSV, XLS and XLSX, JSON and SOAP/XML 2.0, as well as the KML, KMZ, GeoJSON, and Shapefile geospatial formats.*

**Licenses:** These are not clearly presented for individual datasets, however, the next software release, as envisaged, will include custom licences for datasets using the template provided by <http://project-open-data.github.io/license-examples/>

**Search:** This functionality is not visually represented on Junar site as it is not a priority (World Bank, 2014) even though the search function is available with limited focus on faceting data. Data may be structured with metadata but not exposed through the interface to the user to permit search results filtration in a more accessible way.

**Analysis and Visualisation tools:** available interactive API permits developers to experiment live on the database to view results as each dataset generates a unique URI. System administrators can, by use of available API creates dashboards and visualisations which also get unique URI.

**Harvesting & Federating:** Publishing Workflow permits some automated collection and management of data from different locations through the application of some uploader scripts to automatically upload a range of file-format such as CSV, XLS, XLSX, KML and KMZ. Junar maintains integration capabilities that enable data collection from REST/JSON or SOAP/XML web services which are linked directly to source databases for real-time, or near-real-time, data collection. Integrated REDATAM+SP software permits harvesting from HTML forms for direct data collection. A simple drawback appears to be caused by the fact that Junar sites run on same servers as SaaS and that data are not federated across the different platforms. However, an extension app produced by Junar for CKAN enables metadata to be read across both platforms.

**Documentation:** is another aspect of good standard put up by Junar for its users because some wiki pages, many of which are customised for particular clients, are available at <http://en.wiki.junar.com/index.php/>. To increase consultation, these pages need to be more visible to developers and the knowledge base at <http://support.junar.com> needs to be more regularly updated also to improve the support offered by the resources.

Junar support leading open data community standards, however, it does not currently support structural metadata even though it support data endpoints, including CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML, all of which are also available via the API.

**Integration and Extensibility:** Junar provides integration facility compatible with Google Docs and Dropbox but available literature indicates that the platform is a proprietary software and the range of

published public APIs are only about downloading data rather than extending functionality or uploading data.

**Visualisation tools:** Junar provide supports for the development of comprehensive data-driven visual dashboards a range of charts including geospatial plotting, and the ability to drag and drop functionality for visual graphics on integrated dashboard.

**Customisation tools:** Junar Interface can be customised by users and possibly add new functionalities.

Table 18: Summary of Junar features

7	JUNAR
<u>Features</u>	<u>Website literature review</u>
<b>Installed instances</b>	20 well-known instance
<b>Metadata, Data and File Format Standards and Schemas</b>	Supports data formats e.g. CSV, XLS, XLSX, KML and XML. Allow metadata to be attached to datasets
<b>Flexible search facility for datasets</b>	Junar categorises data & adds metadata to improve searchability. Limited search service
<b>Social Media, Collaboration and Social Sharing tools</b>	Support for internal/public interaction, share and distribute functionalities. Junar API promotes social networking systems; analyse & report on feedback. Allow sharing with popular social media networks
<b>Dataset Publishing workshop</b>	Simplified data publishing; easy-to-use platform to collect data of any format from any location no need for conversion. A workflow mechanism optimises data publishing via web interface.
<b>Harvesting, Federation and Cataloguing</b>	Give opportunity for developers to access data from their application through Junar API. Harvest data from REST & SOAP services and harvesting of HTML forms. Junar is offered as SaaS but doesn't support federation.
<b>Extensibility mechanisms</b>	Can integrate the data directly into the user's site through a workflow mechanism that optimises data publishing. Limited extensibility.
<b>Data Analysis tools</b>	Analyses of data are limited.
<b>Visualisation tools</b>	Data enhancement tools turn data into visual graphics with dashboards. Tracks popularity and downloads, and presents reports in visualised tables, charts & dashboard & integrate with google analytics.
<b>Personalisation tools</b>	Possible
<b>Customisation tools</b>	Possible
<b>Dataset licensing service</b>	NA
<b>Accessibility</b>	No special feature related to accessibility
<b>Technical Environment</b>	Written in Java and Python
<b>Others</b>	Measures and report the impact of users of OD in the community, e.g. how the data is being viewed, Tracks popularity and downloads, & presents these reports in dashboard formats.

### 3.2.7 OPENDATASOFT (ODS)



Figure 19: ODS web interface

OpenDataSoft (ODS) was founded in 2011 (Paris) by Jean-Marc Lazard, David Thoumas and Franck Carassus, and is currently being used by at least 40 customers selected from public administration, local and regional authorities, transportation and mobility, energy and environment, services, tourism and the media. ODS, in a nutshell, provides services for Data Collection, Exploration and API-supported functionalities that can be performed on datasets including data publishing, use and reuse, share and broadcast, enrich and monitor usage, analytics and security. ODS claims that it breakdowns data silos and secures aggregation of data with cross-referencing of heterogeneous data; leverages data analysis to produce visualised interactive maps, chart and pictures all through innovative and powerful API publishing, monitoring, web activities, mobile application development, etc. (OpenDataSoft, n.d.). Despite the above claims, the ODS trial version (ODS playground as at 2014) , permits for each data upload occasion, a capacity of just around 5 datasets with 100 000 records of files limited to CSV and Excel formats only and without any possibility for embedding visualizations on external web pages (Lindén & Stråle, 2014). It also has reduced number of processors available for data preparation and data extractors.

**Features:** ODS has a number of different APIs including – OAuth2 Support, Query Language and Geo Filtering yet other available APIs groups include – the OData API, Real Time Push API, Dataset Search API, Records Search API, Records Analysis API, etc. The API support ODS’s services include the following (World Bank, 2014) summarised in the table below.

**Data Collection services** – data file upload (with **file format:** CSV, XLS, XLSX, SHP, KML, GeoJSON, OSM, GTFS), remote web services support, custom **connectivity** and data **Federation**; Data Processing – Geocoding, text transformation, joins, numeric operation and indexing; Data Sharing – text search, multi-criteria text search, data cataloguing, linked data capabilities, data catalogue export (CSV, RSS and RDF formats) and data export (CSV, JSON, XLS, SHP and GeoJSON); OpenDataSoft supports DCAT, and INSPIRE for geospatial data. There is enough documentation for testing and working with the API,



which permits HTTP/HTTPS/BasicAuth and present data end point in JSON/P, CSV, RDF, as well as GeoJSON/P;

**Customisation tools** – Customisable GUI and Embeddable widgets, possibility for creating custom metadata templates;

**Analysis and visualisation tools** – The platform maintains open APIs that provide an interactive online dashboard, Geo data visualisation; Analytics and imagery;

**Content management** – hosting customer data, CMS; User Engagement – supporting forums, contact forms and reuse management; Hosting and admin – user management, user groups, cloud hosting, **workflow**, **analytics** and **integration** functionalities, Data search is allowable in Natural language including filtering by a wide range of metadata and data-types with API allowing faceting during search; Managing domain – security and monitoring, Google Analytics and activity log; Collaborative code view – GitHub workflow for teammate discussions, feedback, compare views, text entry formatting and syntax highlighted code and rendered data.

**Harvesting & Federating:** OpenDataSoft can import data from several types of domains and can set processes for removing personal data, performing calculations based on formulae and data collection can be via remote locations or web services. However, there is currently no federation activities currently but the API and metadata traversal means that this should be possible in future (World Bank, 2014).

*Table 19: Summary of ODS features*

8	OPEN DATA SOFT
---	----------------

<b>Features</b>	<b>Website literature review</b>
<b>Installed instances</b>	38 well-known instances
<b>Metadata, Data and File Format Standards and Schemas</b>	Many APIs including – OAuth2 Support, Odata API, Real Time Push API, Dataset Search API, Records Search API, Records Analysis API. File format: CSV, XLS, XLSX, SHP, KML, GeoJSON, OSM, GTFS & ShapeFile
<b>Flexible search facility for datasets</b>	Both Dataset Search API and Records Search API are provided for searching purposes. Text search, multi-criteria text search are possible on the platform.
<b>Social Media, Collaboration and Social Sharing tools</b>	Limited data sharing on popular social media. User Engagement: forums, contact forms & reuse management, Collaborative code view – GitHub workflow supports teammate discussions, feedback, compare views.
<b>Dataset Publishing workshop</b>	Hosting and admin – user management, user groups, cloud hosting, workflow, analytics and integration functionalities etc. Web based UI and workflow for publishing data
<b>Harvesting, Federation and Cataloguing</b>	Data cataloguing, linked data capabilities, data catalogue export (CSV, RSS and RDF formats) and data export (CSV, JSON, XLS, SHP and GeoJSON). Data can be collected from external sources via web services. No federation available but provides cataloguing features
<b>Extensibility mechanisms</b>	Limited extensibility
<b>Data Analysis tools</b>	Leverages data analysis to produce visualised interactive graphics but analysis at basic level
<b>Visualisation tools</b>	Powerful API publishing support data analysis & interactive visualisation in maps & chart & pictures. Analytics and imagery. Geo data visualisation.
<b>Personalisation tools</b>	Custom connectivity and data Federation; Data Processing – Geocoding, text transformation, joins, numeric operation and indexing. Allows UI customization
<b>Customisation tools</b>	Customisable GUI; Embeddable widgets though Limited customization
<b>Dataset licensing service</b>	Licensing information can be added to dataset
<b>Accessibility</b>	No special features related to accessibility
<b>Technical Environment</b>	NA
<b>Others</b>	Remote web services. Good documentation and instructional manuals within the website for users. Simple and easy to use, easy deployment, offered as SaaS

### 3.2.8 CALLIMACHUS

Callimachus open data platform is mainly used by Government, Healthcare, Pharmaceuticals, Publishing and Research Organisation to address their Linked Data needs such as storage, graph, integrated development environment, visualizations and web publishing. The platform is built to meet web standards in terms of:

- Storage: RDF Graphs
- Data processing: XSLT and Xproc
- Templating: RDFa
- Parametrized SPARQL Queries
- Content management and programs: XHTML5, CSS3, JavaScript



Figure 20: Callimachus web interface

## Features:

**Standards and Formats:** Callimachus is RESTful in design to meet website standards especially HTML version 5, CSS version 3 and JavaScript, and linked data standards such as the Resource Description Framework (RDF). For users, it makes data easy to create, view, and update and simplifies the *integration* of new data with existing data compared to using relational databases. The platform has open source documentation including – guides, videos, sample applications and tutorials and in particular, *Callimachus for Web Developers* is a document which is used to teach developers how to use RDFa as a means of annotating HTML tags with data through the addition of attributes and how to transform data using XML technologies is helpful, especially XSLT and the recent XML pipeline language, Xproc. Callimachus platform supports the previous and current versions of the major browsers in the market – Windows Explorer, Chrome, Firefox and Safari.

**Extension/Personalisation:** However, there are some limitations with the Safari 6 & 7, and Internet Explorer 9, 10 & 11. (Callimachus, n.d.). In Safari 6 & 7, Callimachus does not support Drag and drop of file, logging in with email address via digest access and does not retain the login state in all cases whereas in Explorer 9, 10 & 11, it does not support drag and drop of file, client-side validation and cannot upload files in .docbook format – thus, the .docbook documents must be created within Callimachus system.

**Customisation tools:** Callimachus provides environment for various user groups; for example, the *Admin User group* and the *Super User Group*. While the Admin group manages the resources, the super user group have special privileges to view and modify content in the Callimachus folder tree as well as the core of a running Callimachus instance. Under *file support* and *history*, Callimachus supports binary and text files (which are used to store Callimachus Archive Files) and Zip files; and it

maintains a chronological log for the changes made to resources on the platform respectively. An overview of the features of the two 3Round Stones platforms – Callimachus (community-based) and Callimachus Enterprise (commercial) are presented in while the summary of the features of Callimachus open data platform is presented in below.

Feature	 Callimachus	 Callimachus Enterprise
	Community-based	Commercial
Support		
Linked Data Publication	✓	✓
In-Browser Application Creation	✓	✓
Enterprise Management	✗	✓
Cloud Deployments	✗	✓
User Profiles, Social Sharing	✗	✓
Document and App Management	✗	✓
Open Annotation Support	✗	✓
External Datasources	✗	✓
Shared Deployments	✗	✓
Realms (Virtual Hosts)	✗	✓

Figure 21: Features of 3RoundStones' Platforms. Source: <http://3roundstones.com/products/>

Table 20: Summary of Callimachus features

9	CALLIMACHUS
---	-------------

<b>Features</b>	<b>Website literature review</b>
<b>Installed instances</b>	Not available
<b>Metadata, Data and File Format Standards and Schemas</b>	Web standards for Storage: RDF Graphs, Data processing: XSLT & Xproc, Templating, RDFa, SPARQL Queries, Content management: XHTML5, CSS3, JavaScript, RESTful. Support text files and Zip files.
<b>Flexible search facility for datasets</b>	NA
<b>Social Media, Collaboration and Social Sharing tools</b>	Provides wiki pages for collaboration and sharing
<b>Dataset Publishing workshop</b>	Uses templates format for data collection and wiki pages for publishing & workflow
<b>Harvesting, Federation and Cataloguing</b>	For users, it makes data easy to create, view, and update.  NA
<b>Extensibility mechanisms</b>	Simplifies integration of new data with existing data compared to using relational databases. Open source project
<b>Data Analysis tools</b>	NA
<b>Visualisation tools</b>	NA
<b>Personalisation tools</b>	Limited personalization
<b>Customisation tools</b>	Provides limited customization for various user groups; e.g. the Admin User group & the Super User Group.

<b>Dataset licensing service</b>	NA
<b>Accessibility</b>	No special features related to accessibility
<b>Technical Environment</b>	Java based application
<b>Others</b>	Has open source documentation: guides, videos, sample apps & tutorials particularly for Web Developers. Supports previous & current versions of major browsers. Maintains a log. Publish data as linked data.

### 3.2.9 DATATANK

DataTank provides a software platform with data handling and information tools mainly for local governments to deal with data verification, fraud investigations and streamlining problems. Through fraud detection solutions and the ability to create a holistic view of data across departments, users can increase revenue and make significant savings through efficiencies (DataTank, n.d.). DataTank

platform, based in the UK, uses the financial bureau data with latest technology combined with manual human investigation to produce valued, ISO-certified services for their customers.

**Features: SPD Profiler** – this is a Single Person Discount profiler for the identification of a Single Person Discount fraud through the validation of claims and identification of fraud. This service helps local authorities to save money by avoiding the conduct of annual cold canvas of SPD claimants. **Fraud Profiler** is a service whereby users apply the DataTank special software to manage their fraud investigations and **School Administration Checker** – is another service whereby various schools (primary and secondary) use DataTank software to check the validity of individual applications at time of submission. It is also useful for processing a batch of applications quickly in one go; and in both cases, the software substantially reduces the time and effort it normally take to verify if the applicants' parents or guardians are resident in the individual school's catchment area. DataTank helps its local authority customers *to view, analyse and interpret* their data and also to helps them to connect datasets from different departments and then across-tabulate and process the data in many ways that reveal *relationships, patterns and trends*. Lastly, **Connect Localism** is a service use in England and Wales to offer inter-connections between the council and the Council Tax Boards (CTB) in order to understand the impact of Council Tax Schemes (CTS). This service helps council authorities to understand and to adapt tax changes in order to create policies in line with Localism and Welfare Reform. Below contains a summary of the platform features.

Table 21: Summary of Datatank features

10	DataTank
<u>Features</u>	<u>Website literature review</u>
Installed instances	4 well-known instances
Metadata, Data and File Format Standards and Schemas	Support CSV,XML and JSON file formats
Flexible search facility for datasets	Limited filtering by dataset name

<b>Social Media, Collaboration and Social Sharing tools</b>	NA
<b>Dataset Publishing workshop</b>	Provides tools for ETL
<b>Harvesting, Federation and Cataloguing</b>	NA
<b>Extensibility mechanisms</b>	Open source project
<b>Data Analysis tools</b>	Creates a holistic view of data across departments to help its local authority customers' analyse and interpret their data. Reveal relationships, patterns and trends.  NA
<b>Visualisation tools</b>	Creates a holistic view of data across departments. Reveal relationships, patterns and trends.  NA
<b>Personalisation tools</b>	Limited personalization
<b>Customisation tools</b>	Limited customization
<b>Dataset licensing service</b>	NA
<b>Accessibility</b>	No special features related to accessibility
<b>Technical Environment</b>	PHP based application
<b>Others</b>	Information tools for governments to deal with data verification, fraud investigations and streamlining problems. Convert data in to REST API

### 3.2.10 SEMANTIC MEDIAWIKI

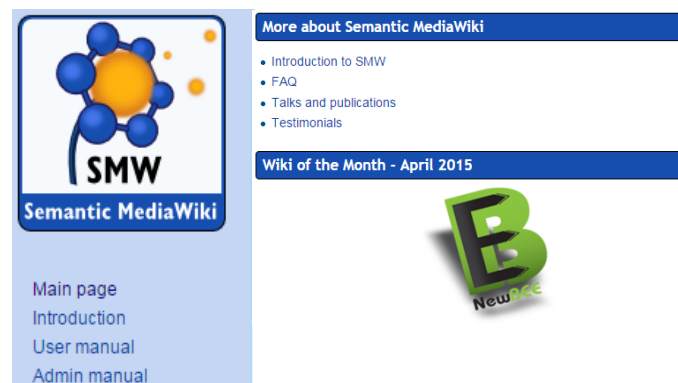


Figure 22: Semantic MediaWiki web interface

Semantic MediaWiki is extension to MediaWiki and can work as collaborative open data platform. Semantic MediaWiki adds semantic capabilities to MediaWiki platform. It allows semantic annotation to be added to MediaWiki content. Information in Semantic MediaWiki is presented in both human readable and machine-readable format. Semantic forms extension can be used to create forms that allow easy entry of structured data into MediaWiki. Data can also be imported from CSV, XML and JSON formats. Data in Semantic MediaWiki is stored in a triple-store and SPARQL query language is used to query the data stored in RDF. There is limited support for data visualization. Semantic MediaWiki provides text search over content stored in MediaWiki. Licensing information can be added to each page. Semantic MediaWiki provides full version control capabilities. MediaWiki platform is a highly customizable and extensible platform. The summary of the Semantic MediaWiki platform is presented in the table below.



Table 22: Summary of Semantic MediaWiki features

11	Semantic MediaWiki (SMW)
<u>Features</u>	<u>Website literature review</u>
Installed instances	NA
Metadata, Data and File Format Standards and Schemas	CSV, XML, JSON formats, RDF. Information in Semantic MediaWiki is presented in both human readable and machine-readable format. Support SPARQL queries
Flexible search facility for datasets	Provides text search over content stored in MediaWiki. SPARQL query language is used to query the data stored in RDF. Free text search over data with limited filtering
Social Media, Collaboration and Social Sharing tools	Semantic MediaWiki is extension to MediaWiki and can work as collaborative open data platform. Allows wiki style collaboration
Dataset Publishing workshop	All data created within SMW can easily be published via Semantic Web. Data can via WikiText or can be entered via web forms or imported using CSV or XML.
Harvesting, Federation and Cataloguing	1) No stated cataloguing services is provided 2) Limited federation and harvesting
Extensibility mechanisms	SMW allows other systems to use its data seamlessly. Based on MediaWiki, allow extensions.
Data Analysis tools	NA
Visualisation tools	Limited
Personalisation tools	Wiki style user interface allows users to organize content according to their preferences
Customisation tools	MediaWiki platform is highly customization and extensible platform. Allows customization
Others	1) Open source 2) Based on MediaWiki 3) Easy editing
Dataset licensing service	Licensing information can be added to each page
Accessibility	Easy to find relevant content
Technical Environment	Written in PHP

### 3.3 GENERIC ARCHITECTURE OF OPEN DATA PLATFORMS

To understand the requirements for Open Data Platforms Architecture we analysed a selection of existing Open Data Platforms. The selection was made by analysing the usage of the platforms as well as the analysis of the latest publications concerning Lined Open Data. Each platform was analysed based on publicly accessible documentation, such as publications, press releases and projects websites.

Based on the survey results discussed above, Open Data Platforms can be represented as three-layered system. This pattern was observed across all of reviewed Open Data Platforms. This layered architecture is shown in and described below.

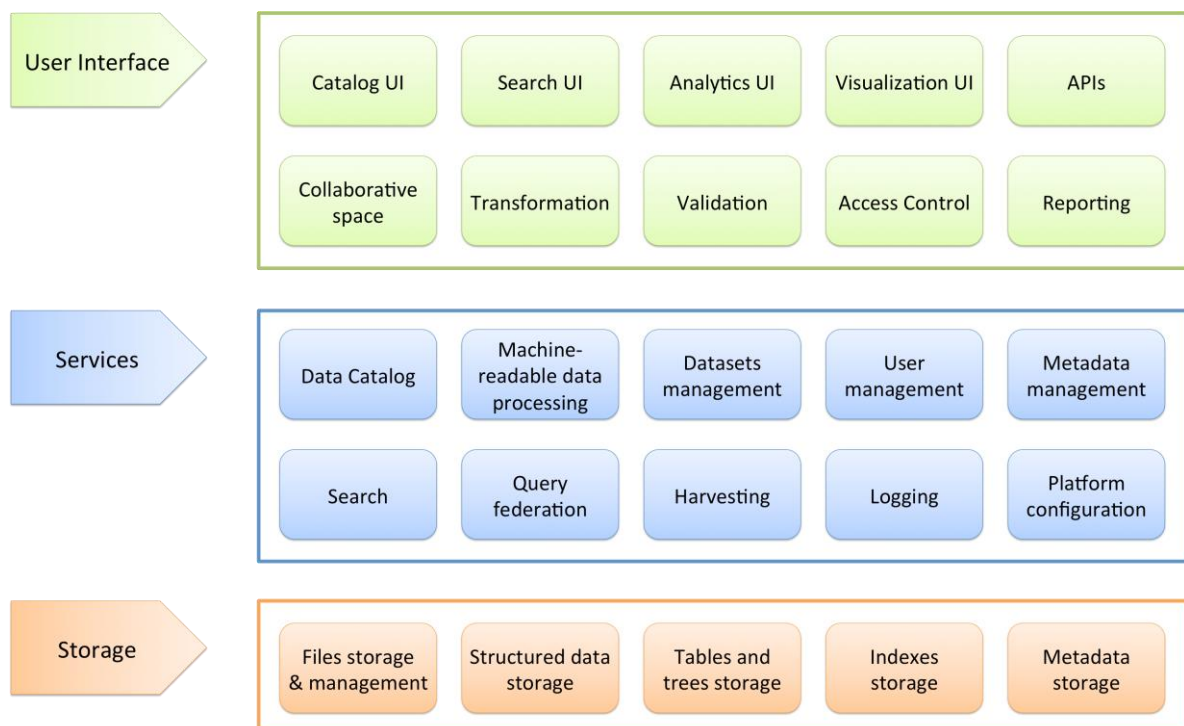


Figure 23: Architecture of Open Data Platforms

The layer at the bottom handles data, which can be implemented with file store; index and data store such relational, no-SQL or RDF store etc. The middle layer can viewed as layer of high level of services provided on top the data which may include catalogue service, federation and harvesting, service for handling machine readable data, search and service for handing plugin interface. The top layer provides services to users and applications, which includes user interface for the catalogue, access control, personalization and customization, analysis and visualization, collaboration tools, user interface for search and API for external applications. Details of each layer are as follows:

**Storage layer:** it is concerned with persistence of data and information and provides all the tools for data storing and efficient retrieving of Open Data. This layer is responsible for storing the files,

structured data, tables and trees as well as the indexes and the metadata. Data can be stored directly in file system storage or in structured data store.

Component name	Role / description
Files storage & management	The main role of this component is persistent storage of raw data files. It may be implemented as a local or remote datastore. Files referenced by URL that need to be downloaded for analysis and consumption are considered as remote file storage.
Structured data storage	Structured datastore allows efficient retrieval and querying over data. File and records stored in platform can be indexed for efficient retrieval for searching.
Tables and trees storage	Indexed and structured storage of tables and trees. Most common option is a SQL database.
Indexes storage	Indexes storage, such as various schema elements, search indexes, vocabularies, concept schemes, etc.
Metadata storage	Persistent metadata storage. This component may include various kinds of metadata, including provenance information.

**Services layer:** it provides services on top of Storage layer that can be exploited by the User Interface layer. Data Catalog services are used to list the details of datasets and associated metadata stored in platform. Search service uses the index to search relevant content. Platform extensions services allow external applications to use the platform services. All these services have the corresponding features in the interface layer.

Component name	Role / description
Data Catalog	A user interface that allows browsing, exploration and querying of a collection of dataset metadata records.
Machine-readable data processing	Process and serves the available data in a format that can be understood and consumed by a computer.
Datasets management	Allow to create new datasets and manage them.
User management	Stores and manages the data about user profiles. Users may need the ability to register, edit their user profile, and view profiles of other users. This component also controls the access, permissions and group memberships which can be configured.
Metadata management	Adding and editing of metadata records, such as provenance information, modification date, license and so on. This component also includes i.e. quality assessment.
Search	This component is responsible for the retrieval of information according to the user queries and actions. It is directly connected with the data stores. It allows full-text search over stored data, at sufficiently fine granularity. This includes metadata search, and searches all the data, potentially returning any resource.
Query federation	Query federation allow data integration between multiple instances of the data portal.
Harvesting	Harvesting services allow integration of data from other portal instances and other, external data sources.
Logging	It provides logging services throughout the system. All components should log to a single system in a consistent way. This component should help to monitor the system status and issue appropriate alerts in the case of system failures.
Platform configuration	This component allows to modify the platform parameters (such as timeouts, limitations) and the available functionalities. It could be implemented through command lines and configuration files.

**User Interface layer:** User Interface layer – i.e. CKAN user interface - provides basic portal functions such as access to the data, search interface and personalization and customization features etc. Search feature allows users to quickly find information stored at the portal, while analysis and visualizations features allow users to explore, analyse and visualize various types of data, such as tabular and geospatial data. Various APIs allows external application to consume services offered by the platform.

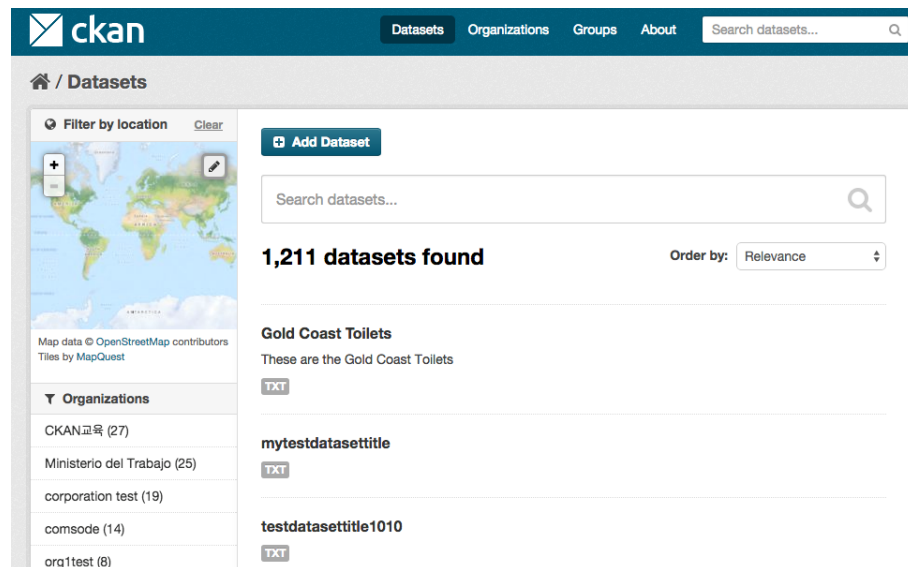


Figure 24: Screenshot of search and catalogue feature of CKAN

Component name	Role / description
Catalog UI	<p>A user interface that allows browsing, exploration and discovery. Basic functionalities are category browsing, dataset rating, data downloading or browsing by category, etc.</p> <p>It may provide additional information that facilitates discovery, e.g., featured datasets, additional labels, etc.</p>
Search UI	A user interface that allows querying the collection of dataset metadata records. Basic functionality is the text search.
Analytics UI	<p>Comprises any number of approaches to generating new insights from data: data mining, data fusion, spatial linking, statistical analysis, clustering, and so on.</p> <p>Usually a knowledge of these techniques is necessary for a user to derive value from this interface.</p>
Visualization UI	Allows to visualise the datasets (i.e. map view, chart, tabular view) as well as the results of the analysis for non-expert users. This interface helps with making data more accessible.
APIs	Allows external application to consume the services offered by the platform in easy programmatic access.
Collaborative space	Display of user profiles; user registration; permission management. Basic functionalities allow users to communicate (messaging system) and discuss about the datasets.
Transformation	Allows to export the data in additional formats. An example is XML to CSV conversion. It should support the data refinement and cleansing.
Validation	Syntactical validation of raw data files for selected raw data formats, e.g., XML parsing, CSV parsing. It should gracefully handle i.e. syntax errors.
Access Control	May be implemented as the Access Control Logic and restrict access to some private datasets and is used for securing access to the system.
Reporting	<p>It provides reporting services throughout the system.</p> <p>It may display usage statistics, data summary, queries summary, datasets quality summary and so on</p>

### 3.4 PLATFORMS EXTENSIBILITY

This highlights our findings on the extendibility features of the different platforms provided to determine the feasibility of implementing some of the desired features discussed in the previous section. Given that four of the platforms are open source software systems making, it possible to freely modify them to address specific implementation needs. Other features provided on the platforms to enable adaptation and extension include: provision of APIs and libraries, support for website branding, and connectors and plugins. More details on the extendibility features of the platforms is included in the summary of the 11 platforms below.

#### **Summary of extensibility features of the platforms**

CKAN – an open source platform can be freely modified by users to meet their specific requirements. The platform provides two kinds of extension mechanisms - core extensions and the external extension. The core extensions are preinstalled with CKAN platform at installation time and only needs to be enabled when required. Examples core extensions in CKAN include datastore, multilingual and stats extensions. In addition, the plan allows for users to define custom fields for metadata schema. A guide is available at <http://docs.ckan.org/en/latest/extensions/> to support developers in extending the platform. There is also an vibrant online community supporting CKAN use and extension.

DKAN – it is a Drupal-based open source platform that is maintained by NuCivic (an open source enterprise). As DKAN is fundamentally a Drupal distribution, extensions or “modules” (over 10,000) already available for CKAN can be used to extend the platform. NuCivic maintains a number of modules specifically designed for DKAN but only available in the commercial Enterprise edition (NuCivic Data Enterprise). Therefore, in general, DKAN-specific extension (or modules) will have to be developed by the users. A guide is available at <http://docs.getdkan.com/dkan-documentation/extending-dkan> to support developers and there is a vibrant Drupal Community that also supports DKAN in addition to the more professional services offered by NuCivic.

Socrata - is a platform focusing on Open Data and services, it uses RDF metadata to describe the datasets, presented in Dublin Core and DCAT, it includes API Foundry for creating and deploying RESTful APIs on top of the data, data on the Socrata platform is accessible through the Socrata Open Data API (SODA), which provides RESTful interface for searching and reading data in XML, JSON or RDF; Socrata’s documentation is well-developed and presented, the developer portal including

numerous libraries for working with software as diverse as the R statistical platform, Scala, Ruby and Java, amongst others. Their developer portal is available at <http://dev.socrata.com/> ; Given the breadth of interactivity possible via the API, extending Socrata is straightforward. A library of existing extensions, released under various open source licenses (including the liberal MIT license) is available on their Github repository at <https://github.com/socrata>.

PublishMyData - is Swirrl's LinkedData publishing platform, It provides a fully hosted, as-a-service solution for organisations that need to publish Open and Linked Data, the core PublishMyData platform as an open source product, it offers RDF as the mechanism for metadata. This is extensible and underlies common metadata formats, There is fairly good documentation on using the API (<http://opendatacommunities.org/docs>) although there isn't very much public information on the interface for the SaaS or for the open source version of the software, The community edition of Swirrl's PublishMyData permits full customisation ([http://github.com/swirrl/publish\\_my\\_data](http://github.com/swirrl/publish_my_data)) while the online platform SaaS can also be customised or integrated into other systems.

Information Workbench - is a platform for building systems that works with all kinds of semantic data, it supports a variety of formats (RDF, N3, Turtle, N-Triples, TriG, TriX), Due to its flexible architecture the Information Workbench provides comprehensive software development kit for building applications and support extensions at different levels that make it adjusted and improved.

Enigma – is a platform that centralizes, mines and relates big public data about companies, people and locations, currently offering one of the largest and broadest repositories of public data, it provides API for accessing public data. The documentation for building applications with enigma and support is available online.

Junar - is a specifically Software-as-a-service platform offering one of the leading open data platforms. The system is able to import and use a wide variety of data formats and, as with all SaaS offerings, is useful to users looking for rapid deployment and the ability to develop and present insight from their data very rapidly, Junar provides an interactive API for each of their sites. This permits developers to experiment live on the database to see what results they can achieve, Junar is proprietary software and the range of published public APIs are only about downloading data rather than extending functionality or uploading data. While the userinterface can be customised, and new functionality written, The documentation for Junar is available on a set of wiki pages, many of which are customised for particular clients ([http://en.wiki.junar.com/index.php/Main\\_Page](http://en.wiki.junar.com/index.php/Main_Page)). This is not particularly easy to read and of limited value to developers.



OpenDataSoft - is a hosted software solution for open data publishers, it's written in Python with Django as the web framework, uses Exalead for search functionality, Hadoop for data processing and MongoDB as its data store, Data can be viewed and downloaded in different formats through the web interface or through the OpenDataSoft API, it has a limited free trial version, Access to the API can be public or protected with HTTP basic authentication or API keys. The API provides functionality for searching and lookup of datasets and there is documentation on how to use it (<http://public.opendatasoft.com/api/doc/>), but there is little on the software itself, from user to administration. Most guidance is provided via videos on the main site.

Callimachus - is a framework for data-driven applications based on linked data principles, it's allows Web developers to easily create data driven applications for the Web, In addition, Callimachus builds on either Sesame or Mulgara for RDF storage, AliBaba a RESTful object-RDF library and uses a template-by example technique for viewing and editing resources. One of the interesting aspects of Callimachus is that templates are parsed to build SPARQL from RDFa markup and then filled with query results, Callimachus community provides support and documentation for platform, It's open source and documentation is available on their website (<http://callimachusproject.org/documentation.xhtml?view>)

DataTank - is an open source tool that publishes data. These data can be stored in text-based files such as CSV, XML and JSON or in binary structures such as SHP files and relational databases. The DataTank will read the data out of these structures and publish them on the web using a URI as an identifier. It can then provide these data in any format a user wants, no matter what the original data structure was. In practical terms, this means that you can provide a JSON feed on a certain URI based on data somewhere on the web say, a CSV output from a google spreadsheet, the source code is publish on github (<https://github.com/tdt/>) and a comprehensive documentation is available on their website (<http://docs.thedatatank.com/>)

Semantic MediaWiki - is an extension of MediaWiki – the wiki application best known for powering Wikipedia, It's written in PHP, Semantic MediaWiki is queryable via a SPARQL interface and is able to return JSON data serialisation. Note, though, that the API only queries the database. Extending the software is done via independent modules that must be plugged into the software itself; MediaWiki is a platform in its own right and a vast number of software extensions have been written to enhance it. Similarly, the active developer community has written up comprehensive documentation that is available to support any custom extension or UX work that may be required ([https://semantic-mediawiki.org/wiki/Help:User\\_manual](https://semantic-mediawiki.org/wiki/Help:User_manual)).

Platforms	Extensible	Open Source	Extension Mechanisms	Guide	Customisable	Maintenance Community	Additional Info
CKAN	√	√	Core Extension External Extension	√	√	Yes & OKFN	<a href="http://docs.ckan.org/en/latest/extensions/">http://docs.ckan.org/en/latest/extensions/</a>
DKAN	√*	√	DKAN-specific Modules  Drupal Module  Custom Module	√	√	Drupal Community , NuCivic	<a href="http://docs.getdkan.com/dkan-documentation/extending-dkan">http://docs.getdkan.com/dkan-documentation/extending-dkan</a>
Socrata	√*	X	API Foundry	√	√	Socrata	<a href="http://dev.socrata.com/">http://dev.socrata.com/</a>
PublishMyData	√*	√*	Offers API and Querying	√	√*	Swirll	<a href="http://docs.publishmydata.com/developers/">http://docs.publishmydata.com/developers/</a>
Information Workbench	√*	√*	Supports Extensions	X	√	FluidOps	<a href="http://www.fluidops.com/en/support">http://www.fluidops.com/en/support</a>
Enigma	X	X	Offers API	X	X	Enigma	<a href="http://enigma.io/solutions/api/">http://enigma.io/solutions/api/</a>
Junar	√*	X	Offers API	X	X	Junar	<a href="http://www.junar.com/">http://www.junar.com/</a>
Open Data Soft	√*	X	Offers API	√	X	OpenDataSoft	<a href="https://public.opendatasoft.com/api/doc/">https://public.opendatasoft.com/api/doc/</a>
Callimachus	√	√	Offers API	√	√	Callimachus project	<a href="http://callimachusproject.org/docs/1.4/callimachus-reference.docbook?view#Callimachus_REST_API">http://callimachusproject.org/docs/1.4/callimachus-reference.docbook?view#Callimachus_REST_API</a>
DataTank	√	√	Offers API	√	X	iMinds	<a href="http://docs.thedatatank.com/">http://docs.thedatatank.com/</a>
Semantic MediaWiki	√	√	Offers API and Supports Extensions	√	√	Semantic MediaWiki community	<a href="https://semantic-mediawiki.org/wiki/Ask_API">https://semantic-mediawiki.org/wiki/Ask_API</a>

√\* - limited feature, usually due to proprietary nature of the platform

### 3.5 SUMMARY

It this section we have provided the summary of the selected Open Data Platforms: available features, architecture, extensibility of the platforms and the technological overview. Table with the general summary of ODP features is available in Appendix 3.

## 4 PERCEPTIONS OF STAKEHOLDERS ON OPEN DATA PLATFORMS

---

We present in this section the summary of the data obtained from interviews and workshop sessions on barriers and limitations of current open data platforms as well as desired features to address some of the identified shortcomings. Categories of stakeholders engaged include open data consumers, enablers, suppliers and mediators. Section 4.1 presents the barriers while Section 4.2 the suggested features.

### 4.1 BARRIERS TO THE USE OF STATE-OF-THE-ART OPEN DATA PLATFORMS

We briefly discuss some examples of the barriers identified by stakeholders in this section. For each barrier we: 1) specify a generic class (i.e. coded each barrier instance) for the problem such as “Non-relevancy” or “Poor awareness” and then 2) associate it with a high level transparency construct, e.g. Accessibility and a more specific construct such as “Availability”. This coding is based on the models described in Section 2. Examples of barriers identified include difficulty in locating datasets of interest, poor context for available data on platforms and poor user interface design for current open data portals. More information on some of the barriers associated with use and adoption of state-of-the-art platforms are presented in the table below.

*Table 23: Excerpt from Data on Shortcomings of State of the art platforms*

Barrier	Stakeholder	Generic problem	Top Transparency Construct	Lower Level Transparency construct
Available open datasets are not 'relevant' or 'speaking to' people's	Consumer	Non-Relevancy	Accessibility	Availability
Open data vs Eincodes (Postcodes), lack of open look-up profile, missed opportunity for open data generation	Enabler	Poor Awareness	Accessibility	Publicity
Metadata problems	Supplier/Mediator	Poor Data Quality	Informativeness	Clarity
There is a lack of useful data	Consumer	Non-Relevancy	Accessibility	Availability

Shortage of technical resources to collect data	All Stakeholders	Data Capture from Source	Accessibility	Availability
Difficulty in finding data - potential data dump rather than good standards for cataloguing, describing, linking data	Consumer	Poor Data Quality	Understandability	Conciseness
Reliability of data feeds and keeping them updated; old data is gone	Consumer	Poor Data Quality	Informativeness	Currency
Poor service design and management	Supplier/Mediator	Poor Platform Usability	Usability	User-Friendliness
Information spread out over multiple organisations, lack of one	Supplier	No Data Consolidation	Understandability	Integration
Poor information management	Supplier/Mediator	Poor data management practices in agencies	Auditability	Controllability
Inadequate technical expertise to produce data in a usable format	Supplier	Poor Data Quality	Usability	Data Format
Lack of available accredited open data training courses	All Stakeholders	Poor Data Literacy Skills	Accessibility	Data Literacy
Dilution of information available to the public	Supplier/Mediator	Poor Data Quality	Informativeness	Integrity
Data on screen may be displayed in a technical way or use unfamiliar technical language	Supplier/Mediator	Technicality of Data Presentation	Understandability	Comprehension
Citizens may not always have up to date browsers on their computers	Consumer	Technical Interoperability	Usability	Operability
Minimal publicity about data available leading to lack of awareness of its	Supplier/Enabler	Poor Awareness	Accessibility	Publicity
Data is in a dense form and requires design input to make it accessible	Consumer	Technicality of Data Presentation	Understandability	Comprehension
Lack of information about the circumstances of data production	Consumer	Poor Data Quality	Informativeness	Metadata quality and Provenance
Lack of user-friendly file-formats, Lack of user-friendly interface	Consumer	Usability of data	Usability	Operability - data formats
Lack of engaging activities/information for those users who arrive to a page without a clear	Consumer	Weak user engagement	Usability	User-Friendliness

Lack of examples available for smart use of open data	Consumer	No smart use example	Usability	User-Friendliness
Lack of access to necessary software / hardware to utilise Open	Mediator/Consumer	Poor access to open data platforms	Accessibility	Resource constraints
Lack of sufficient broadband / bandwidth to successfully interact	Enabler	Poor access to open data platforms	Accessibility	Resource constraints
Level of openness and licences for use in commercial remit	Enabler	Openness of data	Usability	Openness
Quality of data, right formats to the right audience e.g. spreadsheets for 'tourists' and feeds/API	Supplier	Poor Data Quality	Usability	User-Friendliness
Usability; need preview, mapping, visualisation, multiple data layering	Consumer	Usability of data	Usability	User-Friendliness

## 4.2 SOLUTIONS AND DESIRED FEATURES FOR FUTURE OPEN DATA PLATFORMS

This section captures some examples of the solutions and concrete platform features suggested to address three categories of needs on of open data end-users – 1) information needs, 2) social and collaboration needs, and 3) understandability, usability and decision making needs. As in the case of the barriers described in Section 4.1, suggested solutions and features can associated with concrete transparency constructs and sub-constructs described in Section 2. In fact, the categories of the needs specified earlier are directly linked to the open data transparency and social interaction on open data. Examples of suggested solutions include making available specific datasets related to immediate communities of stakeholders, datasets on key indicators of neighbourhoods such as crime statistics, health, and environment. Dataset rating, comments on datasets, collaborative curation of datasets and prioritization of requested datasets through voting were also suggested under social and collaboration needs. In the area of understanding, usability and decision making, users requested for customisable dashboards, map based search and query facilities, modelling tools as well as data integration tools, support for linked data for comparing datasets. Table 2 presents more examples of suggested features.

Table 24: Excerpt from Desired Features in Future Open Data Platforms

<p>Information needs</p> <ul style="list-style-type: none"> <li>Inventory of local business people – support local enterprise</li> <li>Key indicators for my neighbourhood (social, crime, environment, health, etc.) for informed decision making</li> <li>Local info of all kinds – planning, sports, cultural, commercial, social, councillors</li> </ul>
<p>Social and Collaborative Needs</p> <ul style="list-style-type: none"> <li>Anonymity</li> <li>Closed loop, share results of interactions &amp; collaborations</li> <li>Contact tools for finding PA, forums, public participation, network, social media interaction, twitter, facebook</li> <li>Dataset rating &amp; ranking, Calendar, wall style fast feedback, live chat, comments on dataset, blogs, collaborative editing, curating, adding metadata for dataset</li> <li>Diversity of engagement – creativity, inclusion, new knowledge &amp; value</li> <li>Embed data for viral travel of data + its conversations</li> <li>Expert facilitation</li> <li>Live webcast with feedback, newsfeed for decision,</li> <li>Mission/vision statement for discussion</li> <li>Original data location – show paths to where it is shared</li> <li>Prioritisation of data request based on needs/voting</li> <li>Project management tool</li> <li>Reward system, gamification, acknowledgement</li> <li>Verification/traceability of account</li> </ul>
<p>Understandability, Usability and decision-making needs</p> <ul style="list-style-type: none"> <li>Modelling and stimulations</li> <li>Animations &amp; interactive visualisation; Predictive analytics</li> <li>Animations, pictures, browsing exploration experience</li> <li>APIs,</li> <li>Customisable Dashboards, personalisation</li> <li>Data availability over several portable devices; Customised display – pull in from other platforms + layer data</li> <li>Data integration</li> <li>Data mining tools &amp; analysis tools for information extraction to support decision-making</li> <li>In-file data descriptors</li> <li>Interactions, 'rate my service', submit suggestions on map + get feedback</li> <li>Interactive graphical representations as transparency enhancing tools, promote easy reading, understandability, making sense of data</li> <li>Linked data for comparison</li> <li>Map + zoom Vs recovery</li> <li>Map based search &amp; queries</li> <li>Metadata management</li> <li>Modelling tools, layered maps</li> <li>Personalisation – search with filter, especially with memory, notifications &amp; updates</li> <li>Polls and surveys</li> <li>Public or anonymous profile options</li> <li>Q &amp; A mechanism</li> <li>Question &amp; answer, feedback mechanism monitored up-to-date</li> <li>Scheduling services – identify what is logged, actioned or closed</li> <li>Statistics under-pinning policies</li> </ul>

## 5 SUMMARY OF FINDINGS

---

In this section, we present the findings from the analysis of literature review and primary data gathered from workshop and interviews. In total, eleven platforms were reviewed and evaluated in the study including: CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, DataTank, OpenDataSoft, Callimachus, DataTank and Semantic MediaWiki. As shown in Table 16. Five of these platforms are open source while the remaining six are proprietary platforms that provide limited number of open source components for community use. Three of these platforms are also offered as Cloud services (Software-as-a-Service) and one as an online service. The Semantic MediaWiki is a specialised platform for publishing textual contents based on semantic models.

Table 25: *Summary of Open Data Platforms*

	Standalone	Cloud Service	Online Service
Proprietary	<ul style="list-style-type: none"><li>○ Socrata</li><li>○ PublishMyData</li><li>○ Information Workbench</li></ul>	<ul style="list-style-type: none"><li>○ Junar</li><li>○ OpenDataSoft</li></ul>	<ul style="list-style-type: none"><li>○ Enigma</li></ul>
Open Source	<ul style="list-style-type: none"><li>○ CKAN</li><li>○ DKAN</li><li>○ Callimacahus</li><li>○ Semantic MediaWiki</li></ul>	<ul style="list-style-type: none"><li>○ DataTank</li></ul>	

### 5.1 TRANSPARENCY-SUPPORTING FEATURES ON OPEN DATA PLATFORMS

We investigated the features available on state-of-the-art platforms by analysing contents from scholarly literature and documents describing the platforms and also based on our systematic exploration of selected instances of these platforms. The following set of criteria was employed in the evaluation:

- Metadata management and supported several file formats
- Search and Indexing service
- Integration with social media sites like Twitter and collaborative tools like GitHub
- Supports part of open data publishing workflow as well as catalogue management
- Harvesting and federation of dataset catalogs
- On-platform applications for data analytics
- Support rich visualizations of datasets
- Personalization through different end-user settings
- Customisation through the use of different harvesting models and user
- Support for datasets licensing and
- Support for user accessibility



Socrata, CKAN, DKAN and Semantic MediaWiki stand out by providing full-fledged features that support at least 9 of the 12 criteria used in evaluating the platforms (see Table 1). Other platforms support between 1 to 7 fully-fledged features. Overall, while features the use of social media channels, customisation and personalisation of platform features are common place in state-of-the-art platforms, *support for metadata schema adaptation, options for visualisation of datasets and accessibility (including at granular level) to datasets are limited*. Features like availability of publishing pipelines or workflows are visualisation still relatively limited on existing platforms. Whereas, personalisation and customisation feature are very common features across platforms. However, it must be noted that in terms of social media integration, these platforms simply allow a link to social media accounts. Personalisation in the context of this evaluation is only limited to end-user ability to change the behaviour of the platform based on preferences and does not extend to the aspects like the recommendations of datasets to end-users based on relationships with other users or preferences.

## 5.2 PERCEPTIONS ON SHORTCOMINGS OF OPEN DATA PLATFORMS

Our analysis showed that the most common barrier to the use of open data platforms and open data *is perceived poor quality of open data* available on the platforms. Poor data quality according to stakeholders is associated with poor metadata, failure to use the right format for different audience and difficulty in locating data of interest. Other barriers identified are related to non-relevancy of available datasets, usability of platforms and data available on the platform and lack of example of prior use of available datasets.

Figure 25: Perceived Barriers to Use and Adoption Open Data Platforms

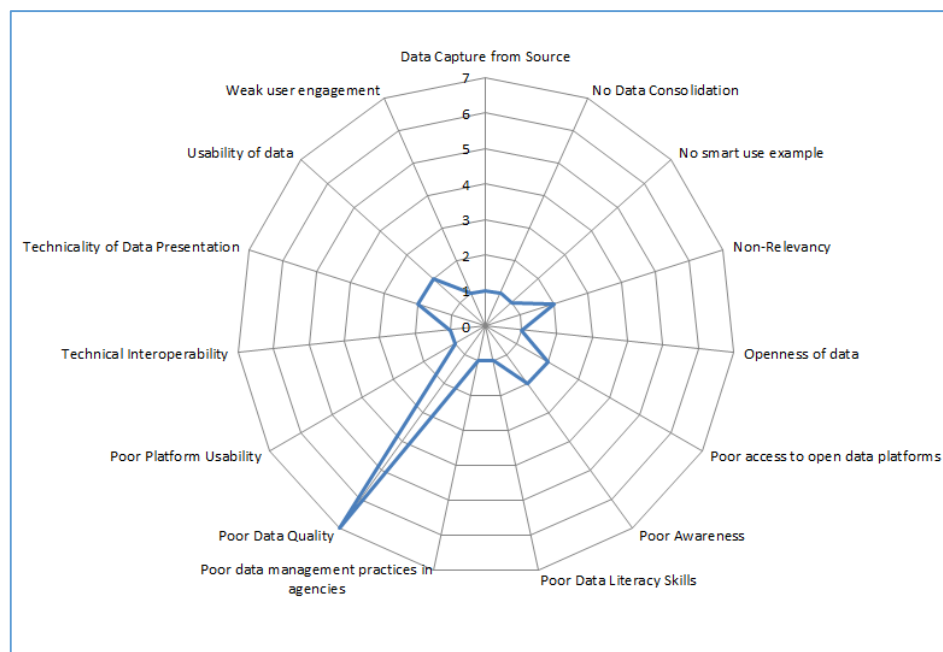


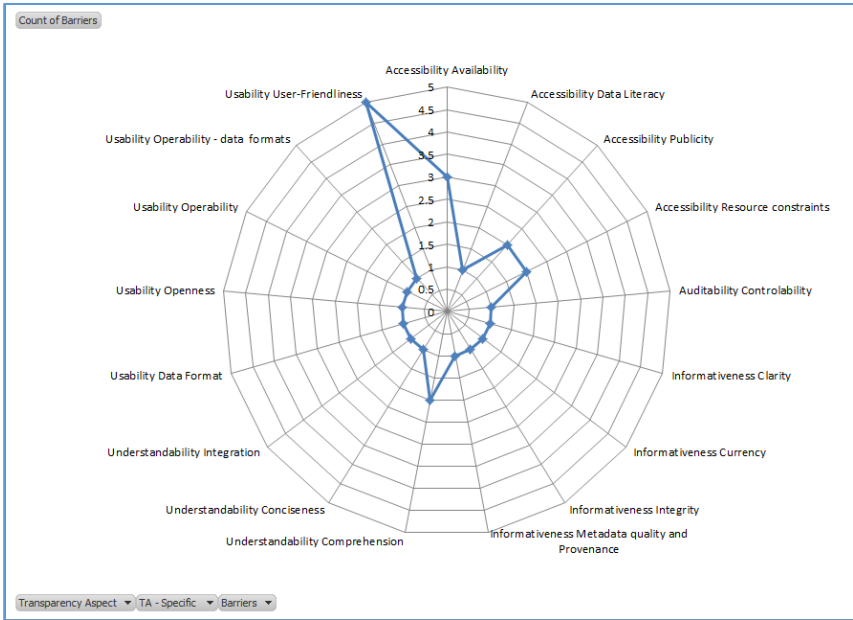
Table 26: Summary of Platform Features

FEATURES	CKAN	DKAN	SOCRATA	PUBLISH MY DATA	INFO WKBENCH	ENIGMA	JUNAR	ODS	CALLIM	DATATK	SMWIKI
DATA, METADATA & FILE FORMAT STANDARDS	●	●	●	●	●	●	●	●	●	●	●
SEARCH & INDEXING	●	●	●	●	X	●	●	●	X	●	●
SOCIAL MEDIA, SHARING & COLLABORATION	●	●	●	●	●	X	●	●	●	X	●
PUBLISHING WORKFLOW	●	●	●	●	●	●	●	●	●	●	●
HARVESTING, FEDERATION & CATALOGUE	●	●	●	●	X	X	●	●	●	X	●
DATA ANALYSIS	●	●	●	X	●	●	●	●	X	●	X
VISUALISATION	●	●	●	X	●	X	●	●	X	X	●
PERSONALISATION	●	●	●	●	●	X	●	●	●	●	●
CUSTOMISATION	●	●	●	●	●	NA	●	●	●	●	●
LICENSING FOR DATASET	●	●	●	●	X	X	X	●	X	X	●
ACCESSIBILITY	●	●	●	●	●	NA	●	●	●	●	●
EXTENSIBILITY	●	●	●	●	●	●	●	●	●	●	●
TECHNICAL ENVIRONMENT	Python	PHP, Drupal	Scala	Ruby on rails	Java & Web apps	NA	Java & Python	NA	Java	PHP	PHP
OTHERS	Good manual Simple to use	Easy to use platform	Tracking & Measure of performance	Flexible, cloud-based, easy to use	R stat, support transparency, linked data	Reliable, scalable, large OD Analyses	Track & measures user impact on OD	Remote web services; easy deployment	Guides, videos, tutorial. Linked data	Deal with fraud, aids transparency	None

● denotes full-fledged solution, ● denotes limited solution, x denotes that solution is not provided, NA denotes information not available

The figure below the associated transparency issues that are related to the above barriers:

Figure 26: Data Transparency attributes related to the Perceived Barriers



*Stakeholders Perspectives on Barriers* - The study adopted a typology for open data stakeholders involving four categories including 1) Consumers – end-users that directly use the datasets published on the platforms; 2) Enabler – entities in the ecosystem like civil platform developers, data analysts and wranglers that provide the necessary infrastructure and services for both suppliers and consumers; 3) Mediators – entities in the open data ecosystem that support customers in accessing and using the published datasets. They include Civil Society Organizations providing training support to end-users or accessing and using open data on behalf of the public; Supplier – includes all organizations, government agencies in particular that involved in producing and publishing open data. Our analysis of the data gathered from workshops show that the problem of poor data quality is a shared concern for *all categories* of stakeholders. *Enablers* in addition consider poor access of end-users to open data platforms and poor awareness about open data as major barriers that should be addressed. Mediators and Suppliers consider poor data management practices in government agencies, poor usability of platforms, and technical presentation of data as major issues to tackle. Supplies in addition consider poor awareness of the availability of open data as a problem. The concerns for the different stakeholder categories are shown in Table 27 below.

Table 27: Stakeholders Perception of Barriers to Use of Open Data and Open Data Platforms

No	Category	Concerns
1	Consumer	<ul style="list-style-type: none"> <li>○ No smart use example</li> <li>○ Non-Relevancy</li> <li>○ Poor Data Quality</li> <li>○ Technical Interoperability</li> <li>○ Technicality of Data Presentation</li> <li>○ Usability of data</li> <li>○ Weak user engagement</li> </ul>
2	Enabler	<ul style="list-style-type: none"> <li>○ Openness of data</li> <li>○ Poor access to open data platforms</li> <li>○ Poor Awareness</li> </ul>
3	Mediator	<ul style="list-style-type: none"> <li>○ Poor access to open data platforms</li> <li>○ Poor data management practices in agencies</li> <li>○ Poor Data Quality</li> <li>○ Poor Platform Usability</li> <li>○ Technicality of Data Presentation</li> </ul>
4	Supplier	<ul style="list-style-type: none"> <li>○ Poor data management practices in agencies</li> <li>○ Poor Data Quality</li> <li>○ Poor Platform Usability</li> <li>○ Technicality of Data Presentation</li> <li>○ Poor Awareness</li> </ul>

### 5.3 DESIRED FEATURES FOR FUTURE OPEN DATA PLATFORMS

The desired features contributed by stakeholders for next generation open data platforms were captured under two categories: 1) Social and Collaboration, and 2) Understandability, Usability and Decision making needs. Dataset rating and feedback on datasets, Wall style feedback, collaborative curation of datasets, prioritization and voting on dataset requests, reward system and gamification are some of the features expressed under the social and collaborative needs. To enable better understandability, usability and better decision making with next generation platforms, users requested for customisable dashboards, data mining tools and custom visualization tools, support for linked data and map based search as well as question and answering features. The cloud-tag below () was generated from the contributed solutions and features to identified stakeholder needs and barriers. Figure 27 shows relative distribution of features across three categories.

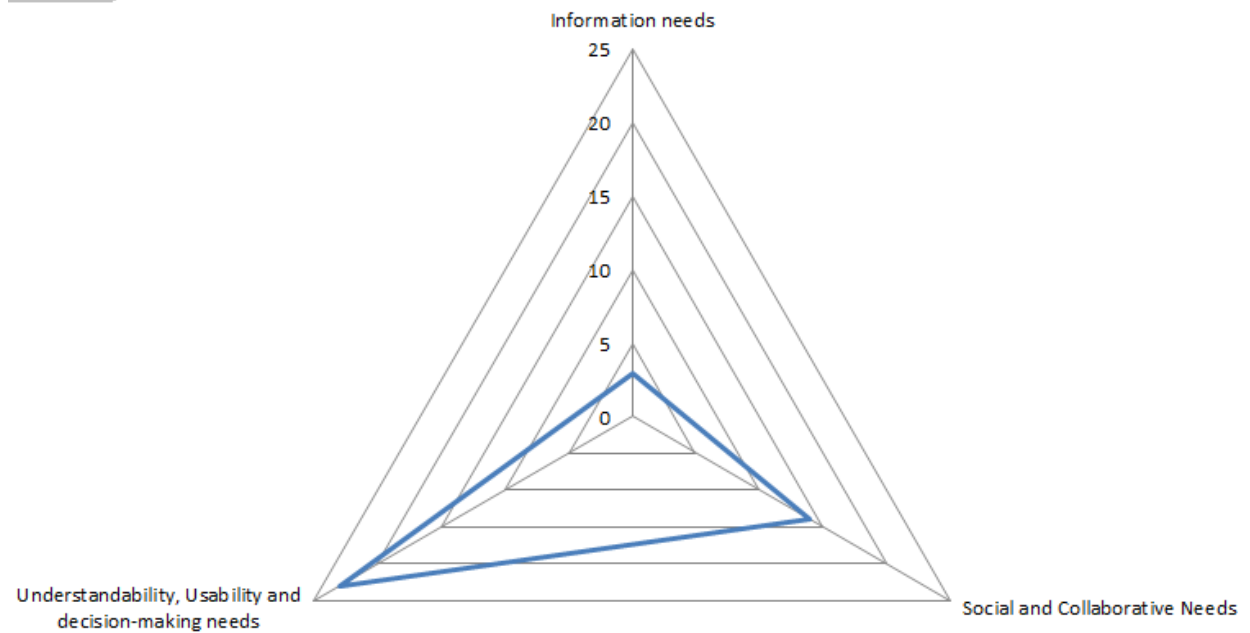


Figure 27: Number of Features in Desired Features Categories

Figure 28: Keywords generated from desired features for Open Data Platforms



## 5.4 EXTENSIBILITY OF OPEN DATA PLATFORMS

Based on the four detailed criteria for extensibility of platforms, CKAN, DKAN and Semantic MediaWiki are the most extensible providing free and open source codes, rich set of extension mechanisms and open architecture, guide to support developers in building such extensions and support for additional fields in the metadata schema. However, Callimachus and DataTank being open source could also be modified as desired albeit at a much higher cost compared to the above that provide explicit extension mechanisms. The detailed table of extension features is presented in the table below.

Table 28: Availability of Extensibility Mechanism in Open Data Platforms

Platforms	Extensible	Open Source	Extension Mechanisms	Guide Available	Customisable Metadata
CKAN	●	●	●	●	●
DKAN	●	●	●	●	●
Socrata	•	x	●	●	●
PublishMyData	•	•	•	●	•
Information Workbench	•	•	●	x	●
Enigma	x	x	•	x	x
Junar	•	x	•	x	x
Open Data Soft	•	x	•	●	x
Callimachus	●	●	•	●	●
DataTank	●	●	•	●	x
Semantic MediaWiki	●	●	●	●	●

● denotes extensive solution, • denotes limited solution, x denotes that solution is not provided

## 6 DISCUSSION

---

In addition to providing good evidence base for determining a suitable open data platform to adopt as base platform for the Route-To-PA project, the study also seeks to contribute to the start-of-the-art in the area of assessment and evaluation of open data platforms in general and in the context of government transparency. While there are a number of studies evaluating open government data programmes, very few address open data infrastructures. In this sense, given the richness of data sources and analysis with respect to selected platforms, we argue that the study addresses a critical knowledge gap in the open data community. In fact, the comprehensive information collected through workshops on barriers and needs of stakeholders provide robust input to the discussion about the nature of next generation open data infrastructure.

Given the depth and scope of our study, major findings from our study as discussed in Section 5 subsumes findings and proposals from the few existing studies in this area. For example, studies like those of Alexopoulos et al. (2014)<sup>42</sup> argues for the need of feedback loops in open data platforms to enable discussion, feedback mechanisms for open data platform users. Consistent with findings from the interview in their study, our study also revealed stakeholders' strong interest in social features dataset rating and feedback on datasets, wall style feedback, collaborative curation of datasets, prioritization and voting on dataset requests. In fact, these specific features strongly agree with the novel web 2.0 functionalities proposed by these authors.

The results from World Bank study on Technical assessment of Open Data Platforms for National Statistical Organizations<sup>16</sup> show the need to improve technical documentations of open data platforms, ensure public APIs and endpoints are interoperable, ensure that metadata and URI's conforms to W3C standards, allow natural language search and metadata faceting for browsing, provide dashboard and visualisations for user engagement and developing data engagement tools for improving data-quality and reuse. While most of these recommendations are related to technical aspects of the platforms, user-engagement and ease of search and browsing of data is included. Our findings elaborates on social and engagement aspects of open data platforms through the needs articulated under the Social and Collaboration as well as the Understandability, Usability and Decision making needs. Another source of validation for our findings (even with differences in our set of criteria) is found in the study documented in the Thesis on Evaluation of Platforms for Open Government Data<sup>43</sup> with findings that largely agrees with our favourable evaluation of Socrata and CKAN as the most advanced platforms (See Table 26).

Perspectives by interviewed experts also provide valuable insights into serious limitations of current platforms and desired features. According to these experts, the inability of open data portals to provide access to datasets in a style similar to how ordinary end-users uses the Google Search engine to find information of interest. They also pointed out the non-interactive (one-way) nature of the platforms making them unable to support bi-directional interactions. The fact that open data portals do not really provide answers to end-user questions was also highlighted as one of the barriers to user engagement. The difficulty in meeting this kind of needs was raised by another expert that explained the need for easy data integration on open data portals as answers to queries or questions are found within one single dataset. Finally, experts argue that value proposition for open data publishers is relatively weak. These insights in our opinion are very important for the open data community.

There were a few challenges encountered in the development of the deliverable. Given that there has been no past study looking to evaluating. One conceptual challenge is to how effectively link transparency constructs to evaluation a technical artefact like an open data platform. Another challenge was how to streamline contributions from the different pilot

---

<sup>42</sup> Alexopoulos, C., Loukis, E., and Charalabidis, Y. 2014. "A Platform for Closing the Open Data Feedback Loop based on Web 2.0 functionality," *Journal of eDemocracy and Open Government* (6:1), pp. 62–68 (available at <http://www.jedem.org/article/view/327/270>).

<sup>43</sup> Stråle, J., and Lindén, H. 2014. "An evaluation of platforms for open government data," Thesis, Computer Engineering Department, KTH, Sweden

workshops even though the workshop design was largely same, with some freedom provided to each pilot for addressing local or specific contextual requirements. Finally, it was difficult getting data from these workshops, consolidating them a coherent whole within a very limited time that was allowed for timely submission.

## 7 CONCLUSION

---

This report on “State-of-the Report and Evaluation of Existing Open Data Platforms” documents the findings from our investigation on regarding existing open data platforms. This report complements existing reports as it focuses on evaluation of the platform from perspectives of open data transparency. Other existing reports have focused largely on the technical aspects of the platforms. In addition, the complementary analyses of the stakeholders input on barriers and desired features provide a pragmatic context for the technical evaluation. Apart from the evaluations, we have also synthesized technical architectures for open data platforms based reviewed materials and our exploration of open data platform instances. Guided by our findings, we conclude as and recommend as follows:

- That a few state-of-the-art open data platforms such as CKAN, Socrata, DKAN, Semantic MediaWiki provide well-developed features to support good data transparency and quality when publishing datasets. With three of these platforms are open-source and explicitly provide extension mechanisms, they arguably stand out as choice base platforms for building next generation open data platforms. CKAN, DKAN and Semantic MediaWiki in particular have a very vibrant developer community that could provide the necessary support in any further development of these platforms.
- Despite these features provided by some of these platforms as highlighted in above, lessons end-user perspective, there are still significant challenges that must be tackled for these platforms to be adopted and used as desired by public administrations and other stakeholders. One of the barriers that stand out in this area is the perceived poor quality of datasets published on these platforms. Consequently, platforms developers would have to directly address aspects of open data quality such as poor context and provenance for published datasets and non-viable data feeds. Feature to explicitly rate datasets in different data quality dimensions could be useful in this regard.
- From the stakeholders’ perspectives, social features such as dataset rating, voting and wall-style feedback on datasets and advanced analytics tools such as customisable dashboards, custom visualisation tools should be considered in future enhancement of open data portals. This is congruent with findings from technical evaluation of state-of-the-art platform features.
- Open and extensible base technology platforms are available for innovation relating the development of next generation open data platform with features described above. In particular, CKAN, DKAN and Semantic MediaWiki are candidate base platform for such innovation activities.



## BIBLIOGRAPHY

- Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). Designing a Second Generation of Open Data Platforms : Integrating Open Data and Social Media. *E-Gove, LNCS 8653*, 230–241.
- Antikainen, M., Mäkipää, M., & Ahonen, M. (2010). Motivating and supporting collaboration in open innovation. *European Journal of Innovation Management*, 13(1), 100–119.  
<http://doi.org/10.1108/14601061011013258>
- Baldwin, C. Y., & Woodard, C. J. (2009). (2009). *The architecture of platforms: A unified view*. Harvard Business School.
- Bonsón, E., Torres, L., Royo, S., & Flores, F. (2012). Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly*, 29(2), 123–132.  
<http://doi.org/10.1016/j.giq.2011.10.001>
- Boyd, M. (2014). ENIGMA OPEN DATA PLATFORM SECURES \$4.5M FUNDING. *Programmableweb*, January(January 13, 2014), 20–23. Retrieved from <http://www.programmableweb.com/news/enigma-open-data-platform-secures-4.5m-funding/2014/01/30>
- Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The State of Open Data: Limits of Current Open Data Platforms Categories and Subject Descriptors. In *Www*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.309.8903>
- Callimachus. (n.d.). Getting started with Callimachus. Retrieved March 18, 2015, from Getting started with Callimachus
- CKAN. (n.d.). CKAN, the world's leading open-source data portal platform. Retrieved March 14, 2015, from <http://ckan.org/>
- Cleland, B., Galbraith, B., Quinn, B., & Humphreys, P. (2013). Platform Strategies for Open Government Innovation. *Department of Management and Leadership, University of Ulster, United Kingdom*.
- Colpaert, P., Dimou, A., Sande, M. Vander, Breuer, J., Van, M., Mannens, E., ... Dimou, A. (2014). A three-level data publishing portal. Athens: European Data Forum. Retrieved from [http://2014.data-forum.eu/sites/default/files/pdf/edf2014\\_submission\\_43.pdf](http://2014.data-forum.eu/sites/default/files/pdf/edf2014_submission_43.pdf)
- DataTank. (n.d.). About DataTank. Retrieved March 16, 2015, from <http://www.datatank.co.uk/about-us.php>
- Duval, A., & Brasse, V. (2014). How to ensure the economic viability of an open data platform. *Procedia Computer Science*, 33, 179–182. <http://doi.org/10.1016/j.procs.2014.06.030>
- Eberius, J., Braunschweig, K., Thiele, M., & Lehner, W. (2012). Identifying And Weighting Integration Hypotheses On Open Data Platforms Categories and Subject Descriptors. *Wod*, 22–29.  
<http://doi.org/10.1145/2422604.2422608>
- Edlich, S., Singh, S., & Pfennigstorf, I. (2013). Future mobile access for open-data platforms and the BBC-DaaS system. *Proceedings of SPIE - The International Society for Optical Engineering*, 8667, 866710.  
<http://doi.org/10.1117/12.2002871>

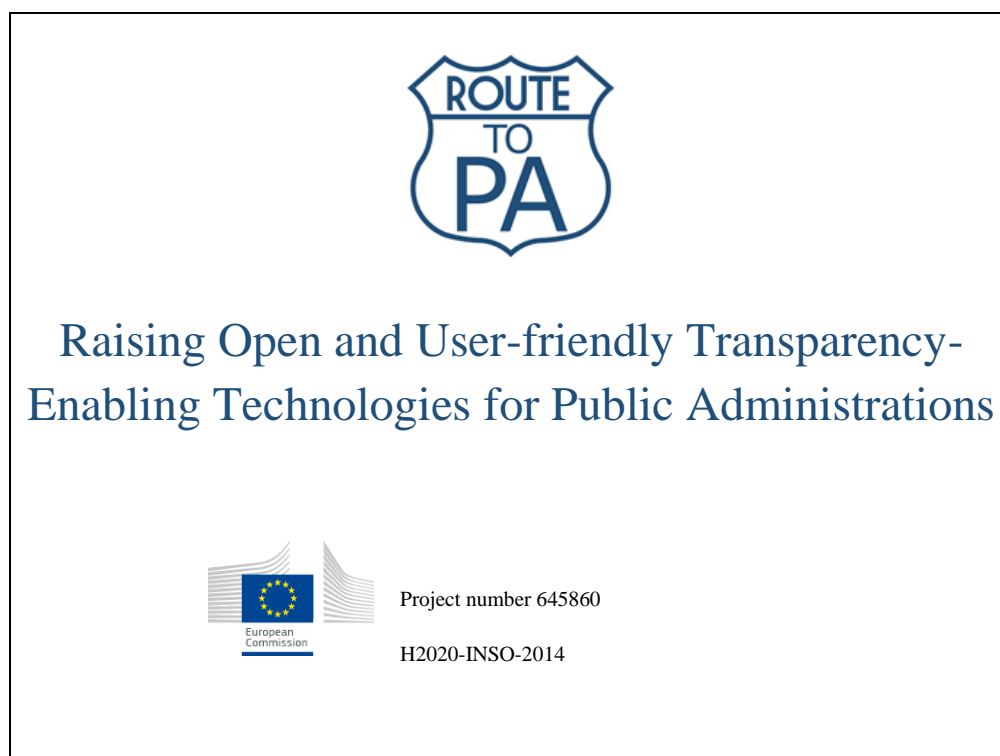
- European Commission. (2015). What are European Technology Platforms ? Retrieved March 13, 2015, from [http://ec.europa.eu/research/innovation-union/index\\_en.cfm?pg=etp](http://ec.europa.eu/research/innovation-union/index_en.cfm?pg=etp)
- Fishenden, J., & Thompson, M. (2012). Digital Government, Open Architecture, and Innovation: Why Public Sector IT Will Never Be the Same Again. *Journal of Public Administration Research and Theory*. Retrieved from <https://markthompson1.files.wordpress.com/2012/02/j-public-adm-res-theory-2012-fishenden-jopart-mus0221.pdf>
- Halonen, A. (2012). *Being Open About Data: Analysis of the UK open data policies and applicability of open data*. London. Retrieved from [www.finnish-institute.org.uk](http://www.finnish-institute.org.uk)
- Hoppin, A., Byrnes, A., & Couch, A. (2013). Open-Source Open Data Platforms The Proprietary SaaS Competition - Circa 2013. Retrieved from [http://www.w3.org/egov/wiki/images/f/f1/W3C\\_OpenSource\\_OpenData.pdf](http://www.w3.org/egov/wiki/images/f/f1/W3C_OpenSource_OpenData.pdf)
- Iemma, R., Morando, F., & Osella, M. (2014). Breaking Public Administrations ' Data Silos: The Case of Open-DAI, and a Comparison between Open Data Platforms. *JeDEM*, 6(2), 112–122. Retrieved from <http://www.jedem.org>
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), 258–268. <http://doi.org/10.1080/10580530.2012.716740>
- Lapi, E., Tcholtchev, N., Bassbouss, L., Marienfeld, F., & Schieferdecker, I. (2012). Identification and utilization of components for a linked open data platform. *Proceedings - International Computer Software and Applications Conference*, 112–115. <http://doi.org/10.1109/COMPSACW.2012.30>
- Lindén, H., & Stråle, J. (2014). *AN EVALUATION OF PLATFORMS FOR OPEN GOVERNMENT DATA*. Kth School of Technology and Health Handen, Sweden. Retrieved from <http://www.diva-portal.org/smash/get/diva2:723341/FULLTEXT01.pdf>
- Margetts, P. D. and H. (2010). The second wave of digital era governance. In *American Political Science Association Conference, 4 September 2010, Washington DC, USA*. Unpublished. Retrieved from <http://eprints.lse.ac.uk/27684/>
- OpenDataSoft. (n.d.). OPENDATASOFT IS THE #1 TURNKEY SOLUTION DEDICATED TO PUTTING BUSINESS USERS' DATA TO GOOD USE. Retrieved March 16, 2015, from <http://www.opendatasoft.com/>
- Programmableweb. (n.d.). Enigma API. Retrieved March 15, 2015, from <http://www.programmableweb.com/api/enigma>
- Rouse, M. (n.d.). Platform. Retrieved March 6, 2015, from <http://searchservervirtualization.techtarget.com/definition/platform>
- Russell, Kristin, M. P. (n.d.). Citizen and Government Collaboration Made Easy. Retrieved March 10, 2015, from <http://www.socrata.com/products/open-data-portal/>
- Swirrl. (n.d.). Features. Retrieved March 15, 2015, from <http://www.swirrl.com/publishmydata#features>

- Taatila, V. P., Suomala, J., Siltala, R., & Keskinen, S. (2006). Framework to study the social innovation networks. *European Journal of Innovation Management*, 9(3), 312–326.  
<http://doi.org/10.1108/14601060610678176>
- Tiwana, A., Konsynski, B., & Bush, A. a. (2010). Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research*, 21(4), 675–687.  
<http://doi.org/10.1287/isre.1100.0323>
- Walther, U. (n.d.). Informatin Workbench - How it works. Retrieved March 18, 2015, from  
[http://www.fluidops.com/en/portfolio/information\\_workbench/](http://www.fluidops.com/en/portfolio/information_workbench/)
- Wasko, M., & Faraj, S. (2000). “It is what one does”: why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, 9(2-3), 155–173.  
[http://doi.org/10.1016/S0963-8687\(00\)00045-7](http://doi.org/10.1016/S0963-8687(00)00045-7)
- World Bank. (2014). *Technical Assessment of Open Data Platforms for National Statistical Organisations*. World Bank, Washingto DC. Retrieved from  
<http://documents.worldbank.org/curated/en/2014/10/20451797/technical-assessment-open-data-platforms-national-statistical-organisations>

## APPENDICES

---

### APPENDIX 1: REPORTS OF INTERVIEWS WITH ODP STAKEHOLDERS



## **Evaluation of Existing Open Data Platforms**

### **Interview Protocol for Stakeholders**

(Draft, version 0.2, 21042015)



WISE&MUNRO



## Abstract

1. **About the interview:**

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date \_27/April/2015

Time \_10:55 am

Location \_Insight Centre for Data Analytics, NUI Galway

Name (Interviewer) \_Ed. Osagie and Waqar Mohammed

2. **Notes to interviewee:**

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 13 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 30 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

### 3. **Introduction:**

*Question coverage:* Our questions cover 3 major areas:

1. platform challenges
2. desired platform features and priorities of the features, and
3. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:

1. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
2. Platform developers or ODP service providers (midstream community)
3. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

***Peripheral data collection about the interviewee's and his/her company or organisation:***

Name Niall O'Brolchain

Company or organisation Insight Centre for Data Analytics, NUI Galway

Stakeholder group Academic/Researcher/Promoter of Open Data; Involved application of Open Data (OD)  
Institute node in Insight

Position/designation Researcher

Typical task at work Managing projects, research, involved in making applications

Interviewee's signature (permission to record detail of interview)

\_\_\_\_\_Date:\_\_\_\_\_



#### 4. Interview proper:

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [state stakeholder/user group here] of ODP and/or (if a company/organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

#### *Interview questions:*

##### A) Challenges:

##### 1. Challenges associated with the use of open data portals and platforms.

##### a) Access to open data published or available on the platform:

- i. *In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[You can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

---

#### **Note (1)(a)(i): Talking about challenges**

The answer to this question depends on different points of view talking about different users. The challenges facing users are many, for example Google, though not exactly Open Data portal, but if you can find more datasets very quickly on it than on specialised Portals with tools for Open Data (OD) then there is no need to go to the OD portals unless you have proper added value on the standard search engine for OD Portal before people can be encouraged to use it.

##### **i) Searchability of data:**

The Open Data Platforms (ODPs) and Portals should be designed in a way that makes it easy and simple to find what you want such as datasets. In the currently existing ODPs/Portals, it is not uncommon to drill down up to 6 – 8 processes before find arriving at the actual datasets of interest to users. This is frustrating and not sustainable.

---

---

**ii) Standard Format of Data:** This is another key challenge area for OD concept and ODPs/Portals. There is currently no uniformity of data formats in most of the ODPs and this situation hinders standard presentation of data across the platforms in the ecosystem.

---

---

**iii) Analytics on Datasets:** Basic analytics or possibility of analysis of data should be available on the portals. But the point here is also about, the quality of datasets presented on the portals in the first case. Most of the dataset now on the ODPs are rubbish with many in pdf format that does not support analysis and another key thing is that are people actually using these datasets, downloading them or linking into the data resources for any purpose. If analytics to make sense of data are not available or even the dataset formats do not allow further manipulation of the datasets, then people would not be interested in using the datasets. So the dataset on the portals should a matter of quality consideration rather than quantity that nobody is interested in, what they don't like or want or unable to use.

---

---

**iv) Licencing:** This is another big challenge in OD practices – the question here – can people actually use them? How easily it is to obtain licences? So the ease of obtaining data publishing licences is a challenge in the industry right now.

---

*ii. From the problems given, which of them do you consider as the fundamental obstacle to the use of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder and also if the same problem has a general industry impact]

---

**Note (1)(a)(ii):**

---

Open data is really not for non-technical people in my opinion. In terms of the portals, we are dealing with raw data and not visualisation. Basically, the portal is place/store for data and the key problems is that the data should be structures, be in formats not just pdfs that nobody can use. So the structure of data including format, metadata, etc. should be provided to enable people use the data. By inference, the current generation of ODPs are still not completely able to deal with the problem of structured data presentation with standard formats including metadata.

---

*b) In terms of understanding the published datasets*

*i. In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

---

**Note (1)(b)(i):**

There are lots of barriers currently affecting making sense of the data published on the ODPs/portals. The most important one is **badly structured datasets/data without proper heading and references** to support user understandability. Furthermore, lack of proper description of the data is another major obstacle to making sense of the data on portals. An example of properly structured dataset is one that, if presented a table, should have the heading for the dataset itself and a heading each for the rows and columns that define the properties or parameters that the figures of data appearing the table represent. In addition, the table should include the URL to the origin of the dataset; and also, **data whether videos or pictures should be presented on the ODP but should be held in linked format with URLs** to their origins. Summarily, Lack of proper data structuring, no lacking to source or other data, use of non-standard data formats, poor data description with little or no metadata are currently problems affecting understability and making sense of data on existing platforms.

---

- ii. *In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

---

**Note (1)(b)(ii):**

Interoperability of platforms (in relation to extensibility, data harvesting, publishing and linking) is in a terrible state. In this first generation (stage) of development of this new OD concept, people are just throwing into the portals or platforms ‘rubbish’ datasets. However, there are some good data on a few portals. The really issue is that there are far too much work left for the data users to do now. On the contrary, there are some portals that are well deigned but several require a lot of drill down activities up to 5, 6, 7 levels to reach the required data causing a waste of time. Thus **Navigation** around the portals to reach the needed data is a major issue as well as **Linking** to the dataset of the portals. The question you face is that – how do you link to the data on the portal? Do you know if the data you are looking for is there on the portal or not, is there any url for it? Do you just download or can you just extend link to it? What do you do? It should be easy to find out what is and where is the url for linking to the dataset on the portal. Simply put, it is not easily found/seen on the portals currently existing, how to link to other portals from a particular portal.

---

- c) Government transparency and accountability enhancement are some of the main goals of ODPs

The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government

transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.

- i. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

---

**Note (1)(c)(i):**

I wouldn't say there is a hindrance in that they don't make things worse; nothing is hindering government transparency as such. But, in terms of enhancing it, yes, there is a lot that can be done, features to improve upon to enhance government transparency. For example if government published data to the public then there is transparency than if you can't find the data. In my opinion, I can't see how not publishing government data or publishing rubbish data or data published not found by users or impossibility of data analysis would make transparency worse of in so far as these situations do not make transparency any better. Similarly the format of data publishing may not necessarily affect transparency of government. For me, it's about baseline thing – from zero. Either support transparency from the point of no transparency or not. Nevertheless, if people provide more open data with better quality, these might improve transparency.

---

- ii. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

---

**Note (1)(c)(ii):**

**'Trust of Open Data':** I think the key thing with OD is that they should be set up data to enable use and re-use (reusability), set up datasets in a way that support frequent updating on a regular basis. As a user of OD, you need to know that the portal contains the up-to-date data (current dataset) for your purpose. ODPs are not usually the sources of open data and if the portals/platforms lack frequent upload of data, users need to know about that fact and also to know about the location of better dataset. It is not proper for data publishing to

---

entertain lying about the source and accuracy of data published. So **data accuracy** is an important issue so do issues relating to the **political or government policy, social and societal** when talking about the concept and practice of OD and use and adoption, performances and promotion of ODPs.

In addition to the above, lack of clarity of the concept and practice of OD, training for skills, and awareness of the public about OD and what it can do has impact on the issues raised in this question. Damage can be done usability, adoption and with the publication of poor datasets on the portals. For example, people are just not familiar with the concept of open data; they don't understand the meaning but if you explain to them that open data means that government should publish more of their activities of governance for citizens to see and know how public money is spent, they become interested. So, awareness again has much role to play but people are now aware of the concept at the moment. Thus by inference, where these present problems in the current generation of ODPs, it is expected that community participation, concept promotion, adoption and performances of the platforms would be low.

The solution certainly entails more training, concerts, seminars, education and use cases to explain the OD and the benefits of its applications such as improvements brought to the society. People would like to see the examples of the good impacts of the practice of open data in the society.

---

---

B) Desired features and their prioritisation:

2. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

---

**Note 2:**

I use local authority data, and in this case I am interested in **visualisation** of map data. People can really **understand** data relating to their community and hope to be able to **supply feedback** to the Public Administrators (PAs) on the demand of the data. A situation whereby people can actually note something happening within their environment and be able to **comment** on the issues, problems or challenges observed; and be able to **share their comments** between themselves (citizens) and with the PAs as well as have the possibility to **track** the progress of the way issues are being resolved or what is being done by the PAs about a particular matter of concern to citizens. Also important is creation of **interaction** and **collaboration** as a way of carrying out or **supporting the new governance** approach and in such a way that you can **measure** or **quantify progress**. The processes of using OD on the portal are important as well as the analysis of data available on the infrastructure are all relevant and important features of ODP.

---

- 
3. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

---

**Note 3:**

The answers to this question drive down to **social media applications** – link to social media networks and the possibility for carrying out **analysis of social media contents** in relation to specific question or problems or societal challenges are desirable features on platforms. **Visualisation** as mentioned earlier along with **layered map** and desirable features in terms of people being able to view specific area of OD, and being able drill down through data should be useful capabilities. So, platform data should be presented in layered forms. Furthermore, availability of a **forum** is useful as a channel for somebody who wants help on questions, for instance, how to use a specific portal or a tool on a portal. In addition using **videos and other multimedia materials** – **pictures and audios** as documentary material or tutorial for explaining how to use data portals and platforms or for explaining the basis of about these infrastructures are very important strategy for encouraging use and adoption of the concept and practice of OD. Again these materials can be used to explain the ‘bounce’ rate of visitors to ODPs/portals so that they can be improved upon looking at the reason why visitors click in and click out without visiting another page.

---

- 
4. What feature(s) of the current platforms would you say is/are performing to your (or users’) expectation – either in general, applicable to all platforms or specific to the platform you use?

---

**Note 4:**

The important tool to recon with here is that most ODPs have well-performing **data harvesting and publishing** tools. Storage capability is another good feature on the portals that actually store datasets as opposed to storing the links to data sources. In terms of better performing platforms as an IT infrastructure, most people would like to use CKAN, although it is hard to say one particular platform is better than the other due to configuration. However, CKAN tends to attract more data users.

---

- 
5. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

---

**Note 5:**

**Data Searchability:** Considering government transparency enhancement, searchability of data is important. However, if not improved upon, searchability will not per se do any damage to transparency. But if people can't find the data that they want, it will not improve transparency situation or status. *In a scenarios whereby lot of datasets have been published but people can't reach the datasets, then the whole aim of publishing has been defeated because people can't find the data in the first place, so is transparency not impacted negatively?* To this scenario question, the interview insisted that lack of searchability doesn't necessarily hinder transparency but where searchability is improved, it may enhance transparency.

---

C) Other features and issues:

6. Given the opportunity, which one area of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens' attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

---

**Note 6:**

This certainly has to do with **Real-time** data which definitely need attentions. I would like to see data being updated on a constant basis with better facilities especially as we move towards Smart Cities. By inference, this relates to the freshness of data like streaming of data as it being generated so it is being uploaded to platforms where users make use of them in real time to make real-time decision for example traffic data and weather data – a situation whereby the more frequent the update the better and more useful is the data.

**Analysis and Visualisation tools and functionalities** are other areas to improve. There are visualisation functionalities at the moment, however, they need improvement to say, **3D style visualisation** capabilities allowing users to view more things visually and easily with **illustrations** that are time-saving. I do believe that **multimedia tools or capabilities** provide good functionalities to express, display and explain things as opposed to descriptions or explanations through the use of texts and tables.

---

7. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens' participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

---

**Note 7:**

---

At the moment, existing ODPs do not support the stated objectives. A few examples of platforms are trying to see how they can help attain the objectives mentioned but these objectives are generally more of theory than practice at this generation of open data concept and practices. There is the need, maybe, to apply use cases to drive down theory to actual practice in the future.

---

---

8. Give your general remark on the technological state-of-the-art of the existing ODPs?

---

**Note 8:**

---

In the case of general remarks on existing open data platforms, I think that it is a great start of the concept and practice of the technology of open data and open data platform and portals. However, we have a long way to go. *On the question of what area of ODPs most disappoint you in terms of features and performances;* I believe realistically, that the technology level is not too poor as it is under evolution. Having said the above, I think the current Route-To-PA project is a brilliant project that aims to bring the concept and practice of OD and ODP to reality. It is a great way to improve the OD practice and adoption.

---

---



5. **The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded Niall O'Brien Date: 27<sup>th</sup> APRIL 2015

55 mins. Spent on this Interview

**5.     The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded \_\_\_\_\_ Date: \_\_\_\_\_



## Raising Open and User-friendly Transparency-Enabling Technologies for Public Administrations



Project number 645860

H2020-INSO-2014

### Evaluation of Existing Open Data Platforms

#### Interview Protocol for Stakeholders

(version 0.2)



WISE&MUNRO





## Abstract

### 1. About the interview:

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date: 24/04/2015

Time: 11am – 12am

Location: Insight Centre for Data Analytics, Lower Dangan, Galway

Name (Interviewer): Arkadiusz Stasiewicz, Mohammad Waqar

### 2. Notes to interviewee:

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 13 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 30 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

### 3. Introduction:

*Question coverage:* Our questions cover 3 major areas:

4. platform challenges
5. desired platform features and priorities of the features, and

6. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:

4. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
5. Platform developers or ODP service providers (midstream community)
6. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

***Peripheral data collection about the interviewee's and his/her company or organisation:***

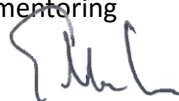
*Name:* Eoin Mac Cuirc

*Company or organisation:* Central Statistics Office, Databank & Dissemination

*Stakeholder group:* mediator/expert/publisher

*Position/designation:* Assistant Principal

*Typical task at work:* data publication, dissemination activities, mentoring



24/4/15

Interviewee's signature (permission to record detail of interview)

\_\_\_\_\_ Date: \_\_\_\_\_

**4. Interview proper:**

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [state stakeholder/user group here] of ODP and/or (if a company/organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

***Interview questions:***

A) Challenges:

9. Challenges associated with the use of open data portals and platforms.

a) Access to open data published or available on the platform:

- ii. *In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[You can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

---

**Note (1)(a)(i):**

i) CSO publishes all the data as the Open Data. They have to make it available in Open Data format. Data is available through the StatBank, JSON-stat API. There is a plan to make it available in the RDF format

ii)

<http://www.cso.ie/en/>

<http://www.cso.ie/en/databases/>

<http://www.cso.ie/webserviceclient/>

iii) The most significant issue is the correct license: right to access the data. When you access the data you need to have it available in the format that you are able to engage with.

- 
- ii. *From the problems given, which of them do you consider as the fundamental obstacle to the use of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder and also if the same problem has a general industry impact]

---

**Note (1)(a)(ii):**

As part of job (data dissemination) CSO look at the users in three main categories: **tourist** - person that doesn't really know about data. portals or IT tools, they just looking for a number. You need to publish your Data to let them find that number. i.e. we're in Galway and they want to Know what is the population of Galway; **farmer** – a research company, Know what is the population of Galway; **farmer** – a research company, insurance company, bank- they want to suck all available data; they have to have the data available in database format / series of data over time. (StatBank, JSON service); i.e. want the data about Galway population over time **miner** – interested in unit record data; they want the data at individual record level.

Note: different users want the data and you have to pull the data in different way. Earlier we had a piece of paper, now interactive tools. You want just a one source of data and make different outputs. Data published in a form of table, PDF, website, RDF, JSON have to be exactly the same data (consistency). The



---

data available at CSO is structured in the same way (population, business) and is consistent among different services. They have a portal and break the data and show you the data that is available through different datasets. i.e. if you look for the data about pigs, you'll have the list of the datasets available (national, international) and Comparable. Different types of users require the data with different tools. The idea of Open Data change the PDF publication into StatBank and more flexible Data outputs and allow customisation for specialised software. Supports machine-to-machine communication.

---

---

b) In terms of understanding the published datasets

iii. *In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

---

**Note (1)(b)(i):**

---

Key issues in general and what CSO is trying to achieve:

The key is the metadata. When you are talking about the metadata you have to be consistent. Everything should be called using the same vocabulary. You need to have the same blocks of data among services. You have to be sure that i.e. given observation is defined by the same methodology behind:

Where the data come from, how was that data generated, what sort of survey was done, what was the sample, what was the questions that were asked. It needs to be consistent among the domains. Not just the CSO but government, public users need to be sure that Cork is Cork, not a county, part of the city etc. It needs to be clear. Clear guidelines need to be published. Once you setup the link to the data it needs to stay there over time.

---

---

iv. *In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

---

**Note (1)(b)(ii):**

---

CSO doesn't have the knowledge what is the best way to put Linked Open Data out there. Seeking for help at Insight. Currently there is support of machine-to-machine communication. Seeking help in OpenCube project. High interests in actual data linking an international level (i.e. Across Europe, international standards) Learning process. How to link different datasets between publishers. Limitation: how to join the data together among publishers?

---

---

- c) Government transparency and accountability enhancement are some of the main goals of ODPs

The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.

- iii. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

---

**Note (1)(c)(i):**

If you are technical person and you understand the data, then you are able to engage with the open government partnership and open data movement, but if you are not in that situation. i.e. you are in one of the departments you go along the department policy and schemas, you may not have the skills in your department to publish the data.

Open Government movement tries to change that with clear guidance. Is no support and standards it is difficult to start the publication process Governments need to engage with public to create clear vision and rules. Which data you want and which data shall we publish. High data quality required. To link the data you need to publish using same identifiers and correct structure. Data expert needs to manage it correctly. There is no point to publish the data that is rubbish.

**'Eoin' between the datasets can occur as:** Mr Eoin, E., Owen, Eoin which doesn't allow to link the data properly.

**To make it transparent:** if you have the data and the data is confidential to make it transparent you still need a set of metadata that states it, shows that it is available and who to contact for more details.

**Put metadata that is open:** this is what we have, this is the license, this is accessibility requirements etc.

---

- iv. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

---

**Note (1)(c)(ii):**

---

The CSO puts all the statistical data at StatCentral. Eoin is visiting schools, universities and gives a lot of talks. The questions who is aware that there is a portal cso.ie? Who is aware that statcentral.ie exists? Who is aware that data.gv.ie exist? The most people doesn't know that the data is available for free! All have access to it in open format. Ignorance? Education? Fundamental action is to make people aware that the data is out there, where can they access it and how do they access it in easy way. (it have to be really simple!) Is it better to have the data quicker and not as accurate or wait with the publication while the data is more accurate? The argument: the data can be cleared later. Finals will come later. Good approach as long as the user is aware of it.

PROMOTION make less licenses (sort them out) – which license should we use to make our data as open data? there's a lot of confusion! I want to publish = I want to allow other to re-use. Even commercial. The language used for legal is way to difficult! should be: please use the data or please add note about the data source Where to publish? There should be centralised space. You can't find a bit here, a bit there and it doesn't help!

---

B) Desired features and their prioritisation:

10. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

---

**Note 2:**

---

Most important: how the user can find the data.

1. How the data is structured? Is it clear structure?
2. On that structure – is there a clear description what is the structure = Metadata. Easy search engine requirement is most important.

In statistics there are clear classifications, which make it easier. User need to know: How accurate is that figure? How old is this figure? All counted or sampled? Aggregation was made?

---

11. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

---

**Note 3:**

CSO is using PC-Axis – all statistical offices follow Nordic model. It is not expensive to use and they have the support – how to use. There is a working group meeting once per year. General discussion.

There is a need to have it for Open Data movement: this is the tools, those are the people, there is the platforms, this is the endpoint, this is the search engine you should use.

General manual / handbook needed. Intuitional tools are required. All you need is a video shows how to use the tools / structure. Right metadata should be used. How can we convert our data? What tools should be used? How do we update i.e. the classifications? How to link data with users if they are using different URIs? Central management required.

---

---

12. What feature(s) of the current platforms would you say is/are performing to your (or users') expectation – either in general, applicable to all platforms or specific to the platform you use?

---

**Note 4:**

In relation to Open Data CSO is happy of their achievements:

Clear structure, unified form, follows Eurostat patterns. What if you do not have standard type of data? Video, recordings – how to make it available? I.e. National museum – how to publish their objects? How is the data around these objects available? Same story with castles, churches, heritage – how do we know about Data about physical landscapes, rivers, mountains, shipwrecks, boundaries. Who is managing all those standards? How to publish that kind of complicated data?

Different organizations that are in charge – how to make sure that they know what they are doing.

---

---

13. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

---

**Note 5:**

The most important thing is to get the data published and to make it easy to publish by someone in open data format. Any tools that make it happens would be desired. Some kind of audits / rulebook / education would be very helpful:

This is where you are, this is the data that you have now and this is the way You shod follow to make your data an Open Data. Step by step process with video tutorials and tools description as well as User-friendly description of licenses to follow.

---

C) Other features and issues:

14. Given the opportunity, which one areas of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens’ attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

---

**Note 6:**

Waterfall approach. Tim Berners-Lee 5 stars, start with any data online (single star) and climb up To five stars <http://5stardata.info/>. Some publishers might be fine to stay with i.e. 2 or 3 stars and that’s fine. Sometimes tools are there – i.e. spread sheets - and users are fine with it. It would be great to have roadmap for achieving 5 stars.

---

15. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens’ participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

---

**Note 7:**

It depends on the area of data that you are looking at, i.e. Ireland – do we have the open maps available? High quality maps of the cities are created but not available as open data with open access. There is political will and commitment to address it. How long will it take? What pieces of data are important for open data movement to take it forward. Decision making – tries to engage communities, but how do you know who is interested in open data? who to ask for opinion? all wives? all kids? How to actively engage with the people? The priority is to engage those who are not using the data at the moment. Why are they doesn’t they use it and the available tools?

---

16. Give your general remark on the technological state-of-the-art of the existing ODPs?

---

**Note 8:**

We are in the learning phase – work in progress. How to do it? What is the best way to do it? Personal level, organization level, national / international level. Idea is smart, but how to connect all the dots?

---

---

5. **The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.



24/4/15

Interviewee's signature against responses recorded

\_\_\_\_\_ Date: \_\_\_\_\_



Raising Open and User-friendly  
Transparency-Enabling Technologies for  
Public Administrations



Project number 645860  
H2020-INSO-2014

**Evaluation of Existing Open Data Platforms**  
**Interview Protocol for Stakeholders**



WISE & MUNRO



**1. About the interview:**

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date \_27/April/2015

Time \_06:00 pm

Location \_Insight Centre for Data Analytics, NUI Galway

Name (Interviewer) \_Ed. Osagie and Waqar Mohammed

**2. Notes to interviewee:**

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 13 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 30 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

**3. Introduction:**

*Question coverage:* Our questions cover 3 major areas:

1. platform challenges
2. desired platform features and priorities of the features, and
3. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:



1. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
2. Platform developers or ODP service providers (midstream community)
3. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

***Peripheral data collection about the interviewee's and his/her company or organisation:***

Name **\_Ed Curry**

Company or organisation **\_Insight Centre for Data Analytics, NUI Galway\_**

Stakeholder group **\_Academic/Researcher/Promoter of Open Data; Involved application of Open Data (OD)**  
Institute node in Insight

Position/designation **\_Research Unit Head – Green And Sustainable IT**

Typical task at work **\_Investigations of the design of IT systems, design of software systems, how emerging technologies are designed.**

**4. Interview proper:**

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [Researcher & promoter] of ODP and/or (if a company/organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

***Interview questions:***

(A) Challenges:

1. Challenges associated with the use of open data portals and platforms.
  - a) Access to open data published or available on the platform:
    - iii. *In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[You can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

**Note (1)(a)(i): Talking about challenges**

**i) Data format:** The challenges of dealing with the export of data which assumes a certain level of expertise to export and use data

**ii) Using multiple pieces of data:** Other challenges centre around using multiple pieces of open data from different datasets, e.g. two entries in a data catalog from two different organisations, how to integrate them – the skills and technical knowhow to manage the data together.

**iii) Technical knowhow:** Needed to manage datasets from many sources, to integrate them together are not available, not many people don't have these skills.

**iv) Cataloguing:** Also data cataloguing itself still pose challenges to accessibility of data on OD platforms

- 
- ii. From the problems given, which of them do you consider as the fundamental obstacle to the use of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder and also if the same problem has a general industry impact]

**Note (1)(a)(ii):**

The fundamental challenge facing the use of open data on the portals is that they have not reached the level of sophistication to provide users with answers to their questions. So provide more sophisticated platforms that require less technical knowledge for users to use them. The ability to discover all the datasets for use by a user is difficult to come by. Searchability and discoverability have impact in the use of open data; and although related they are quite different in meaning. While searchability assume you know what you're looking for, discoverability does not necessarily imply knowing what you are want. However, both combine to pose a problem for data accessibility.

- 
- b) In terms of understanding the published datasets*

- v. In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

**Note (1)(b)(i):**

Again, the skill to make good use of the data is what is important. The more raw the data is, the more skills are required to make sense of it or use it. The more usability of the services provided by the portal the better

the usability of the data. Lack of skills hinder making sense of the data provided on the portal. Furthermore, the level of metadata provided affect usability and understandability. Thus more metadata entails better understandability and better it is to make sense of it. So even for technically-oriented person, if the metadata is not available, the person still can't use the data. in addition, how well the data is linked, quality, normalisation, veracity of the data, linked to cross dataset is issues that impact on the making sense of a datasets.

- vi. *In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

**Note (1)(b)(ii):**

To this question the interviewee opted to give a rather more general opinion based on his knowledge of CKAN. Interoperability issue relates a lot to the data format/metadata available in the dataset. So interoperability depends more on the actual dataset rather than the platform itself. However, the actual catalog services of the platform does not necessarily support the integration of data but tags, occurrence of tags across different catalogs do not seem to consider entity management across platform. From my point experience of data catalogs, there are challenges to work across datasets. Summarily, there are insufficient metadata, tagging, cataloguing to support interoperability. Having enough metadata though is supportive but not enough alone to support complete interoperability. It would appear that a lot of the job is left for users to figure out on the metadata.

- c) Government transparency and accountability enhancement are some of the main goals of ODPs

The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.

- v. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

**Note (1)(C)(i):**

On the issues of feature that currently enhance or hinder government transparency and accountability, the interviewee agrees that there are features to talk about. In the first place, the idea of having the open data platforms and portals are good initiatives that tend to create awareness on the concept of openness of government and hence transparency because the culture of awareness of the concept is bringing the picture of transparency as an important issue.

I don't think (personally) that there are at this stage specific features on the platform level that might be hindering transparency. However, the policy which pushes the government to decide what become open and transparent or not might be the issue in this case affecting transparency. By default, what the government uses to decide availability of data is a point in question. When presented with the scenario that assumes datasets are published on platforms but not relevant to anybody, not understandable or usable by the citizens; and again supposed datasets are provided in such a way that users can 'play' with them, manipulate them, marsh them in a way that they may be able to understand the underlying fact facts better, share among other users, comment on them, will this make issues better understandable and increase transparency of the activities of the government? Respondent agrees yes, that point of view is understandable. However, looking at it from the government perspective, they don't really want to encourage people to look at the data, they may prefer to release data not readily understandable, difficult to use and that is quite common with them. But whether trying to up the quality of data or social collaboration on data will actually enhance transparency is hard to point out. But, well, collaboration on dataset may enable some to see more about the actual activities of the government yet here has to be more tools on the platforms to enable such collaboration instead of the platforms just holding data alone.

- 
- vi. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

**Note (1)(C)(ii):**

In terms of the performances of OD platforms, the interviewee suggests the following hindrances:

**i) Licensing:** of OD is an issue as well as the number of existing licences in use. The Irish environment has a restrictive licence on open data living the people debating on which data is open data and which is not. I

think with some OD entrepreneurs wants to start creating businesses and apps on OD but they are worried about sustainability. If the data is release today, what about updates and will it be release tomorrow? So business people don't want to create businesses on data that may not be available going forward. So sustainability is very important for the future and this issue should become part of the contract between portal managers and data suppliers. The incentives at the moment to jump-start the ecosystem of open data are trivia; the hackathons incentives are nice but they are left at a point without being finalised.

**ii) Data Relevance:** Many data are not commercially viable though many support transparency and are good for mainly journalist to engage with thus the commercial relevance is being overplayed.

**iii) Data Privacy:** Privacy is an aspect that is relevant to platform performance because it is also being overplayed people who are always concerned about privacy. However, as long as sensible measures [to ensure data privacy] are in place, that should be enough. From the government point of view, there is a lot of resistance internally for them to release the data. So, I think, incentives are needed to enhance data release, and data and platform usability.

iv) To the question of current platforms lacking the capability to significantly support the re-usability of data, the respondent agrees and states that the main problem is that platform/[portals] are merely catalogs, the common fellow in the street would not understand the availability of government data on the portals. On the other hand, the fellow may be better off with an app that enables him/her to view or see the government data [say in a portable device]. In this sense, portals are not so relevant to the common users of OD but applications that enable them to see the information derived from the dataset. Thus platforms need to supply the apps for the need of the common citizens, users of the open data.

---

(B) Desired features and their prioritisation:

2. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

**Note 2:**

When asked about the most important features of open data platforms, interviewee refers to the his position as a stakeholder involved not really in the use of the platforms as a data consumer but does engage with the platforms and portal in order to study how the systems work. However, from experiences and observations, a portal is only a part of the solution which you may think of as a catalog where you can dump your data. but there is a lot of infrastructure installed in place to make the data available on the platform and to consume the data. What is missing in this system of data in, data out of the platform, is the support system of infrastructure for the movement of data in the platforms or portals. The platforms are unclear on how to use them in publishing or consuming open data. There is no **Best Practice** on how to use the open data, for example, as an entrepreneur; it is not clear where to start from in consuming or building a business on OD. A lot of best practices which should become part of the new features on the platform should be provided instead

of leaving users and entrepreneurs to figure out what to do. Best practice and guides should be provided for users.

---

3. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

**Note 3:**

There is really no current platform that can do all or any more benefit or features than are listed in this question. However, social collaboration is important, and if looking at the example of other communities such as Apache', which is very simple and a strong governance – these are really not technical matters rather instead social arrangements that are very basic and which are currently lacking on OD platform ecosystem. So far government releases data with on the platforms and that is it. I think that government can provide more services or features on the platform but this issue of additional features for better performances is a social thing that government should not be involved in because it centres on the transparency matter and hence should be areas for the public concern. So anything that will enable better governance of the platform and the social collaboration on the platform will be useful in enhancing better use of the platforms.

---

4. What feature(s) of the current platforms would you say is/are performing to your (or users') expectation – either in general, applicable to all platforms or specific to the platform you use?

**Note 4:**

To the interviewee, certain features on current platforms are performing to impressive level. Most important of these is the cataloguing.

i) **Cataloguing** services provided by majority of platforms support multiple data formats, support some metadata and file conversion operations.

ii) CKAN tries to have some additional **plug-ins** embedded in its features and **Extensibility** for CKAN is to some extent positive as well. Cataloguing for CKAN is quite good with a strong community.

When asked to relate performance to data transparency/accountability support on platform, interviewee could not clearly draw a point on current platform that support transparency as a matter of definite attempt at doing so. However, as an open source system, CKAN could be considered supportive of transparency but he adds quickly that he is not quite sure, actual portal implementation for the support of transparency exists.

---

5. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

**Note 5:**

Entity management stands out clearly as one of the capabilities because the management of the various entities that exist in the dataset need improvement so do data usability, searchability and licensing. On the issue of platform governance, he agrees that there is the need to know what data is available from government areas, the quality of service the government has signed up to on the portal e.g. to understand matters surrounding data supply, updating and quality thereof. Secondly, the tools that platforms provide for usability and searchability of data are too basic and are not necessarily providing answers to users' questions especially the non-technically-oriented users. Therefore more effort needs to be made to improve on the platforms tools in a way that users can get answers for their questions.

---

(C) Other features and issues:

6. Given the opportunity, which one area of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens' attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

**Note 6:**

Considering areas of OD platforms that deserve changes, the respondent would like to see changes in the area of quality data service. So data providers should ensure supply of quality data which may support transparency; and as recommendation to data providers, I think they should provide a rating system for datasets in order to enhance the quality of the dataset and further improve trust.

- 
7. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens' participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

**Note 7:**

The interviewee surely believes that current open data platforms do not clearly support the benefits mentioned in the question in his opinion and that the reason for that is not because of the problems of the lack of platform/portals features but because of the exaggerated expectations of the benefits of open data concept.

There is very little open data can do to achieve these benefits because they are very difficult to achieve. Getting people involved in use of platforms is not necessarily platform problems but civic engagement problems. For example, to get certain decision made in the government you have to get input from the civic environment which is not part of the open data concept.

Basically, there is a divide between what the promoters say OD can offer and what is actually achievable through open data practices on the portals or platforms. The interviewee disagrees with this assertion as that is not the case. When people say what OD can do all of the services, they fail to realise that open data concept is not just the platform alone, there are many other aspects to be put in place for open data to achieve some of these benefits. The non-technical but very important issues such as the policy of publishing licensing, quality of data, etc. are all relevant parts

To achieve some of these benefits call for actions that will involve a wider ecosystems – the platforms, civic engagement and government supply of data. Unfortunately, this is not very easy to attain but the biggest challenge is getting people to play their roles, engage the system and get government to buy in. Summarily, the basic thing is to understand what the platform can do and what it cannot, motivate the citizens who are interested in transparency yet hesitate to go to the portals/platforms to use the data provided.

---

8. Give your general remark on the technological state-of-the-art of the existing ODPs?

**Note 8:**

In general, the data management is complex and costly. Companies are willing to participate in open data concept with high expectation but require lot of funds and managing the data is not that easy. The expectation of the user to see platform assume Google style searching capability is not comparable with the technology of existing open data platforms. The data integration problem is not something you can fix very easily. In the case of the next generation platforms, efforts should be geared towards great improvements in entity management, natural language interfaces. Moving away from just a portal infrastructure to a more user-friendly interface for user interaction and the portals need to be able to support variety of applications, support services for developers to develop applications such as adopting an App store for open data ecosystem.

---

**5. The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.



*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded

\_\_\_\_\_ Date: \_\_\_\_\_



## Raising Open and User-friendly Transparency-Enabling Technologies for Public Administrations



Project number 645860

H2020-INSO-2014

## Evaluation of Existing Open Data Platforms Interview Protocol for Stakeholders

(Draft, version 0.2, 21042015)



WISE & MUNRO





1. **About the interview:**

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date: 14 May 2015

Time: 2:15 PM IST

Location: Skype call

Name (Interviewer): Edo Osagie and Arkadiusz Stasiewicz

2. **Notes to interviewee:**

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 19 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 50 - 60 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

### 3. Introduction:

*Question coverage:* Our questions cover 3 major areas:

1. platform challenges
2. desired platform features and priorities of the features, and
3. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:

1. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
2. Platform developers or ODP service providers (midstream community)
3. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

#### ***Peripheral data collection about the interviewee's and his/her company or organisation:***

Name: Bill Roberts

Company or organisation: Swirl IT Limited

Stakeholder group: Mediator

Position/designation: CEO

Typical task at work: Develop OD platform PublishMyData, provide technology for customers e.g. government, local government etc.

Interviewee's signature (permission to record detail of interview)

\_\_\_\_\_ Date: \_\_\_\_\_

#### 4. Interview proper:

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [state stakeholder/user group here] of ODP and/or (if a company/organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

#### ***Interview questions:***

##### A) Challenges:

1. Challenges associated with the use of open data portals and platforms.
  - a) Access to open data published or available on the platform:
    - i. *In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[If you can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

#### **Note (1)(a)(i):**

i) **Meaning of Data:** Once you find some data on OD platforms it is hard to determine the meaning of data, because most common way of publishing open data is to publish it as CSV on to the web in a data catalogue and usually with fairly minimal documentation, so you can download that, and you can see table of numbers and with some column headings, often you can guess what the headings means but you can't always be certain, also you don't know enough about how the data was collected.

ii) **Trustworthiness and Quality:** Often we don't know how trustworthy the data is and what process was used to create it and how well maintained is that process is, so these are kind of meta-data and quality issues.

iii) **Data Integration:** Its unusual to find all the data you want use is in one file, if you are downloading dataset for somewhere then it maybe that you need to download several other related datasets and then you are left with job of trying to join them together without consistent identifiers for things, so obviously that's the one of reasons behind our interest in using linked data as technology. It's often a problem for a user is that if you want to collect multiple datasets on a topic then its often not described consistently so you are left with big task of data integration.

- 
- ii. *From the problems given, which of them do you consider as the main obstacle to the access, use and reuse of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder and also if the same problem has a general industry impact]

**Note (1)(a)(ii):**

Probably, the issue of meaning and structure of data not been clearly defined, if you want use that data you may have to make your own judgements about what it means and that puts lot responsibility on data consumer to try to work it all out.

- 
- b) In terms of understanding the published datasets

- i. *In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

**Note (1)(b)(i):**

Lack of documentation and meta-data, we are linked data specialist so linked data is good technology for that, but the main problem is that data publishers needs to provide to good documentation and need to process the data.

- ii. *In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

**Note (1)(b)(ii):**

There are two levels of interoperability, one is that the dataset catalogue level that is perfect but not too bad, most data catalogues includes similar metadata, probably will be using DCAT vocabulary, so interoperability of catalogues is not perfect but is not too bad. Interoperability of data representation itself is much less good, is generally poor, again what we are trying achieve with linked data approach is to make interoperability of data itself better.

- 
- c) Government transparency and accountability enhancement are some of the main goals of ODPs. The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.
- i. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

**Note (1)(c)(i):**

For government transparency, the purpose for open data is disseminate information to people about what government is doing, yes certainly open data platforms have been helpful in this process, but



the challenge is to make open platform better in getting information to the people who want it in a form that they can understand and use. But there also some questions related to standardizations and education of the data publisher to get them to release information in useful form.

- 
- ii. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

**Note (1)(c)(ii):**

Getting the data owner to give it enough priority I suppose is a lot of the problem, the technical aspects of the platforms are not perfect and the biggest issue is get data owner to understand what they can achieve by doing it, people tend to not really understand the process and they want to do as cheaply as possible, fair enough, they want to get good value for money but I think its difficult to make business case for open data platforms and showing what its benefits are, Its quite difficult to measure and quantify the benefits of open data, a clear explanations of expected benefits to data owners so that they may publish more data, that is a challenge for sure. Platforms are not prefect at the moment, especially data integration support, the way data is represented or the approaches to put some effort to represent data clearly using linked data to clearly define meaning and structure of the data, that takes extra effort for the data publisher, the challenge is to provide some tools to make that process simpler and more reliable for people who are not technical experts.

---

B) Desired features and their prioritisation:

- 2. What is your interest in open government data? Why are you interested in using OD/OGD?  
To what extent is your interest (or information) currently available through open government data (OGD) platforms?

**Note 2:**

I think am not right person to be asked this question, because our role is of intermediary in publishing process, generally geographic reference data, transport, health, demographic and economic data interests lots of people. UK open data users group website have request mechanism where people can ask for data that is not currently available, so there are some examples what

people are looking for. One of the most popular one we worked with was the index of multiple deprivation in the UK which dataset about deprivation and relative levels of poverty and so on, very widely used.

3. Is there information which is produced but not currently disclosed that you would like to have access to? What about information which is not currently collected or generated?

**Note 3:**

In the UK it is mainly the question about licencing actually, the issue that comes very commonly is that there is information available around geography and addressing but its not currently openly licenced so it can't be used in open data, so its only available for commercial licencing.

There is lot of useful data available from ordnance survey and royal mail, but some of this data is not open data and that has limitation on what other people are able to publish. In particular there is no open addressing dataset which links individual addresses to locations.

The most examples I know of are ones where data exists it just not easily accessible, so there may well be other types of data that is not collected that would be useful but that are not the most important problems for me at moment.

4. If the information that you want is not available, what kinds of mechanisms might you currently use to request this data?

**Note 4:**

These are usually a combination policy and cost from the organization that owns them, it comes back to clearly explaining the value that would have been created if the information was more widely used, the problem with ordnance survey is that currently get some of their income from licensing

this data to be bought, so if they give it away everyone than they won't that income, so then that had to come from another perhaps taxation or something. To make the case the new value created from doing is more than then lost income and more than the cost of maintaining it to high standard because that's obviously one of the issues if you want quality data, it costs money to prepare it and somebody has to pay for that. So it's about trying to measure the value created by open data.

5. Are there other mechanisms that you'd like to see or which you think would help users of government information to get the data they need?

**Note 5:**

There some good thing happening in that area, so one is helping the citizen to know the things that are happening in other countries, so everybody can learn from successes of other places, things like open government partnership is doing a good job over there and also some of the campaigning organization such Open Knowledge Foundation is doing good job. Open Data Institute is doing good and they making relationships lots of other countries, just generally raising awareness about the best practices theses areas, no country is doing a perfect job everyone can learn from other people, certainly there are some countries that are identified as leaders in field so it would be good for people in countries that are doing less in open data to learn from ones that doing more and then they can demand from their government well look this other country is making all this stuff available why don't you. I think the pressure from citizens and business would make it more

6. When you use government information, what kinds of other data would you typically combine it with?

**Note 6:**

On thing is combing geographical and statistical data that's a very common example, you got statistics about some place but you want to know how different places relate to each other. Common thing is combine information about same place that's produced by different organizations, maybe you want know information about you city some of that will come from city government and some of it come from national government, so you want to get best picture of what is happening in your city around economic, crime, health or other situation that information will come from different organizations.

7. Would it be useful to you if users of OGD platform could share enhanced or transformed data with other users?

**Note 7:**

Sure there will examples where that would be useful such correcting mistakes, filling the gaps in datasets etc.

8. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

**Note 8:**

**API:** It is important for data to be available through as an API as well as by download; because there is some kind of data where downloading data isn't good idea either because the data is too big or it changes too frequently. Moreover if you want to get data from different dataset and want to combine together then getting data through API gets important. Once you get the data than you need some sort of standard representation of it, so standard representation for data is next important thing.

**Standard Representation:** It kind of technical so most users are do not thinking why I need an API and standard representation but in fact that enables all kind of other thing to happen that are currently difficult.

9. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

**Note 9:**

Some of my pervious answers are relevant here that better API and standard representation is important. Better tools for owners of data to process it, clean it and transform is important, because there is always a lot of work to do to tidy up the data before you publish it, you want to be able to automate some of that and to make it repeatable and offer provide some guidance to data owner to help them in represent the data in good way that could be programmed in to tools. An interesting

application area that kind of goes beyond open data is to combining open data with personal data for particular purpose is very useful, that obviously not really a feature of open data platform because the personal data will not be in the open data platform but I suppose kind of management of add test control to particular data collection people can efficiently use their own data together with open data from outside.

Mostly collaboration is social issue. Making open data platforms more user friendly so to make it easy for users for find the data they want and make it easy for them to understand it some of that is user interface design and visualizations, and also make it easy get the data out, so if you make it easy for people more people will do it. Lots of attention to detailed usability of the software helps. Most of the communication and sharing are social issues really so it not about software it about how you get people interested in getting together and helping them to achieve something.

10. What feature(s) of the current platforms would you say is/are performing to your (or users') expectation – either in general, applicable to all platforms or specific to the platform you use?

**Note 10:**

Its difficult to answer it in general way really, I don't have anything useful to say, you can go to an individual platform and say I like that and but I don't like that, it only possible give a very detailed answer.

11. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

**Note 11:**

I think I pretty much answered this question already, I will be just repeating my self.

C) Other features and issues:

12. Given the opportunity, which one area of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens’ attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

**Note 12:**

I would like more people to use linked data because I think that’s is good solution to problem we were talking about standardizing ways of representing data, so getting more people to produce more quality data to maintain it properly and use standards to representing it. Governments needs to published more data, even in places like UK where there is quite a lot of open data there still more information that would be useful and addressing the quality of data at least for some particular useful datasets that are commonly used as reference data you want to make sure that is well maintained, clearly available and with consistent identifiers for things that other people can use.

13. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens’ participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

**Note 13:**

Most the platforms are little bit one way they are about broadcasting rather than encouraging and gathering contributions so we could probably add some features to platforms to encourage people to participate. There are different functions for different kind of platforms sometimes it useful for something to be source of official information so in that case you don’t want user generated content on the site but there are other kind of platforms where you do want to encourage people to contribute so its kind of depends on what kind of data and what sort purpose you would like to

achieve. A lot of time those mechanisms might exist but it's kind of a social issue for getting people to use it. Why for citizens to systematically contribute information is something that in most open data platforms but could be very useful.

A lot of government data effectively contributed by citizens already because it comes from various kind of surveys where people go around and systematically ask people list of questions and get the answers and then collect the results so that is what essentially a census is, it's user generated content that being heavily processed by the government so that already happens, that's the way large proportion of government data is created. The challenge with this process is that it takes a long time and there could be technology-based approaches which allow gathering of information from people more frequently and in more fine-grained way, I think challenges there are mostly process and statistical ones, working out if you have got good sample of people is representative either biases in your sample because for whatever reasons some people may have better access to technology and more inclined towards providing information that gives you bias in your information which is difficult to understand when you are analyzing it. There is not simple cure to this, it's deep problem how to systematically deal with contributed information to provide something that would be useful.

14. Give your general remark on the technological state-of-the-art of the existing ODPs?

**Note 14:**

It's long way to go, handling bigger volumes of data, better user interface design, better API and data standards. There is quite a lot to be done in user interface and interaction design to make it easier for people to find information and that a lot underline technology to drive it. I would be quite nice have Google style interface where just type your question in and get the answer but that answer kind of being driven not just by text index instead by the numbers that are available, that would be very useful. Letting users to manipulate information in their own tools, making easy for people to build their own visual ways of looking at data but a lot of it is user interaction design.

---

**5. The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded

\_\_\_\_\_ Date: \_\_\_\_\_





## Raising Open and User-friendly Transparency-Enabling Technologies for Public Administrations



Project number 645860

H2020-INSO-2014

## Evaluation of Existing Open Data Platforms Interview Protocol for Stakeholders

(Draft, version 0.2, 21042015)



WISE & MUNRO





### 1. **About the interview:**

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date: 18 May 2015

Time: 1:00 PM IST

Location: Skype call

Name (Interviewer): Edo Osagie and Arkadiusz Stasiewicz

### 2. **Notes to interviewee:**

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 19 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 50 - 60 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

### 3. **Introduction:**

*Question coverage:* Our questions cover 3 major areas:

7. platform challenges
8. desired platform features and priorities of the features, and
9. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:

7. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
8. Platform developers or ODP service providers (midstream community)
9. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

***Peripheral data collection about the interviewee's and his/her company or organisation:***

Name: Paul Hermans

Company or organisation: ProXML bvba

Stakeholder group: Mediator

Position/designation: CEO

Typical task at work: Analysis and Development of data transformation pipeline

Interviewee's signature (permission to record detail of interview) \_\_\_\_\_ Date: \_\_\_\_\_

**4. Interview proper:**

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [state stakeholder/user group here] of ODP and/or (if a company/organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

**Interview questions:**

A) Challenges:

17. Challenges associated with the use of open data portals and platforms.

a) Access to open data published or available on the platform:

*iv. In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[If you can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

**Note (1)(a)(i):**

First, I need to say that I have only experience with two open data platforms that CKAN and DataTank.

**Poor Usability:** I still find it for users very difficult to access data in these environments; there are issue with user interface, it is hard to find and search for data. . It is very difficult to have a look at the data, because the visualizations are underemployed in most of the open data portals that I know. It doesn't need to be high-level analytics, it just needs to show fragment of data in nice table or as basic visualization in a chart. For example if you look at Flemish government data portal there are lots excel sheets over there but excel sheets are very difficult to read, even not visualizable as table in CKAN, those are major issue to make an assessment about what the data is about and how it looks like.

**Documentation:** Very hard to know what is data about, missing or limited metadata, no information about the provenance where does the data comes from and how it is collected, no human explanation on data model and fields used.

*ii. From the problems given, which of them do you consider as the main obstacle to the access, use and reuse of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder an also if the same problem has a general industry impact]

**Note (1)(a)(ii):**

It depends little bit on type of user. For general public the main issues are the ones I have mentioned already, I would also like to have storytelling for non-technical users. On the other hand for developers its sometimes difficult because there are no standardized API to work with, there should be more emphasis on standardized API.

---

b) In terms of understanding the published datasets

vii. *In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

**Note (1)(b)(i):**

That's similar to what I have already said. Very difficult to see the data in tubular form or in simple graphic, you don't have any explanation about the data model or meaning of the fields, there is no background where the data came from and where it being used for.

viii. *In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

**Note (1)(b)(ii):**

That's the subject that we study very well in Belgium, because in Belgium we have very complex structure of the state, there are regions at federal level, provinces, big cities and so on, all of them have different systems, for example the larger cities in Belgium are working with DataTank, the Flemish government level has CKAN integration with DataTank, at the federal level its a custom Drupal development, we tried to solve interoperability at catalogue level using DCAT, at moment there is no support for interoperability at data level. Interoperability at data level is big challenge because every source has its own way of defining classes, tables, property and fields and use different serialization for exchange, so its very hard.

---

c) Government transparency and accountability enhancement are some of the main goals of ODPs. The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.

vii. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

**Note (1)(c)(i):**

Yes, in fact there two level there. You can have transparent data publishing without using open data portals; the open data portal is just repository to get open data published elsewhere at least this the case in in Flemish, but the main problem according to me is that if you look into the publish data you can't judge them good enough or understand them due to lack of provenance or lack of documentation, so the data isn't clear enough to make judgement so its not really transparent because you don't understand the data. The quality generally affects the quality. Sometimes data is good but you don't understand the data as you lack surrounding information. Analytics and human explanation of data can play role in enhancing transparency.

---

viii. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

**Note (1)(c)(ii):**

It depends on which user group; if you are looking to developers I have the impression that developers know about open data and they find their way to open data portal to find relevant datasets that they can use to build a useful applications, I won't say every developer knows that but that amount is growing. For the general public though open data is not yet on their horizon. To be honest, I don't have any idea what to do about it.

---

B) Desired features and their prioritisation:

18. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

**Note 2:**

As mediator in publishing process, the API offered by the platforms is very important to us to automate the process of data publishing.

19. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

**Note 3:**

I don't see it directly, most of the time the first thing for getting data in the open is that you change the existing processes and procedures to collect and publish data, publishing it as open data should become part of those processes. That's the first thing, off course if you add some collaboration and editing facilities in the open data portals it will help but first thing you need to address is that open data publishing becomes part of your traditional workflow. For some very specific datasets that are very close and relevant to normal users for example the place of bus stops, that people are checking in their own neighbourhood and are able to find errors and communicate those errors but as it was until now its being address as every dataset has a contact person and people can send their remarks to that contact person. I know that iMinds at University of Gent are working on trying to offer an interface to enable collaboration at the data level within the dataset itself, but I haven't seen it yet, I don't know how easy or how difficult it is to use for a normal user. My impression is that whatever we have already at moment is sufficient for most of the use cases.

20. What feature(s) of the current platforms would you say is/are performing to your (or users') expectation – either in general, applicable to all platforms or specific to the platform you use?



**Note 4:**

From my own point of view I don't have any complaints and the reason is that we are in the business of publishing open data and I have to say that CKAN API and also the DataTank API are sufficient for what we need do, so that's good about open data platforms, the quality and richness of the API is good.

On the API level I haven't encounter any needs that are not addressed in context of uploading datasets and adding and managing metadata on the dataset. Once again this very narrow view for rest I have already explained, I have lot of remarks on the usability of open data platforms.

21. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

**Note 5:**

This relates to things previously said. There must be more explanation around the datasets, related to dataset itself, data structure, data model, the provenance and some other background information that is useful to know. The ability to have quick look at data in very user-friendly way, the ability to have some visualization in automatic way, as for the moment in CKAN you can build yourself some graphs but you need to choose by yourself x-axis and y-axis, which is way difficult for a normal person, so there must be some intelligence that by looking at dataset the open data portal is intelligent enough to build automatically most easy to understand visualization.

C) Other features and issues:

22. Given the opportunity, which one area of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens' attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

**Note 6:**

What is still problem I think is that open data movement is being followed by people who already convinced with its usefulness. It may sound very cynical for lot of people the interest in open data is for keeping their own jobs and the only thing that counts is how many dataset are published, and what is done with those data and how much of success it is an instrument for as many as of target groups is of less concern, its not the case for all but true of a majority. The bridge that needs to be crossed to make it that everyone has interest and find it useful tool is still very wide. I don't have any solution for that, the only that I would like to test out by myself to investigation the option of data journalism, if we are able to succeed in writing very interesting stories using open data that attracts attention of all sorts of people to open data that can raise the interest open data, interesting stories can make open more visible and know in the world.

23. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens' participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

**Note 7:**

In principal and in theory that's all correct, but I am afraid I haven't seen much of it in realty and I can't explain why. One of the thing is that open data portals are not user friendly enough but there is still the bridge to be build between the few that are aware and doing interesting things with it and larger audience, larger group of citizens and interested persons. It comes back to about not having enough storytelling about data on the open data portals but if you want short answer I consider it very low.

24. Give your general remark on the technological state-of-the-art of the existing ODPs?

**Note 8:**

The APIs are done well but can be made better. The current generation of platforms lack intelligence, the platforms should for example be able find automatically related datasets, at the moment if you want to create a relationship between datasets you have to put it by yourself, it would be good if system could detect these things automatically. The same for visualizations as well, the platform should be able propose automatically best visualizations; better natural language processing like better keyword detection would be nice to have. These are the few things that are missing.

---

**5.     The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded \_\_\_\_\_ Date: \_\_\_\_\_



## **Evaluation of Existing Open Data Platforms**

### **Interview Protocol for Stakeholders**



WISE&MUNRO



**1. About the interview:**

**Project:** Route-To-PA Project: Work package 2, Deliverable 2.1, Task 2.1: “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”

Date \_21/May/2015

Time \_10:15 am

Location \_Insight Centre for Data Analytics, NUI Galway

Name (Interviewer) \_Ed. Osagie and Waqar Mohammed

**2. Notes to interviewee:**

First, I would like to thank you for your participation. I believe your input will be valuable to this research that aims to identify salient issues to consider in developing next generation open data platforms.

The interview process starts now.

Confidentiality of data/information collected in this interview is guaranteed. The data/information gathered will be used for the purpose of Route-To-PA project stated below

Number of interview questions: There are 13 questions covering the three major question areas (A), (B) & (C)

Approximate length of interview time: 30 minutes.

**Purpose of research:** To gather data from industry stakeholders regarding the current state-of-the-art of existing open data platforms in order to meet the demand of the Route-To-PA Task 2.1: *The “State-of-the-Art Report and Evaluation of Existing Open Data Platforms”*

**3. Introduction:**

*Question coverage:* Our questions cover 3 major areas:

4. platform challenges
5. desired platform features and priorities of the features, and
6. other features and issues surrounding ODP capability to support the enhancement of government transparency, accountability and general adoption.

*Stakeholder coverage:* Stakeholders to be interviewed include:

4. Data suppliers or producers e.g. mainly government agencies, but also businesses (the upstream community)
5. Platform developers or ODP service providers (midstream community)
6. Researchers/Analysts, Data Journalists and Apps Developers (the downstream community)

***Peripheral data collection about the interviewee's and his/her company or organisation:***

Name **\_Adam and Trevor**

Company or organisation **\_ Marin Institute, Co Galway \_**

Stakeholder group **\_Data Generation, collection and Publishing**

Position/designation **\_Adam (Team Leader on Data Management); Trevor (Scientific/Technical Specialist on Geospatial Information Systems)**

Typical task at work **\_Coordinating data management, data project management for the Marin institute (Adam); Developing data Strategy for the geospatial data systems for institute.**

**4. Interview proper:**

*Note to the interviewee:* In this interview, you are required to give information, in most cases, as it pertains to you as a stakeholder or user [Researcher & promoter] of ODP and/or (if a company / organization) as a company/organization having a stake also in the industry.

Where required, please give a general comment or information as it affects the ODP ecosystem.

*Note to the interviewer:* Ask each question clearly, introduce examples and scenarios, terms, topics and keywords provided in the questions to help respondent where necessary. Avoid interjection or interruptions as the interviewee responds to question. Be time conscious.

***Interview questions:***

(A) CHALLENGES:

9. Challenges associated with the use of open data portals and platforms.

a) Access to open data published or available on the platform:

v. *In your opinion, what are the challenges facing users in assessing data published on the current generation of open data portals?*

[You can, consider roles and benefits from the point of view of various stakeholders – the data suppliers/publishers, government, ODP developers or providers; data consumers – Apps developers, data journalists and analysts, and also the ordinary citizens]

**Note (1)(a)(i): Talking about challenges**

**i) Getting hold of the data** – being unable to reach the data due to people's reluctance to release data

**ii) Low quality** of data being published and the lack of enforcement of the data standards

**iii) Discoverability** of data is poor because some datasets are difficult to find in some cases. There may be too many keywords which pose the problems of confusion, clustered access but data access needs to be clear. Also most data on portals at this stage requires too much cleaning effort before you can see the meaning or use it.

**iv) Poor standards of datasets**

---

*ii. From the problems given, which of them do you consider as the fundamental obstacle to the use of open data particularly by the non-technical users and the ordinary citizens?*

[Elaborate on the chosen obstacle particularly as it affects you (the interviewee) as a stakeholder and also if the same problem has a general industry impact]

**Note (1)(a)(ii):**

- Poor data standards, poor level of metadata and documentation about datasets
- Lack of knowledge about source and who to contact about data question for clarification
- Inappropriate table and row headings for excel spreadsheets.
- Differences in definition of terms leading to misperception

---

**b) In terms of understanding the published datasets**

*ix. In your opinion, what are the barriers to making sense of and effectively using published datasets on the open data platform?*

[You may consider, the way datasets are presented – formats, publishing styles, tools for manipulations, etc.]

**Note (1)(b)(i):**

The fundamental obstacle to data access is poor standardisation of practices which does not permit users and other stakeholders of the open data to easily see

- the **how** dataset are generated or collected
- the **who** are involved in the collection of the datasets
- the **where** datasets come from (source of data)

All these compound the inability for users to find answers to some of their questions around open data resources such as the meaning of rows and column headings of an excel spreadsheet dataset. They also go further to affect understandability and usability anyone can make of datasets and the problem can cause misperception of terms, confusion due to inconsistency in definitions of terms, the validity of data, and lack of links to the data collectors in order to ask questions. Another area of serious challenges in OD practices stems from the fact that same kind of datasets from different origin tend to carry different sets of metadata and data descriptions style. The problem appears to relate also to the problem of poor standardisation which in this case makes unification of datasets very difficult. So one can see that standardisation of data presentation is a very important issue.

---

*x. In your opinion, how interoperable are the existing ODPs? Give a general comment on the interoperability of the platforms with reference to extensibility, data harvesting, data publishing, data linking, etc.*

**Note (1)(b)(ii):**

Interoperability of ODPs seems good within sectors in the open data domain. For example, within the environmental and geospatial data subsectors data interchange/exchange is fairly easy. These types of very specific area of data environment are interoperable within the similar data type sub-area such as the Data.gov.ie and the ISDE data publishers. However, making the specific datasets interoperable with other datasets from outside the subsector – say agriculture data with health data or geospatial with agriculture data may prove difficult. Datasets from data.gov.ie and ISDE are fairly interoperable as said earlier partly also because CKAN provides tools that permit easy harvesting of datasets across these portals. The thing with platform interoperability is that a platform has to have data exchange format which other platforms can read, and this function is provided by CKAN through the use of Catalog System for the Web (CSW). When the interviewees were pushed to rate the interoperability performance of ODPs (in terms of poor, average or above average), they are of the opinion that the performance is fair enough because there are some good tools that support the some functionalities. However, factors such as configuration, plug-in of own vocabulary and extensibility that are hindering the performance from going above average level exist.



- c) Government transparency and accountability enhancement are some of the main goals of ODPs

The rationale behind Open Government Data can be summarised into two parts: Open Data advocates propose that making government data available to the public increases government transparency and accountability. Open Data Platform (ODPs) is the technological infrastructure that enables these objectives to be achieved.

- ix. In your opinion, is there any characteristic feature of ODPs that you think might be hindering or enhancing the achievement of government transparency and accountability through the use of existing ODPs?

[You may think of ODP characteristic features such as:

- personalisation of dataset search and data consumption pattern
- quality of datasets through enforcement of data formats, metadata and provenance standards
- recommendations provided on datasets for users based on users' profiles and consumption patterns
- integration of related datasets using linked data, and
- basic analytics on datasets to detect violations of rules]

**Note (1)(c)(i):**

On the question of whether there are features on existing ODPs that might be hindering or enhancing government transparency and accountability, interviewees' first response is that the mere existence of ODPs and the extent of involvement of governments in their development and other supportive roles are testimonies of transparency enhancement because governments in OD practicing countries across the world do in fact publish some datasets that some citizens look at on these platforms. These are positive progress towards transparency and accountability. What is lacking and perhaps hindering progress are the fact that usability of platforms by citizens is still poor and this could be traced to the problems such as platform interfaces which are not clean-looking and non-user-friendly enough to attract site 'traffic' (especially the non-technical users of data) compared with what you can see on Google and similar establish online platforms. The advanced users of data such as developers, researchers, and particularly other members of the mediator group of stakeholder may find the interfaces of existing platforms easy to engage. However, the non-technical ordinary citizen users may want simpler, easier interfaces with plenty of documentation, a neat presentation of catalogs with nicely arranged glossary.

When interviewees were asked to relate platform performances in support of transparency and accountability through features such as personalisation of dataset search, quality of datasets through enforcement of data formats, metadata and provenance standards, recommendations provided on datasets, integration of related datasets using linked data, and provision of basic analytics, respondents quickly ruled out the use of Linked Data in current ODPs. They are not aware of any platform using linked data to support transparency and

accountability but noted that this area is probably an important one to be improved upon in the next generation of platform technology. On the positive side, searchability as transparency enhancing feature is satisfactory but with instances of good cataloguing, nevertheless, the majority of the latter need improvement on their own although Marin Institute as a data publisher does well in cataloguing services.

- x. Considering the larger society in which ODPs exist, is there anything in your opinion from experience or observation, through theoretical knowledge or reasoning that you think might be hindering or supporting the performances of existing ODPs in terms of adoption, popularity and impact on the society?

[You may think about political & social/societal issues such as government policy, citizens' attitude to & skills for adoption of OD, uses & participation on ODPs, OD concept promotion; practitioners' encouragement, rewards and incentives for sustained involvement & contribution, etc.]

**Note (1)(c)(ii):**

Interviewees believe that platforms performances based on current technology are not so bad talking about their cataloguing functionality. Perhaps this is based on the fact that they are generally having very 'light' (non-sophisticated) user interfaces. Even though, they don't usually engage cataloguing services on portals being publishers themselves, they believe that portals such as data.gov.ie and data.gov.co.uk are performing fairly well. When asked to relate the question of performance to political interplay with the domain of OD, social and societal influences, policy effectiveness and citizens' attitude towards adoption and use of OD and ODPs, respondents advocate the notion that government policy is supporting OD/ODPs concepts. To the contrary, they believe that majority of the ordinary citizens outside the OD academic environment, developers, platform providers, researchers and data consumers are unaware of open data concepts. According to them, the lack of awareness is due to the fact that nobody is going around engaging citizens and preaching OD door-to-door. An effective way to promote OD, therefore, is to go to the community dwellers' level to educate the society on the concept and benefits of practicing open data concept. Compare with Ireland, UK is not doing anything better in terms of promotion of OD within the society according to the interviewees.

In terms of using incentives and rewards to attract and encourage OD practitioners, publishers, data owners users and others to improve data release, data consumption and ODPs development and usage, respondents see no direct approach in this regard at the moment whereby political or community leaders are providing encouragements. One of the respondents is currently involved in a programme to promote the growth of the Irish spatial data content, data and metadata exchange with other different bodies in the sector. This kind of promotion is easy because it is easier to try to influence experts in the domain because they know the importance, challenges and benefits of OD than non-OD-oriented people. In their opinion, the best way to influence non-OD-oriented people is by using concrete examples (use cases) however, not with two-datasets

in a portal catalogue but with catalogues that contain huge amount of datasets from several sources or sectors of the economy. Furthermore, OD team leaders from various stakeholder groups (from within the downstream, midstream and upstream categories) should play roles in coordinating OD promotion to the public awareness. In the case of the ocean data subsector, respondent mention the use of citations to give credits to sources of data providers as incentive for recognition of work and encouragement to provide more. They also follow the UNESCO recommendation to grant incentives to data providers.

---

(B) DESIRED FEATURES AND THEIR PRIORITISATION:

10. From your point of view as a stakeholder (data supplier/platform developer/platform resource user) what feature(s) of the existing ODPs is/are most important to you and why?

**Note 2:**

The most important open data features for these interviewees include: **simple data search facilities** including advanced search, and as a publisher – easy method of getting datasets into the portals (uploading) so that repeat operations can be avoided. An additional feature desirable is **simple content management features**. By relating the demand for desirable feature on platforms, one of the respondents recalls that he is missing a kind of customisation functionality in all platforms. This missing feature is supposed to allow him add standard vocabulary to describe datasets in a consistent way and with same meaning of terms across all platforms. As a matter of desirability but no availability, the ODP UIs are reckoned by respondents as not advanced enough; just too ‘straight-forward’ (meaning not sophisticated) to permit additions of features to enrich user experience.

---

11. In your opinion (as a stakeholder in your category) are there additional features, that will enable better supply, use/reuse, collaboration, communication, sharing/distribution of data, commenting, rating, co-creation of services, other transparency-enhancing tools (such as personalisation, standards enforcement, data recommendation to match consumption pattern, integration, basic analytics, etc.) on the platforms currently in existence?

**Note 3:**

As additional features especially those of social influences that need to be integrated into platforms, interviewees look at issues from their point of view as stakeholders and mentioned spatial interactivity referring to possibility of **interacting with data on a map** – for example, a user discovers a link to dataset on a portal and clicks on it; then that takes the dataset and spread it out on an interactive map for the user to view. **Tools to drill down on data provenance**, to the respondents, are additional desirable features which platforms must have. The explanation uses a scenario whereby a raw dataset published on platform is taken and manipulated enriched with pieces of data from other datasets and reincorporates the result into the

platform. The new partially processed dataset becomes another data resource for further uses. Tools for provenance tracking should be able to reveal all aspects of the movement of data and the processes it go through from the initial origin to the current stage in order to show to the latest user how it came about. To the respondent this is more than just data file versioning; it is another useful form of data transparency because it reveals the provenance records, historical movements, manipulations, processing, conversion, and other enhancement activities so done to the original raw data up to the newest form.

When asked to revisit the issues about user communication with one another, sharing of and commenting on data, etc. in order to critique, develop and distribute knowledge around datasets, interviewees quickly respond that platforms of today are lacking good representation of these functionalities despite their desirability. Furthermore, what's there on platform at this stage is mainly raw datasets and that the little amount of user generated contents on platform are sadly not captured for further uses due to the fact that platforms do not really offer opportunities for user-generated content around dataset and as well as capture of same as crowd-sourced data for analysis. On the other hand, data resources on platform are not sufficient and also do not support good interlinking between platforms to drive user interactions, comments and recommendations. Poor data interlinking is seen as a problem from the poor data transport network within the ecosystem. When the interviewees were asked to suggest solutions to the issues of user content not being generated on platforms and when generated, it is not captured for further uses, they suggest the need to introduce more tools especially social media tools for sharing and commenting on data and even for manipulating data. A new idea discussed is to use tools to link datasets on platforms to publications and reports on places or situations where such datasets have been used to produce value. Simply put, link datasets on platforms to previous established use cases which should serve as examples to the public of how datasets could be utilised just to get them thinking in the new data business models.

---

12. What feature(s) of the current platforms would you say is/are performing to your (or users') expectation – either in general, applicable to all platforms or specific to the platform you use?

**Note 4:**

In their opinion about the performance of current ODPs that meets expectation, first, they mention keyword search, keyword linking and suggestions are doing fine as long as those keywords are not too many. Secondly, storage facility or methodology has no problem because the majority are of datasets are store locally on the portals rather than in a huge centralised database, and this situation may facilitate harvesting/federation of data and metadata according to the interviewees from Marin Institute, Galway.

---

13. In your opinion, what type of platform feature or tool or capability would you advice be improved upon, especially those affecting the main goals of ODP (e.g. transparency enhancement) and why? Name one (if any) that requires critical improvement.

**Note 5:**

Open data platform features to be improved upon is definitely more of those that impact positively on transparency of government. For example tools that enable better use of provenance metadata, display and tracking. In order to improve usage of platforms, the designs of their interfaces also need to be improved. From the point of view of geospatial subsector, data visualisation has to be improved so does mapping facility to enable users interact with the data. Generally, cataloguing services can still be enhanced further even though it is rated as part of the good side of the current technology.

---

(C) OTHER FEATURES AND ISSUES:

14. Given the opportunity, which one area of an ODP ecosystem attributes including the environmental factors, would you like to change, and what change would you introduce?

[Think about – government policies, citizens’ attitude to adoption, availability of skills, sufficient understanding of the concept of open data publishing and usage i.e. what to do with the ODP infrastructure, incentives for participation, etc.]

**Note 6:**

The **promotion** of OD concept to the general public need critical improvement to enable people see how the government is trying to enhance transparency of their activities and decisions; and talking the best mode to promote OD and use of ODPs, the respondents opted for use of **documentary media materials** such as radio commentaries (good for Ireland), videos, YouTube videos, and so on. **Skill development** is important because without open data skills, citizens will not be motivated to actually engage the platforms. Therefore various training programmes in third level institutions are recommended for data skills development because you need data-conscious citizens in a data-oriented ecosystems or economy.

When presented with the opinions of other interviewees regarding difficulty in designing viable business cases in the OD domain, the difficulty in determining cost input and revenue stream, what should be your offering and value creation strategy, respondents say the problem is that people involved are often just reluctant to develop the useful business cases and not necessarily the difficulties aforementioned. By inference from the explanations of the respondents, the reluctance to develop open data business cases come not from the difficulty in designing the business case but, first, from the choice of data owners to retain ownership of their datasets rather than release them. Second, data owners, perhaps, do not understand the new data business models so cannot design some. By engaging with open data community, data owners could see how their data are being re-used by other stakeholders including developers, researchers, consumers etc.

However, they might not be able to put a value on this reuse, but might be able to start thinking about other possibilities on how to generate values or revenue stream in an innovative way around the re-use of the data. For example, when developer build apps to support the re-use of data, an idea might come up on how to introduce charges between the supply of data, the app that grants access data and enable users to enjoy benefits of information provided by the data. Thus, data owners need to think in the direction of completely different types of business models based, this time, on data as a resource and not in the direction of the traditional older-style business models based on normal commodities. Furthermore, data owners need education to learn about the new data business model design and how to leverage data as a resource for value generation.

---

15. In theory, ODPs are infrastructures to promote transparency of government activities, to bring about citizens' participation in governance and co-creation of better services (public/private) that suits their needs and the participation in decision-making on issues that affect the society. What is your opinion regarding how well (or otherwise) do the existing ODPs support these objectives?

**Note 7:**

This question seeks to understand the supports offered by the existing ODPs towards the achievement of objectives mentioned in the question. Respondents maintain a view that ODPs do support transparency of government but they are unsure whether the same has been achieved in term of support to co-creation of services and user feedback into the system. Regarding participation of citizens in governance and decision-making, they believe that depends on

- the sphere in question
- whether citizens are actually interested in participation [motivational factor] and
- the availability of data to make informed judgement.

It is possible in some areas to support decision-making, declared by respondents; however, the lack of data will impact negatively on decision-making support in other areas. In the sense of co-creation of services, interviewees believe that certain things are missing in the loop – things [tools] that encourage and enable people to feed back their creations into the system. Furthermore, user-generated contents are not currently being captured and these contents ought to be used as crowd-sourced inputs for decision-making and co-creation of better services. This problem arises as much from the lack of appropriate tools on platform as is from the lack of user-base on platforms to generate the user contents.

---

16. Give your general remark on the technological state-of-the-art of the existing ODPs?

**Note 8:**

In giving their general remarks, respondent first referred specifically to CKAN which, according to them, tends to lead the way of the technology. Generally, the technology is at a stable stage and this is seen as a good thing when viewed in relation to a volatile technological environment – one with unnecessary technological escalation whereby killer technology might not allow enough time for a well thought-through development ideas and designs at a particular stage before moving on to the next stage. In order to lift the current state-of-the-art of ODPs further, respondents recommend:

1. A full understanding of all cataloguing technologies in the market and analysing them, and then taking the best of each system to be unified into a [‘hybrid’ of] cataloguing system for the next generation of platforms. A nice feature to search for anything on the platform is also desirable.
2. Using analytics to see how people use platform system and functions e.g. how they search with keywords, what they search for so as to know what their interests are, etc.

In term of social media/network improvements, the interviewees remark that:

3. The bad situation of the current platforms needs to be upgraded especially in collaboration areas.
4. However, the kind of social media application recommendable for the ODPs is not the type of Twitter or Facebook which gathers too much data on profiles and statuses of users; rather, it should assume a form of question provocation and answer provisioning for users of data resources and other members of the OD ecosystem and the public.
5. The experts within the community (data suppliers, developer, platform managers, researchers, analysts, government agencies) should assist in the provision of answers to questions generated by users.

On visualisation earlier mentioned as important areas for improvement, interviewees were asked to suggest how this feature could be improved. In this case they remark that visualisation on dashboards appear well designed in some existing data portals such as the following Ireland-based organisations which have simple charts and other visual graphics:

- Dublin Dashboard <http://www.dublindashboard.ie/pages/index>
- Irish Ocean Energy portal <http://oceanenergyireland.com/>
- AIRO Data Store <http://airo.maynoothuniversity.ie/>

These graphics enable users to quickly make sense of the data without having to look the rows and columns of figures in tables or spreadsheets. Other important features demanding improvements in next generation platforms should include :

- Real-time data streaming for latest observations and values
  - Videos and YouTube videos and interactive functionalities which are of more values for quick meaning of data to users.
  - Attractive manner of data presentation will enhance users experience as they arrive at a data site.
- 

**5. The interview conclusion:**

*Vote of thanks:* Thank you [name of the interviewee] for the attention granted for this interview. I appreciate your effort and patience in explaining your opinions

*Permission for follow-up:* I seek your kind permission return for further clarification of any unclear responses if necessary.

*Confidentiality:* I wish to reassure you of the confidentiality of your personal data will be upheld as state earlier in this interview.

Interviewee's signature against responses recorded \_\_\_\_\_ Date: \_\_\_\_\_



## APPENDIX 2: GENERAL SUMMARY OF ODP FEATURES

Features	CKAN	DKAN	Socrata	Publish MD	Info Wbch	Enigma	Junar	ODS	Callim	DataTK	SMWiki
Installed instances	116	No info	No info	6	No info	1	20	38	No info	4	No info
Data/metadata/file format standards	CSV, XLS, ArcGIS, Inspire & Geo, CSV, XLS, ArcGIS, Inspire and Geo, DCAT	Support DCAT, INSPIRE, CSV, XML, & RDF. Upload Files in any format	RESTful open API, open data API supports App ecosystem. Files: , CSV, XLS, & XML. DCAT, Geospatial	Linked data standard APIs. RESTful, Turtle, RDF/XML RDF SPARQL, DCAT	Uses RDF format & SPARQL for queries	RESTful APIs, Direct plug-in standard	Supported formats: CSV, XLS, XLSX, KML and XML	Several types of APIs. File format: CSV, XLS, XLSX, SHP, KML, GeoJSON, OSM, GTFS & ShapeFile	RDF XSLT & XProc, RDFa, CSS3 SPARQL XHTML5, JavaScript, RESTful.	CSV, XML and JSON file formats	CSV, XML, JSON RDF, MediaWiki
Search/Indexing	Search API, query/access data, RESTful API, download, keyword search, filter by tag, facet index	Clear search facility, filter by metadata. Search UI, little description for result	Robust search index; allows filtering	SPARQL & other query tools for searching. Limited keyword search on catalogue data	Provides no user interface or API for searching	Augmented search tools. Powerful search UI & API; search for data at record level	Categorises data & adds metadata to improve searchability. Limited search	Dataset Search API, Records Search API, multi-criteria text search	NA	Limited filtering by dataset name	Text search. SPARQL query. limited filtering
SM/collaboration/sharing	social media tools: Facebook, Google+, twitter, etc. support communication collaboration, comment, sharing, RSS, follow	Tools to manage content & community, Supports social media: blog, comment, Drupal, Disqus comments, sharing, collaboration & interaction	Civic engagement, participation & social experience: comments, rating, feedback Connects OD initiatives to the broader app ecosystem	Interactive tools that support collaboration sharing	Uses social network info or other web sources. Wiki style UI for collaboration	No	Interaction, share, distribute Tools, SNS, share, feedback,	User engagements, popular SM, forums, GitHub, discussions, feedback	Yes/ Limited to wiki pages	NA	Semantic media wiki style collaboration
Publishing & workflow	Streamline, import via web UI, update, refine, workflow for groups, public /private, add metadata to upload, access control	Full cataloguing, easy publishing. Editable UUID, Upload using web front end. workflow attach metadata to dataset	Automatic publishing 'push mode'. Configure publishing & workflow, control of dataset, private /share, web-based data upload	Use Linked Data standard to publish. Converts file from CSV to RDF.	Support data integration via semantic model, has connector, file conversion	Discover Public Data – a repository of data from governments, and other organizations	Simplified data publishing; easy-to-use platform. workflow optimises publishing	Hosting & admin, cloud hosting, Integration, workflow for data publishing	Wiki pages for publishing & workflow	Provides tools for ETL	Publishes via Semantic Web, WikiText

Harvesting / Federate / catalogue	Customisable harvesting from Geospatial CSW Servers, existing web catalogues, simple HTML index pages or Web Accessible Folders, ArcGIS, Geoportal Servers & Z39.50 databases. Cataloguing, strong integration & federate	Complete suite of tools for cataloguing and harvesting dataset.	Network creation with regional hubs. Federate with other customers. powerful harvesting. cataloguing	Provides datasets catalogue	No specialized support for harvesting and federation	NA	Access data from application, Harvest from REST & SOAP, HTML Forms. No federation	Cataloguing & export, linked data capabilities, Data from external sources.	Uses templates format for data collection	NA	No stated cataloguing. Limited federation & harvesting
Extensibility	60 extension options, open source, nice Json API, links to external datasets	Has 18,489 extension modules to support customizable functions with easy dataset management.	Scalable cloud platform, co-creation & crowd-sourcing. Nice API & library to easily extend capabilities	Allows development of branded data site, Tools for use with Linked Data. Open Source. website compatible with browsers	Integrates Dataset, links Organisations together. Supports developers' community, allows extension & connectors	Provides API tools for Apps & services	Integrates data Directly into the user's site. workflow optimises limited publishing.	Limited extensibility	Simplifies integration of new data with existing data using JavaScript RESTful & RDF	Open source project	MediaWik allows extensions seamlessly
Data Analysis	Admin dashboard & data members management, no special tool for data analysis	Support Google Analytics, Publishing maps with CartoDB.	Tools for maps & charts. Basic BI tools. Library for statistical package R	NA	Support analysis & visualization, R statistical package	Analyse data, combine/view : Time series & Join analysis.	Analyse & report on feedback. But limited	Basic analysis API, Records Analysis. Visualised interactive graphics.	NA	Creates a holistic view of data across depts, analyse & interpret data.	NA
Visualisation	Basic visualization for tabular data & by charting, mapping & imagery, etc.	Visualization features exist but limited support	Powerful tools for machine readable & geospatial data visualization	NA	Visualization of Twitter followers, visualizations widgets	NA	Visual graphics & reports, tables, charts & dashboard & integrate with google analytics.	Powerful API for interactive visualisation: maps, chart, pictures, Geo data & images	NA	NA	Limited
Personalisation	Themable features, personalisation settings	Theming is available for personalisation.	Personalised sorting, auto-filtering to view & portal admin	Grants administrative control for publishing	Flexible data-driven UI, Self-service, allows users preferences	NA	Personalisation is possible	Personalised connectivity and data Federation & data processing	Limited personalization	Limited personalization	Wiki style UI allows user preferences

Customisation	Customisable harvesting / importing of data, customised extension	Customisation with theming, API and Drupal extension	Tools to customise portal admin & metadata mgmt. data inventory via APIs or data. file	Grants admin control of customisation of platform but limited functions	Develop and deploy custom apps	NA	Customisation is possible	Customised GUI; Embeddable widgets but Limited	limited customization for various user groups	Limited customization	MediaWiki platform is highly customizable
Licensing for dataset	Yes	Yes	Yes	Yes	No	NA	NA	Yes	No	NA	Yes
Accessibility	No info	No build-in support for accessibility, but can be added using Drupal accessibility modules	Uses common best practices to allow accessibility	Linking information can be added as metadata	No	NA	No special features	No special feature	No special feature	No special features	Easy to find relevant content
Technical Environment	Python	PHP, Drupal CMS	Scala	Ruby on rails	Java & web apps	NA	Java & Python	No	Java	PHP based application	PHP
Others	Good manual Simple to use	Easy to use platform	Tracking & Measure of performance	Flexible, cloud-based, easy to use	R stat, support transparency, linked data	Reliable, scalable, large OD Analyses	Track & measures user impact on OD	Remote web services; easy deployment	Guides, videos, tutorial. Linked data	Deal with fraud, aids transparency	None