ROUTE TO PA

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

European Commission

Project number 645860

H2020-INSO-2014

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

(Version 1.0 - Final - 01/06/2015)

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: 31 May 2015

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

TABLE OF CONTENTS

Executive Summary

Opening up government data to the public has been recognized to have a significant impact on enhancing transparency and openness of public sector entities while promoting new forms of accountability and improving citizens' trust in governments[1]. In response to the European Public Sector Information (PSI) directive, many European Union (EU) member states have launched their Open Data initiatives[2] with over 8,000 datasets available on the EU Open Data Portal[3]. However, due to persistent barriers such as: 1) limited access and use of Open Sata by citizens and third-parties, 2) limited budget and resources on the part of government agencies to publish new datasets of high value and 3) weak legislative framework to enable ethical reuse of available datasets[4], the high expectations have not been satisfied. Therefore a need for new approaches arises, to improve accessibility and understandability of Open Data. One of the potentially successful methods involves explicit support for social interaction over published datasets as a means to increase data and government transparency[5]. Such next generation platform could be realized through the integration of Web 2.0 or Social Media technologies with traditional Open Data platforms[6]. However, efforts in this direction are just beginning to gather interest and momentum in the Open Data community. The Route-To-PA Project[7] is a specific instantiation of this kind of methodology.

The Route-To-PA project (Raising Open and User-friendly Transparency-Enabling Technologies for Public Administration) which provides the context for this study; aims to enable the transition into the next generation Open Data platform paradigm by creating tools that will enable citizens to social directly engage over Open Data resources. The project also aims to provide specific explanatory tools that could be incorporated into existing Open Data platforms to deliver greater transparency, quality and understandability of the datasets. *Building such tools and technologies is contingent on investigation and evaluation of state-of-the-art Open Data platforms to determine current set of capabilities and possible expandability of the platforms*. While there are a few existing studies on Open Data Platforms[8], 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. In addition, none of these reports also look at the social aspects of Open Fata. This report attempts to bridge this knowledge gap.

This report aims to provide better understanding and evaluation of Open Data platforms with respect to:

[1] 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[9]:

---

[1] 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

[2] 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

[3] European Union Open Data Portal, available at https://open-data.europa.eu/en/data

[4] 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

[5] 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).

[6] 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.

[7] http://www.routetopa.eu

[8] E.g. the study on Technical Assessment of Open Data Platforms for National Statistical Organisations, 2014, by the World Bank

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

accessibility, usability, understandability, informativeness and auditability, as well as social interaction and collaboration on datasets;

[2] The shortcomings 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 with respect to dataset transparency, social interaction and collaboration on datasets and

[4] 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.

The study relies on the analysis of information gathered from review of literature, survey of eleven state-of-the-art open data platforms, stakeholder interviews, and stakeholder workshops in Dublin (Ireland) and Prato (Italy). The platforms reviewed and evaluated include: CKAN, DKAN, Socrata, PublishMyData, Information Workbench, Enigma, Junar, DataTank, OpenDataSoft, Callimachus, DataTank and Semantic MediaWiki. The first and fourth objectives were addressed through data from platform evaluation (Section 3), while the second and third objectives were based on data contributed from interviews and workshops on perception of stakeholders (Section 4).

To address the first objective, 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. 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 standout 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 in terms of 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.

Table 1: Summary of Platform Features

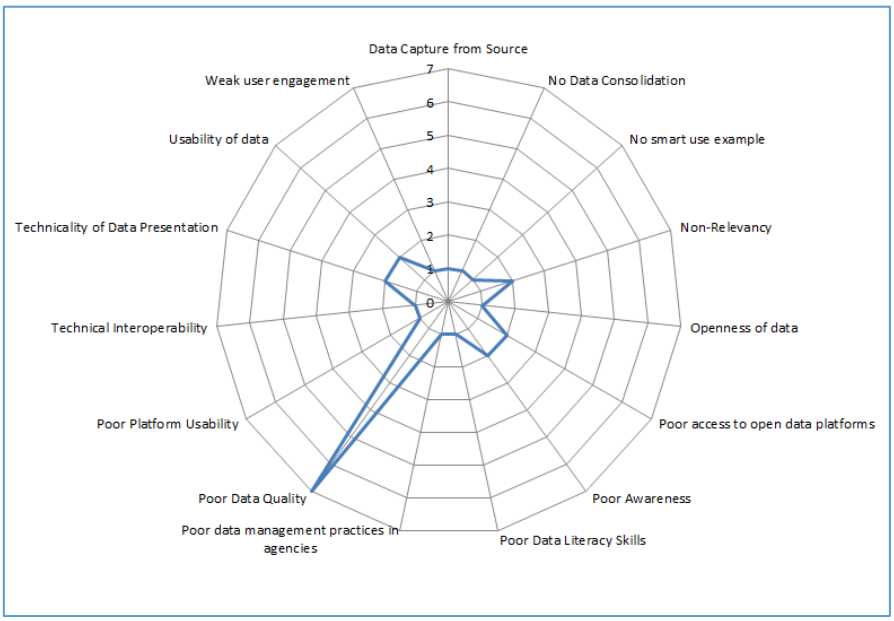
FEATURES	CKAN	DKAN	SOCRATA	PUBLISH MY DATA	INFO WKBENCH	ENIGMA	JUNAR	ODS	CALLIM	DATATK	SMWIKI
DATA, METADATA & FILE FORMAT STANDARDS	●	•	●	•	•	•	•	●	●	•	•
SEARCH & INDEXING	●	•	•	●	X	●	•	●	X	•	●
SOCIAL MEDIA, SHARING & COLLABORATION	●	●	●	•	•	X	●	●	•	X	●
PUBLISHING WORKFLOW	●	●	●	•	•	•	•	•	•	•	•
HARVESTING, 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 performan-ce	Flexible, cloud-based, easy to use	R stat, support transparen-cy, 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 transparen-cy	None

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

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

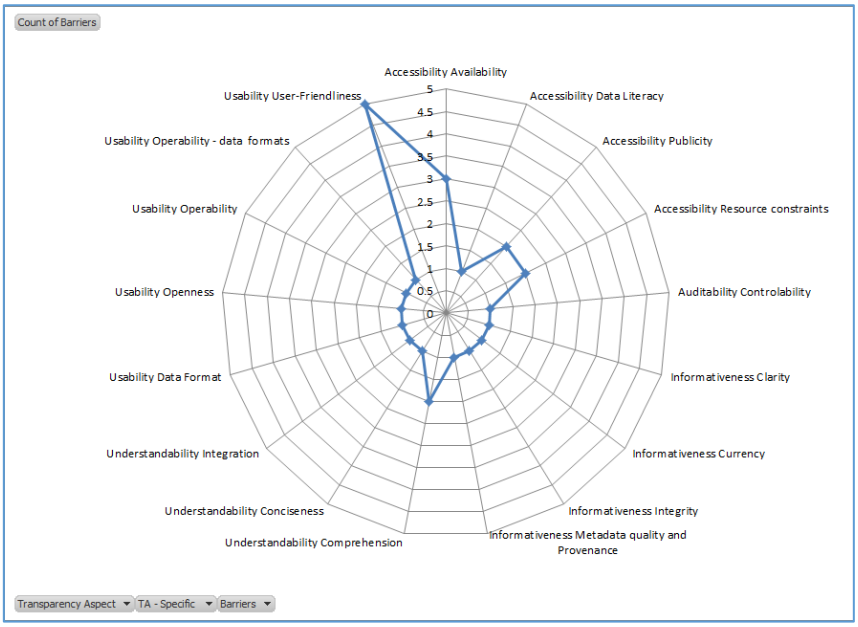
Our analysis showed that the most common barrier to the use of Open Data platforms and Open Data itself *is the perceived poor quality of datasets* available on the platforms. Poor data quality according to stakeholders is associated with poor metadata, failure to use the right format for different audience 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 1: Perceived Barriers to Use and Adoption Open Data Platforms



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

Figure 2: 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. The cloud-tag below in Figure 3 was generated from the contributed solutions and features to identified stakeholder needs and barriers.

Figure 3: 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 standout 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[10], 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 in its use of personal data should in general help to build the trust of citizens and improve accountability of policy makers[11]. 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[12].

In the past few years, Open data programs have featured prominently as a major 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 enable the desired transparency in its different aspects. In fact, while, opening up data, processes and decisions of governments are in general expected to improve transparency, recent studies have showed high quality transparency depends not only on how visible information is made, but on how well it lends itself to accurate inference[12]. Even more recent studies[13] 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 (supplier party) to citizens (recipient party) for the purpose of informing and involving citizens in government decision making, 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 for 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 expects platforms in general to support the open data best practices[14]
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

[10] EU Anti-Corruption Report, Report from the Commission to the Council and the European Parliament, February 3, 2014
[11] European E-Government Action Plan - http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:0743:FIN:EN:PDF
[12] Greg Michener and Katherine Bersch, Conceptualizing the Quality of Transparency, 1st Global Conference on Transparency, 2011
[13] Eliezer N. Mishory, Clarifying Transparency: Transparency Relationships in Government Procurement, Government Procurement Seminar, Chris Yukins & David A. Drabkin, November 4, 2013
[14] Data on the web best practices, http://www.w3.org/TR/dwbp/

data portal by creating tools that will enable 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 data transparency and quality and understandability. *However, building such tools and technologies requires good 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[15], none of these studies specifically address the affordances of these platforms with respect to the quality and transparency of open data published on this platforms about government agencies and public authorities. This report addresses this gap by providing a describing state-of-the-art platform 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.

The rest of the report is organized as follows: Section 2 presents the methodology for the review and evaluation 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 next desired features of next generation platforms. Section 5 summarises the findings from the study while concluding remarks are given in Section 6.

[15] E.g. the study on Technical Assessment of Open Data Platforms for National Statistical Organisations, 2014, by the World Bank

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[16]: 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.

[16] Cappelli et al, Managing Transparency Guided by a Maturity Model, 3rd Conference on Transparency Research HEC PARIS, October 24-26th, 2013

- *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.

2.2 ANALYTICAL FRAMEWORK

In this section, we present models that allow us link open platform features with transparency qualities or attributes of datasets maintained on those platforms and expected impact of greater data transparency on perceived organizational transparency. The whole Route-To-PA innovation is implicitly based on this thesis.

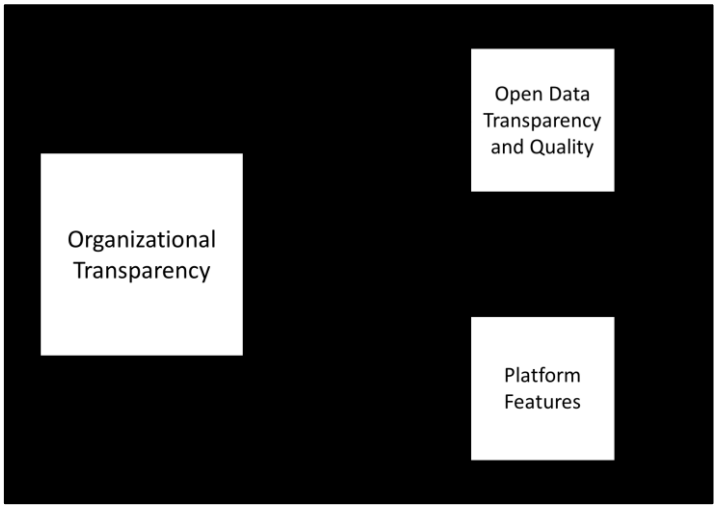


Figure 4: Model for Open-data based Organizational Transparency

Albeit, there are several conceptualizations for transparency construct. For Open data transparency in model presented in Figure 4, we adopt the deconstruction of transparency presented in Cappelli[17]. The conceptualization provided five major aspects and several sub-themes of transparency including: usability, accessibility, auditability, informativeness and understanding (see Figure 5).

[17] Cappelli C et al, Managing Transparency Guided by a Maturity Model, 3rd Global Conference on Transparency Research HEC PARIS, October 24th – 26th, 2013

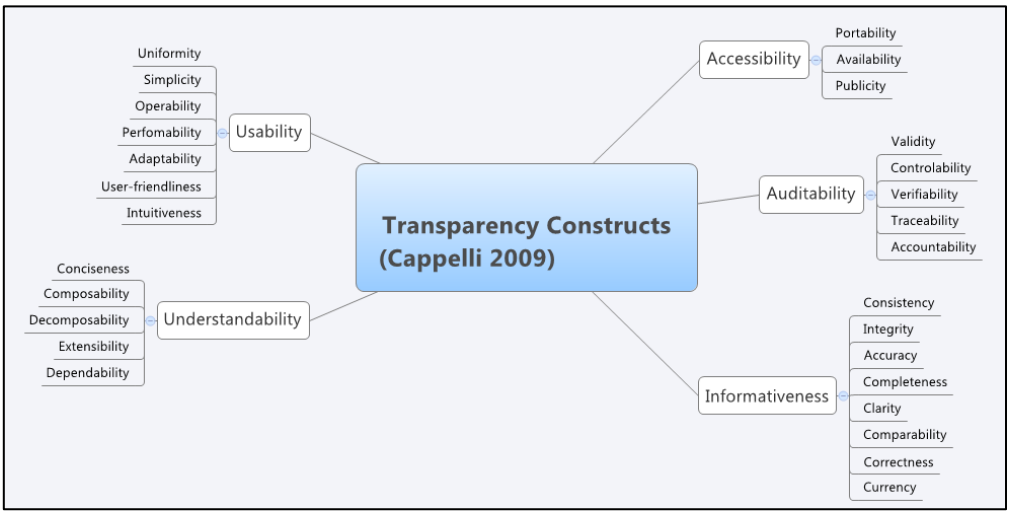


Figure 5: Transparency Construct decomposed into sub-constructs

These constructs and sub-constructs were used to group identified barriers by stakeholders (Question2). The constructs were also used to link barriers to specific platform features. For instance "poor data quality" as a barrier could be associated with the "informativeness" construct and subsequently linked to platform feature such as flexible or extensible metadata schema for instance as enabling technical feature. See Figure 6 below.

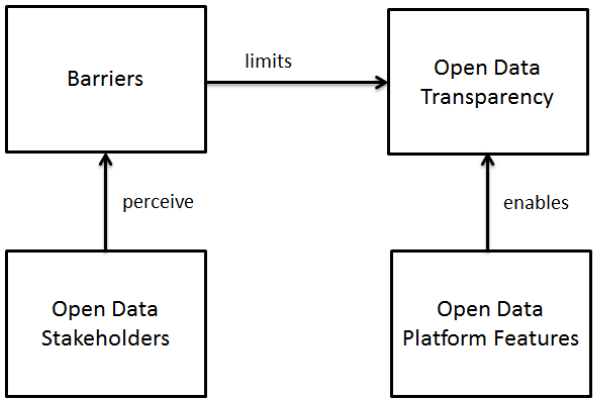


Figure 6: Linking Barriers to Transparency Constructs and Platform Features

2.3 DATA GATHERING

Four data gathering methods were adopted in the study. This uncovers a wide range of sources and materials that will provide sufficient and applicable information necessary to describe the current state-of-the-art of Open Data Platform (ODP). We evaluate the materials from these sources and filter relevant details especially in relation to Social Platform on Open Data (SPOD) and Transparency Enhancing Toolset (TET) in line with the objectives of the project of Route-To-PA. The four methodologies applied in data gathering include – 1) Literature review of books, journals and articles on open data platform; 2) Platform websites literature review

of documents, guides and manuals, online blogs and comments; 3) Workshop on open data and open data platforms and 4) Interview of open data domain stakeholders.

Literature Review - Deliverable 2.1 demands the "State-of-the-Art Report and Evaluation of Existing Open Data Platforms" in order to reveal the status of innovation of open data technology through an in-depth exploration of several sources of information to uncover existing open data platforms profiles. Task2.1 (State-of-the-Art Investigation) specifically requires the investigation of the current state-of-the-art of open data platforms to produce a report on existing platform capabilities or functionalities and an estimation of the current innovation effort in those infrastructures. This exploratory research takes a four-dimensional approach (as explained under section 2.3 below); the aim is to gain a robust base of data and information necessary to produce the required SOTA report. The first part of the report assembles information from available and related books, journal, articles, and conference papers on open data platforms. The second part presents information from selected open data platforms gathered from their webpages through a review of the literatures in form of webpages, blogs, comments and other online materials such as graphics, images and charts. The third part assembles data from field exercises – conduct of workshops and interviews with stakeholders of open data concept and applications.

The information from all 4 sections of data gathering are put together to reveal the state-of-the-art of the existing infrastructures of ODP around the world. The aim of the wide methodology is to deliver sufficient knowledge on the structures, components and tools, for example, API availability which enables various functionalities such as customisation, extensions, workflow, analysis and visualization, sharing and integration, and so on, to be carried out on the platforms. The investigations produce information on data harvesting, publishing, visualisation and dashboards, community services, some hints on stakeholders' adoption and popularity among user communities such as developers and third party stakeholders. In particular, this study sets a list of benchmark features against which the selected open data platforms (ODPs) features are reviewed and described. This approach enables a fair comparison between the selected ODPs on their individual level of advancement in technology and fit-for-use design.

In this desktop exercise, we cover a wide range of documents including books, papers, journal and article which have information open (government) data and open data platforms and portals. They provide several perspectives on the concept of open data and the practice of the concept, the barriers to adoption and the benefits it may offer to transparency and accountability, the implication or support that social media integration on the ODPs may offer to the effectiveness of the ODPs objectives. We evaluate the resource of ideas and theories that are contained in these textual materials, analyse them and summarise findings that are included in this report

Open Data Platform Survey - In this exercise, we examine the contents of the selected eleven ODPs enumerated in section 3.3 to reveal the current information disclosed about the functionalities and capabilities, tool and uses as well as the technology base. Basically, we applied two methods of data gathering in this exercise – first, is the desktop review by visiting each of the websites and reviewing the textual materials presented on the sites; searching online for more user community information such as blogs and comments on

issues and challenges or otherwise about the websites. Secondly, we acted the role of a platform user and approach the platforms from the point of view of the objectives of typical platform users hoping to make the best use of the resources on the platforms. By this method, we actually login in into each of the platform and attempted usage in order to study the back-end or other possibilities that would not normally be seen from just visiting the site through a browser. The objective is to collect more information to augment the desktop version so as to gain more robust information on the features and capabilities as well as the technological base of the platforms for the purpose of describing their state of the art. Data gathered form these exercises are assembled, analysed and summarised into Appendix 3 - General summary of ODP features.

Fieldwork and Workshops - Data gathering from field work involve two set of activities – workshop on open data and open data platform and interview of open data stakeholders. A number of workshops and stakeholder interviews are conducted across participating European Counties with each of them bring in experts, academia, industry specialist (IT expert, open data practitioners, representatives of governments and open data researchers) to brainstorm on the open data (platform) adoption challenges, solutions to the challenges and the suggestion of various groups of requirements necessary for consideration in the design of the next generation open data platforms so as to gain the advantages of improved government transparency and accountability through citizens participation and collaboration for decision-making to improve social services and community development.

Whereas the workshop data gathering take the approach of collective intelligence methodology to gathering useful data through collective thinking and suggestions, the interview section assumes a more personal point of view methodology to sampling opinions as it affects the individual stakeholder. From this exercise, we gather more ideas from the interviewee's perspective, his/her personal interests, challenges/frustrations, current areas of satisfaction and/or disappointment and were as areas he/she hopes to see improvements not only in the use of ODPs but also in the general practice of open data concept including government input and citizens' responsibilities.

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). 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 sematic 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[18] 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.

[18] http://www.w3.org/TR/vocab-dcat/

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[19] 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.

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[20] 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[21] and recline.js[22] 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.

[19] http://enigma.io/
[20] http://www.r-project.org/
[21] http://d3js.org/
[22] http://okfnlabs.org/recline/

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 tis 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 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.

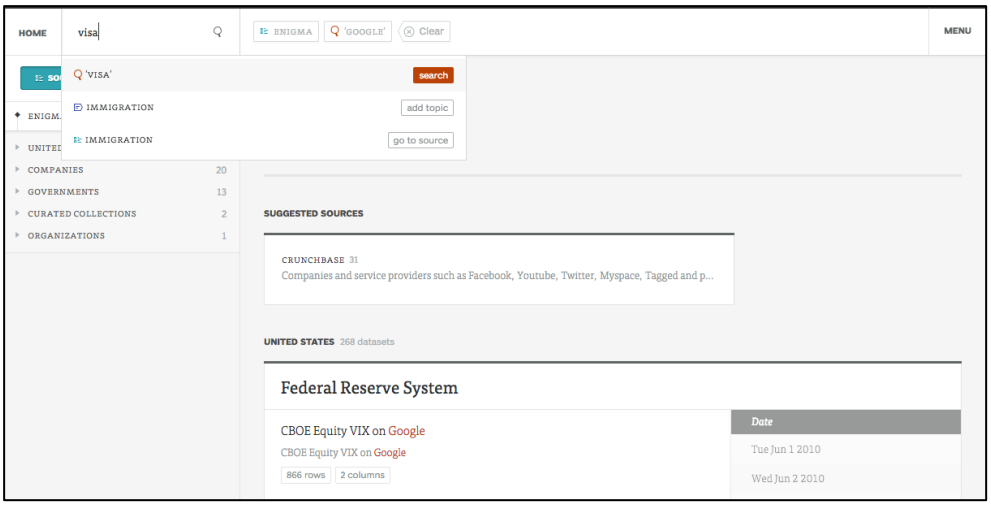


Figure 7: Enigma search user interface

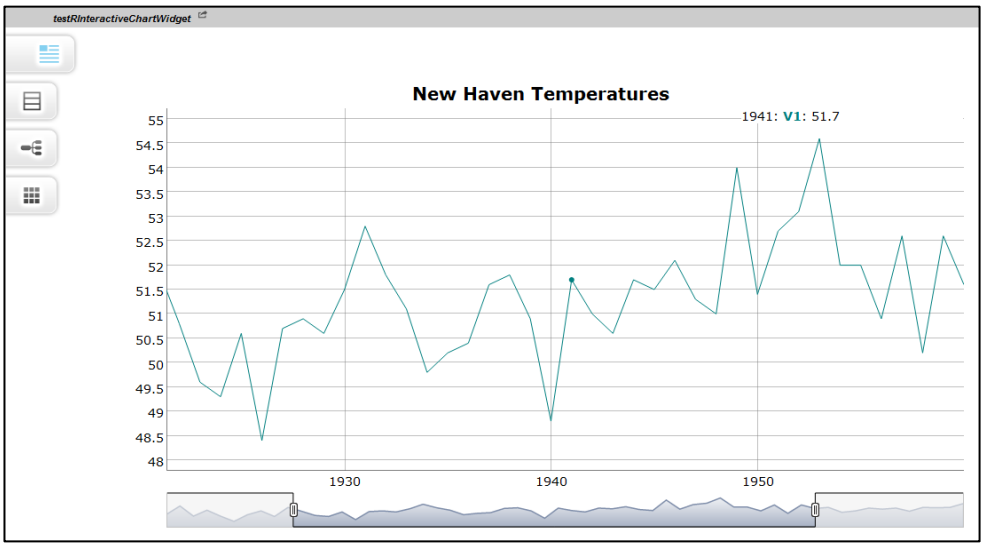


Figure 8: Interactive R chart in Information Workbench.

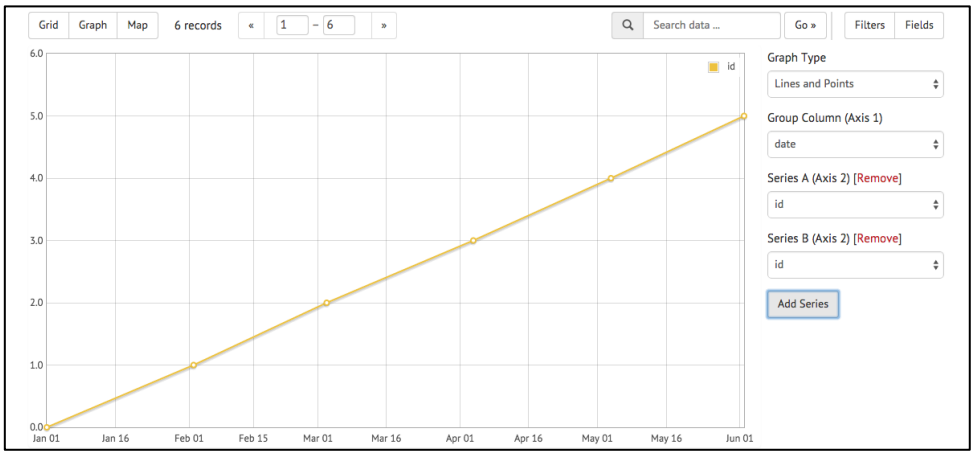


Figure 9: Recline.js visualization library used by CKAN

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 (Iemma 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[23]
- DKAN[24]
- Socrata[25]

[23] http://ckan.org/
[24] http://nucivic.com/dkan/

- PublishMyData[26]
- Information Workbench[27]
- Enigma[28]
- Junar[29]
- OpenDataSoft(ODS) [30]
- Callimachus[31]
- Datatank[32]
- Semantic MediaWiki[33]

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 (Iemma 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 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,

---

[25] http://www.socrata.com/
[26] http://www.swirrl.com/publishmydata
[27] http://www.fluidops.com/en/portfolio/information_workbench/
[28] http://enigma.io/
[29] http://www.junar.com/
[30] http://www.opendatasoft.com/
[31] http://www.callimachus.com/
[32] http://www.datatank.co.uk/
[33] https://semantic-mediawiki.org/wiki/Semantic_MediaWiki

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.

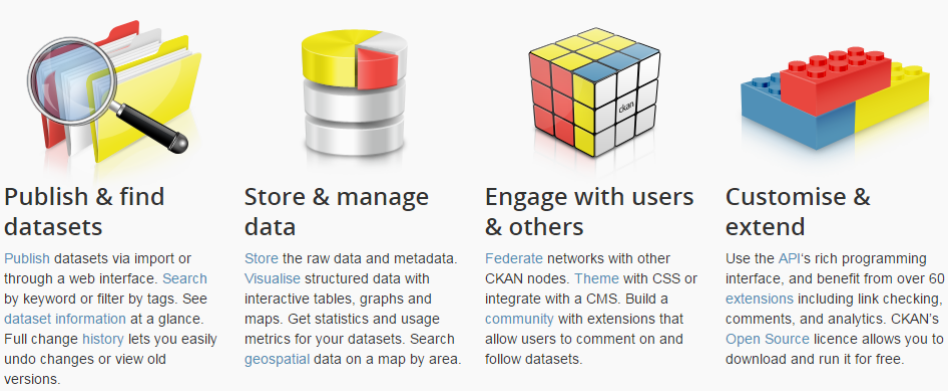


Figure 10: CKAN's main features - extract for ckan.org

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 3: Summary of CKAN features

1	CKAN
Features	Website literature review
Installed instances	CKAN has 116 well-known instances on the web and several other instances.
Metadata, Data and File Format Standards and Schemas	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.
Flexible search facility for datasets	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.
Social Media, Collaboration and Social Sharing tools	CKAN has many social media tools: Facebook, Google+, twitter, etc. for user to communicate & collaborate, to comments, share, RSS feeds, follow, & To-do extensions.
Dataset Publishing workshop	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
Harvesting, Federation and Cataloguing	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.
Extensibility mechanisms	Has up to 60 extension options for users; it's Open source, very extensible platform, has JSON API. Allows links to external datasets
Data Analysis tools	Administrative dashboard for members and data management but no special tools for data analysis
Visualisation tools	Basic visualization for tabular data and also by charting, mapping and imagery, etc.

Personalisation tools	Themable features – personalised settings for users' preferences.
Customisation tools	CKAN has customisable data harvesting models which support importing datasets from users' repositories. Customization using extension
Dataset licensing service	Licensing information can be added during the upload process
Accessibility	No special features related to accessibility
Technical Environment	Build using python programming languages with pylon web framework.
Others	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).

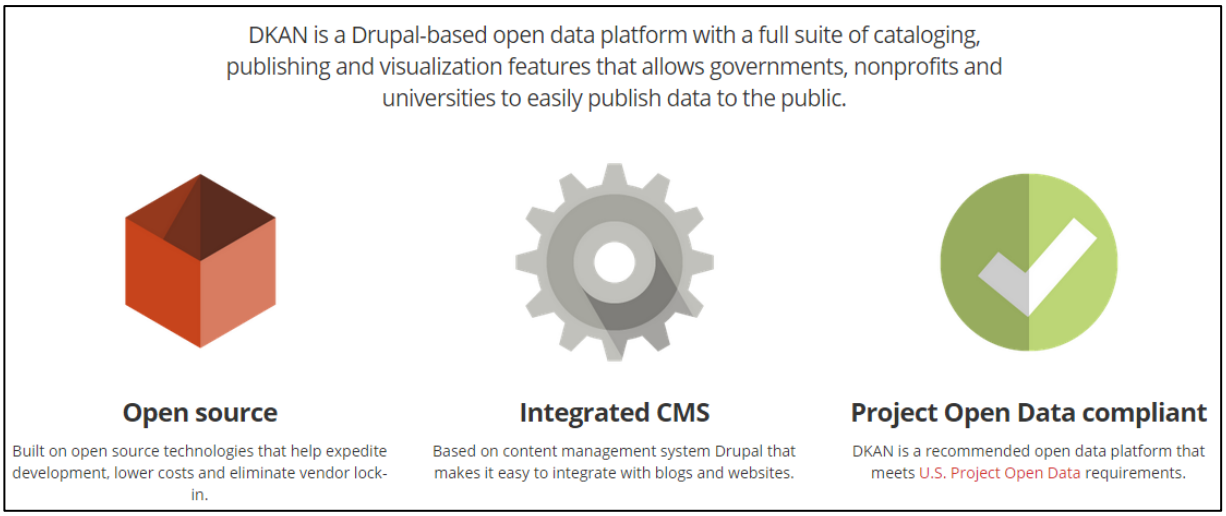


Figure 11: DKAN web Interface. Source: http://nucivic.com/dkan/

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)

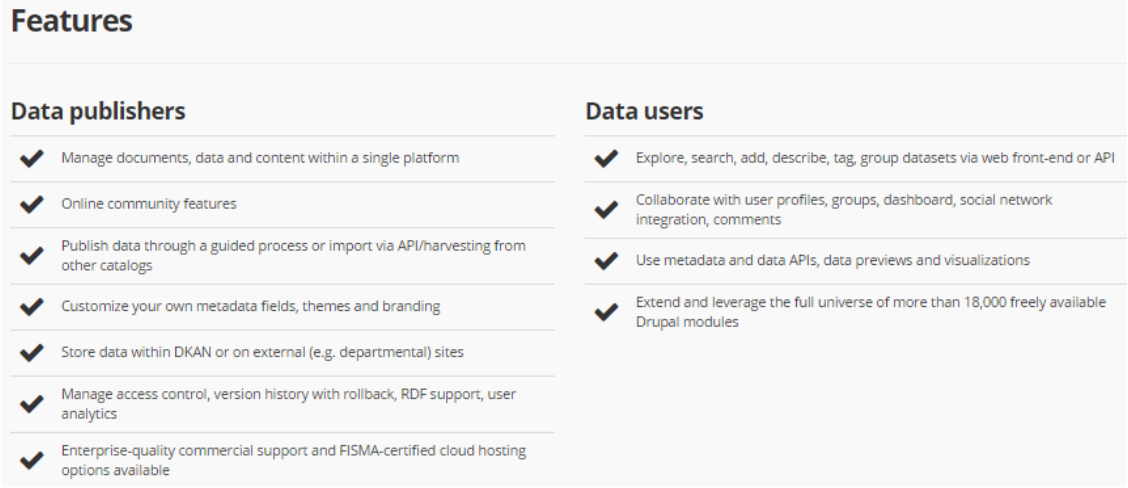


Figure 12: DKAN features in brief. Source: http://nucivic.com/dkan/

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 4: Summary of DKAN features

2	DKAN
Features	Website literature review
Installed instances	No good estimate available
Metadata, Data and File Format Standards and Schemas	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

Flexible search facility for datasets	The search facility is clearly presented; permits filtering by metadata to returns results with title and description DKAN provides search UI and allows filtering on metadata fields
Social Media, Collaboration and Social Sharing tools	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.
Dataset Publishing workshop	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
Harvesting, Federation and Cataloguing	DKAN has complete suite of tools for cataloguing and harvesting dataset.
Extensibility mechanisms	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
Data Analysis tools	No special data analysis functions or tools, support Google Analytics, Publishing maps with CartoDB.
Visualisation tools	Visualization features exist for users to display their dataset in reports but limited support
Personalisation tools	Theming is available for personalisation.
Customisation tools	CKAN can be customized with theming, JSON API and Drupal extension
Dataset licensing service	Licensing information can be added during the upload process
Accessibility	No build-in support for accessibility, but accessibility features could be added using Drupal accessibility modules
Technical Environment	Uses PHP based CMS Drupal
Others	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 13: 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 facetted 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 5: Summary of Socrata features

3	SOCRATA
Features	**Website literature review**
Installed instances	No estimate available
Metadata, Data and File Format Standards and Schemas	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
Flexible search facility for datasets	Searching data on the portal is by a robust search index & allows filtering.
Social Media, Collaboration and Social Sharing tools	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.

Dataset Publishing workshop	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.
Harvesting, Federation and Cataloguing	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.
Extensibility mechanisms	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.
Data Analysis tools	Capabilities for mapping and charting are available. Some basic Business Intelligence services. Library for working with statistical package R
Visualisation tools	Provides powerful tools for visualizing machine readable data in various formats. Visualizing geospatial data
Personalisation tools	Personalised sorting, auto-filtering to create a preferred view. The portal administration allow personalization of portals
Customisation tools	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.
Dataset licensing service	Licensing information can be added to dataset as metadata
Accessibility	Uses common best practices to allow accessibility
Technical Environment	Most of Socrata components are written using scala
Others	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:

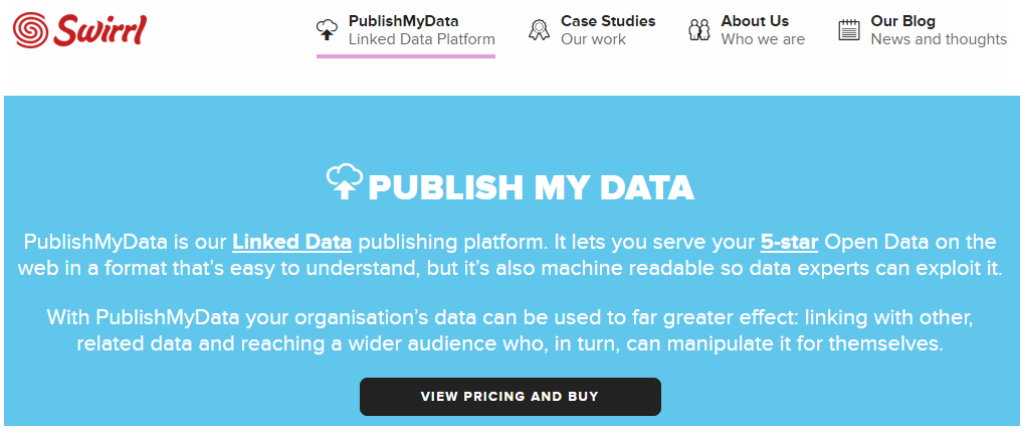


Figure 14: PublishMyData web interface. Source: http://www.swirrl.com/publishmydata

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 JASON, 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) while the online platform version (SaaS) can also be customised or integrated into other systems.

Table 6: Summary of PublishMyData features.

4	PublishMyData
Features	**Website literature review**
Installed instances	6 well-known instances

Metadata, Data and File Format Standards and Schemas	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. 1) Uses RDF files formats as input 2) Supports SPARQL querying 3) Metadata about the datasets is valuable in DCAT.
Flexible search facility for datasets	Provides SPARQL & other query tools for searching functionality but limited keyword search on catalogue data
Social Media, Collaboration and Social Sharing tools	Interactive documentation tools that support collaboration and sharing Provides no special features for sharing and collaboration
Dataset Publishing workshop	Use the standard of Linked Data from W3C to publish data with RDF. Converts file from CSV to RDF
Harvesting, Federation and Cataloguing	Provides datasets catalogue for users
Extensibility mechanisms	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. Community edition is available as Open Source project
Data Analysis tools	NA
Visualisation tools	NA
Personalisation tools	Grant administrative control of data publishing and customisation of platform
Customisation tools	Granting administrative control of data publishing and customisation of platform but limited
Dataset licensing service	Linking information can be added as metadata
Accessibility	Simple & easy to user interface and simple intuitive navigation on linked data.
Technical Environment	Developed using Ruby on rails
Others	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 15: 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 7: Summary Information Workbench features

5	Information Workbench
Features	**Website literature review**
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 sematic data model and API. Has connector to various data formats: CSV, XLS & database connections etc. that allows conversation to RDF.
Harvesting, Federation and Cataloguing	No specialized support for harvesting and federation

Extensibility mechanisms	Able to integrate dataset & link organisation together. Supports developers' community, via SDK for building apps. Allows extension and connectors
Data Analysis tools	Enables users to access, use, analyse and visualise & combine data in flexible manner. Supports R statistical package
Visualisation tools	Visualization of Twitter followers. Provides variety of visualizations widgets
Personalisation tools	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
Customisation tools	Develop and deploy custom apps. User interface is easy to customize, allow custom extension
Dataset licensing service	NA
Accessibility	NA
Technical Environment	Written using java and common web technologies
Others	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 16: 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 **Error! Reference source not found.**.

Table 8: Summary of Enigma features

6	Enigma
Features	**Website literature review**
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.

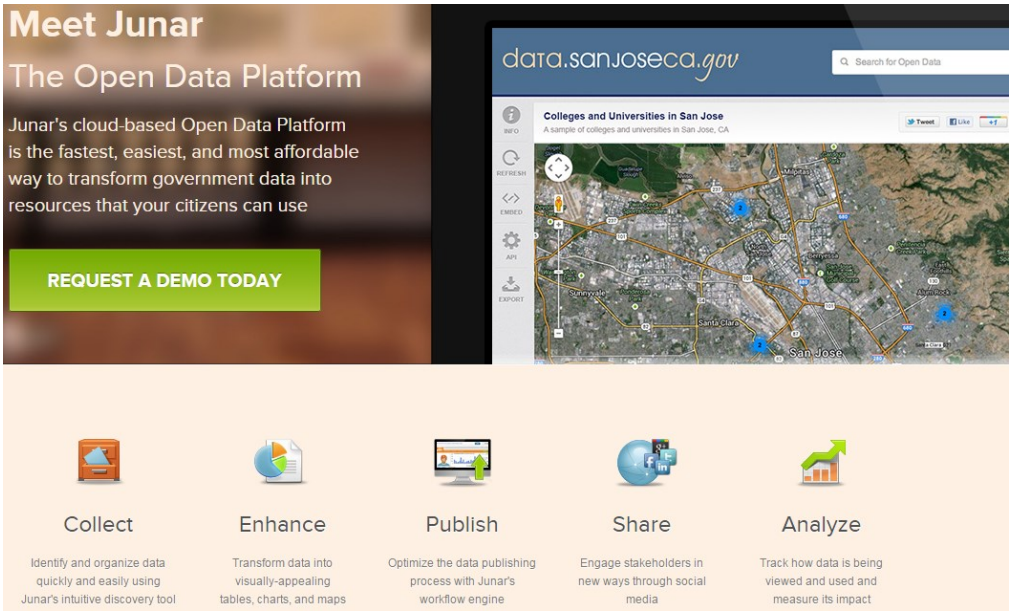


Figure 17: Junar web interface showing features. Source: http://www.junar.com/

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 chats 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 9: Summary of Junar features

7	JUNAR
Features	**Website literature review**
Installed instances	20 well-known instance
Metadata, Data and File Format Standards and Schemas	Supports data formats e.g. CSV, XLS, XLSX, KML and XML. Allow metadata to be attached to datasets
Flexible search facility for datasets	Junar categorises data & adds metadata to improve searchability. Limited search service
Social Media, Collaboration and Social Sharing tools	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
Dataset Publishing workshop	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.
Harvesting, Federation and Cataloguing	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.
Extensibility mechanisms	Can integrate the data directly into the user's site through a workflow mechanism that optimises data publishing. Limited extensibility.
Data Analysis tools	Analyses of data are limited.
Visualisation tools	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.
Personalisation tools	Possible
Customisation tools	Possible
Dataset licensing service	NA
Accessibility	No special feature related to accessibility
Technical Environment	Written in Java and Python
Others	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)

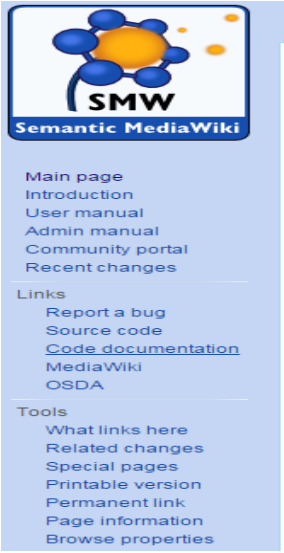


Figure 18: ODS web interface

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 10: Summary of ODS features

8	OPEN DATA SOFT
Features	**Website literature review**
Installed instances	38 well-known instances
Metadata, Data and File Format Standards and Schemas	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
Flexible search facility for datasets	Both Dataset Search API and Records Search API are provided for searching purposes. Text search, multi-criteria text search are possible on the platform.
Social Media, Collaboration and Social Sharing tools	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.
Dataset Publishing workshop	Hosting and admin – user management, user groups, cloud hosting, workflow, analytics and integration functionalities etc. Web based UI and workflow for publishing data
Harvesting, Federation and Cataloguing	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
Extensibility mechanisms	Limited extensibility
Data Analysis tools	Leverages data analysis to produce visualised interactive graphics but analysis at basic level
Visualisation tools	Powerful API publishing support data analysis & interactive visualisation in maps & chart & pictures. Analytics and imagery. Geo data visualisation.
Personalisation tools	Custom connectivity and data Federation; Data Processing – Geocoding, text transformation, joins, numeric operation and indexing. Allows UI customization
Customisation tools	Customisable GUI; Embeddable widgets though Limited customization
Dataset licensing service	Licensing information can be added to dataset
Accessibility	No special features related to accessibility
Technical Environment	NA
Others	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

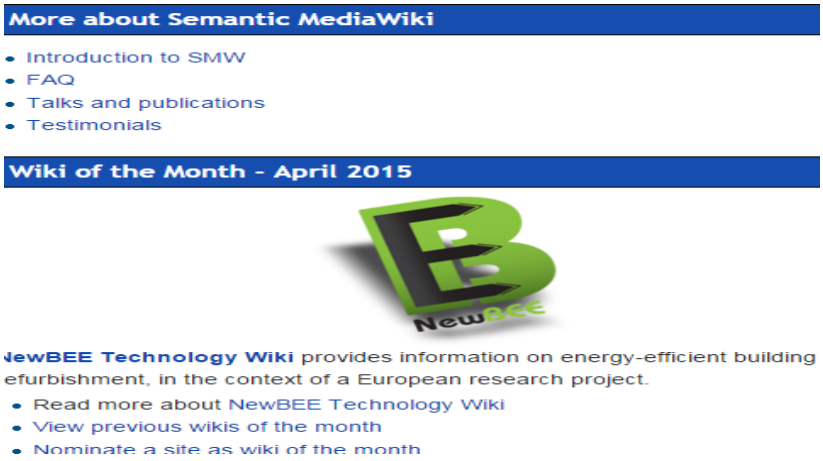


Figure 19: Callimachus web interface

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.

Figure 20: Features of 3RoundStones' Platforms. Source: http://3roundstones.com/products/

Table 11: Summary of Callimachus features

9	CALLIMACHUS
Features	**Website literature review**
Installed instances	Not available
Metadata, Data and File Format Standards and Schemas	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.
Flexible search facility for datasets	NA
Social Media, Collaboration and Social Sharing tools	Provides wiki pages for collaboration and sharing
Dataset Publishing workshop	Uses templates format for data collection and wiki pages for publishing & workflow
Harvesting, Federation and Cataloguing	For users, it makes data easy to create, view, and update. NA
Extensibility mechanisms	Simplifies integration of new data with existing data compared to using relational databases. Open source project
Data Analysis tools	NA
Visualisation tools	NA
Personalisation tools	Limited personalization
Customisation tools	Provides limited customization for various user groups; e.g. the Admin User group & the Super User Group.

Dataset licensing service	NA
Accessibility	No special features related to accessibility
Technical Environment	Java based application
Others	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 12: Summary of Datatank features

10	DataTank
Features	**Website literature review**
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

Social Media, Collaboration and Social Sharing tools	NA
Dataset Publishing workshop	Provides tools for ETL
Harvesting, Federation and Cataloguing	NA
Extensibility mechanisms	Open source project
Data Analysis tools	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
Visualisation tools	Creates a holistic view of data across departments. Reveal relationships, patterns and trends. NA
Personalisation tools	Limited personalization
Customisation tools	Limited customization
Dataset licensing service	NA
Accessibility	No special features related to accessibility
Technical Environment	PHP based application
Others	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

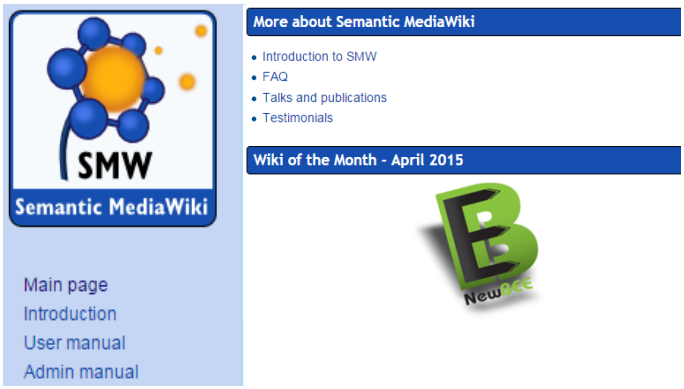


Figure 21: Semantic MediaWiki web interface

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 13: Summary of Semantic MediaWiki features

11	Semantic MediaWiki (SMW)
Features	**Website literature review**
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 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 in 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.

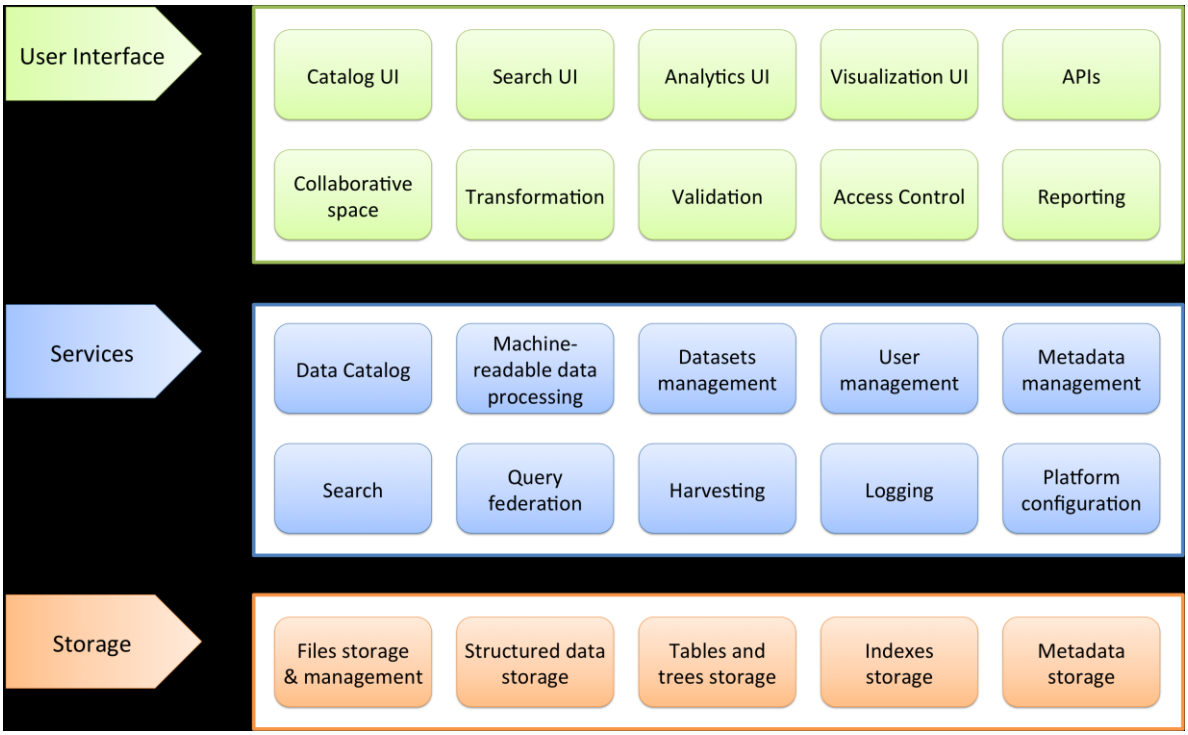


Figure 22: Architecture of Open Data Platforms

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.

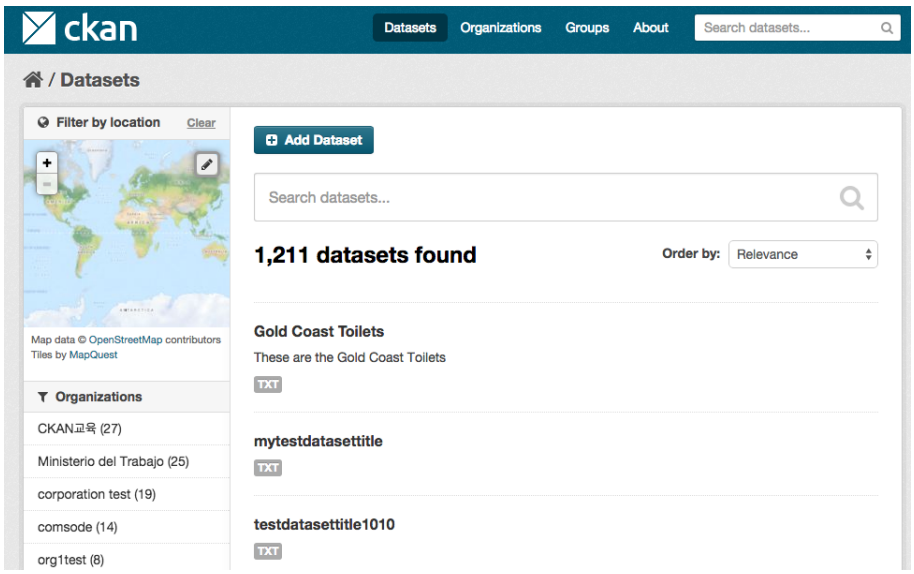


Figure 23: Screenshot of search and catalogue feature of CKAN

Component name	Role / description
Catalog UI	A user interface that allows browsing, exploration and discovery. Basic functionalities are category browing, dataset rating, data downloading or browsing by category, etc. It may provide additional information that facilitates discovery, e.g., featured datasets, additional labels, etc.
Search UI	A user interface that allows quering the collection of dataset metadata records. Basic functionality is the text search.
Analytics UI	Comprises any number of approaches to generating new insights from data: data mining, data fusion, spatial linking, statistical analysis, clustering, and so on. Usually a knowledge of these techniques is necessary for a user to derive value from this interface.
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	It provides reporting services throughout the system. It may display usage statistics, data summary, queries summary, datasets quality summary and so on

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) 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-aservice 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). 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 limitied 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 wrriten 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).

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

√* - 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 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 14: 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 interest	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 off	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 portal	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 existence	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 goal	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 data	Mediator/Consumer	Poor access to open data platforms	Accessibility	Resource constraints
Lack of sufficient broadband / bandwidth to successfully interact with Open Data	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 for data 'miners'.	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 15: Excerpt from Desired Features in Future Open Data Platforms

Information needs
Inventory of local business people – support local enterprise
Key indicators for my neighbourhood (social, crime, environment, health, etc.) for informed decision making
Local info of all kinds – planning, sports, cultural, commercial, social, councillors
Social and Collaborative Needs
Anonymity
Closed loop, share results of interactions & collaborations
Contact tools for finding PA, forums, public participation, network, social media interaction, twitter, facebook
Dataset rating & ranking, Calendar, wall style fast feedback, live chat, comments on dataset, blogs, collaborative editing, curating, adding metadata for dataset
Diversity of engagement – creativity, inclusion, new knowledge & value
Embed data for viral travel of data + its conversations
Expert facilitation
Live webcast with feedback, newsfeed for decision,
Mission/vision statement for discussion
Original data location – show paths to where it is shared
Prioritisation of data request based on needs/voting
Project management tool
Reward system, gamification, acknowledgement
Verification/traceability of account
Understandability, Usability and decision-making needs
Modelling and stimulations
Animations & interactive visualisation; Predictive analytics
Animations, pictures, browsing exploration experience
APIs,
Customisable Dashboards, personalisation
Data availability over several portable devices; Customised display – pull in from other platforms + layer data
Data integration
Data mining tools & analysis tools for information extraction to support decision-making
In-file data descriptors
Interactions, 'rate my service', submit suggestions on map + get feedback
Interactive graphical representations as transparency enhancing tools, promote easy reading, understandability, making sense of data
Linked data for comparison
Map + zoom Vs recovery
Map based search & queries
Metadata management
Modelling tools, layered maps
Personalisation – search with filter, especially with memory, notifications & updates
Polls and surveys
Public or anonymous profile options
Q & A mechanism
Question & answer, feedback mechanism monitored up-to-date
Scheduling services – identify what is logged, actioned or closed
Statistics under-pinning policies

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 16: ***Summary of Open Data Platforms***

	Standalone	Cloud Service	Online Service
Proprietary	○ Socrata ○ PublishMyData ○ Information Workbench	○ Junar ○ OpenDataSoft	○ Enigma
Open Source	○ CKAN ○ DKAN ○ Callimacahus ○ Semantic MediaWiki	○ DataTank	

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 standout 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 24: Perceived Barriers to Use and Adoption Open Data Platforms

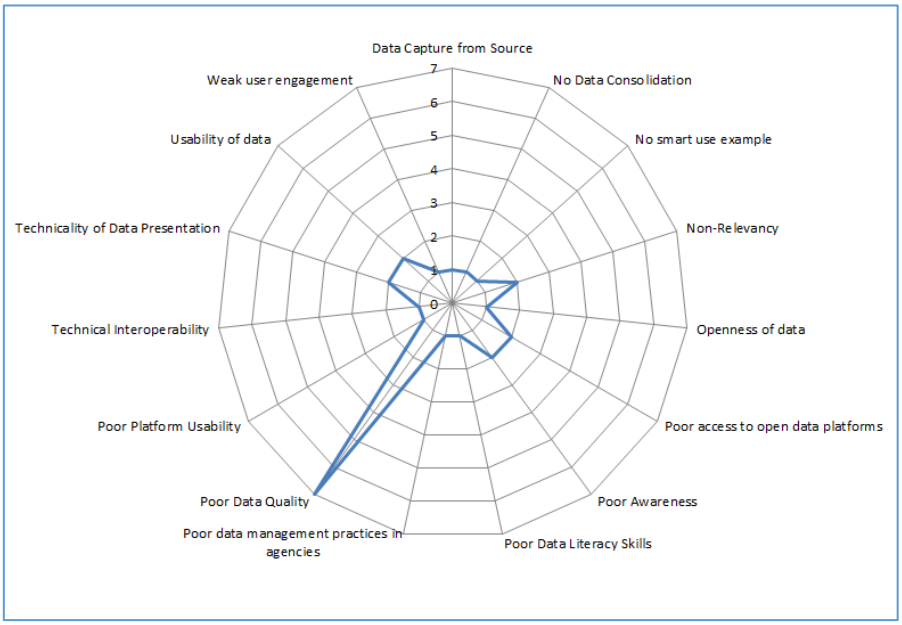


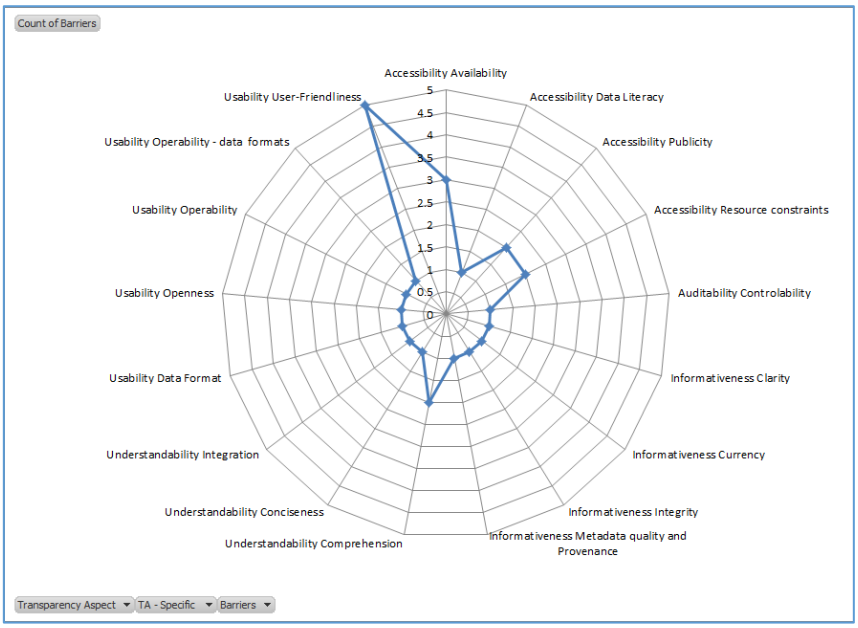
Table 17: Summary of Platform Features

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 25: Data Transparency attributes related to the Perceived Barriers



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 26 shows relative distribution of features across three categories.

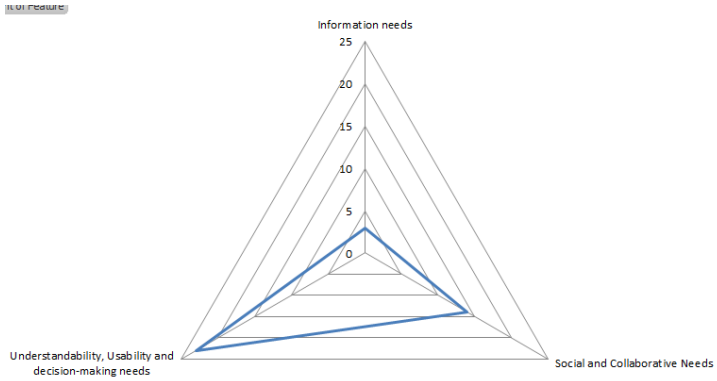


Figure 26: Number of Features in Desired Features Categories

Figure 27: 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 18: 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 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 standout 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 standout 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 including 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.

The deliverable will serve as input into the selection of the base platforms for innovation activities to be carried out in the Route-To-PA project.

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

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

Fishenden, J., & Thompsom, 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

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

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/

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

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

APPENDIX 1A: REPORT 1 (OD PROMOTER / RESEARCHER AT INSIGHT CENTRE FOR DATA ANALYTICS, NUI GALWAY)

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 [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:

7 (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 an 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 linking 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 at now aware 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.

8 (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 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** their 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 are 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 tool 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 _____ Date:__ 27th April 2015

APPENDIX 1A: REPORT 2 (OD MEDIATOR: EXPERT/PUBLISHER AT CENTRAL STATISTICS OFFICE)

ROUTE TO PA

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

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:

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: 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

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 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:

9 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):

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, -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 an 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 are just looking for a number. You need to publish your Data to let them find that number, that is, 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, 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, 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

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):

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 needs to be published. Once you setup the link to the data it needs to stay there over time.

---

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):

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.

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):

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. There 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.

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):

The CSO puts all the statistical data at StatCentral. Eoin is visiting schools, universities and gives a lot of talks. The questions are: who is aware that there is a portal cso.ie? Who is aware that statcentral.ie exists? Who is aware that data.gv.ie exists? The most people don'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 has 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 too 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!

---

10 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:

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?

---

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:

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 platform, 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.

---

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:

In relation to Open Data CSO is happy with 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? That is, National museum – how to publish their objects? How is the data around these objects available? Same story with castles, churches and 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.

---

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:

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:

1 1

---

6. 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.

---

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:

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 are the pieces of data that 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.

---

8. 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.

Interviewee's signature against responses recorded _____ Date: 24/4 15 ____

APPENDIX 2: REPORTS OF DUBLIN WORKSHOP ON ODP

APPENDIX 2A: OPEN DATA BARRIERS AND THEIR RAKING

Conflict / Cooperation	Score	Motivation	Score
Hostility toward monitoring and benchmarking via open data	1	Lack of interest in using open data for any purpose ('Sure why bother?')	2
Conflict between wanting to share data and the data being used as criticism	9	Failure to understand the benefits that Open Data can offer	5
Conflict between different govt. agencies regarding what should be transparent and accessible (relates to different code of ethics and incoherent value systems in different organisations)	2	Unwillingness to educate oneself as to the benefits of open data	0
Perceived lack of government credibility	0	Unwillingness to equip oneself with the skills to utilise open data	0
Ignorance towards research/expert opinion	1	Lack of promotion / marketing surrounding open data initiatives offering motivation to 'get involved'	3
Hostility towards data release as it is seen as a source of power	2	Lack of public drive to get government to change	??
Lack of cooperation between government and public	0		
Open Data vs Eircodes (postcodes)	2		
Conflict and lack of progress due to contrary interests.	??		
Conflict between DPER position on supporting OpenData and establishment of new Eircodes. No free look-up file from eircode to statistical geography (SA or ED) as per NI and UK. Major opportunity for OpenData in Ireland lost.	??		
Total	**17**	**Total**	**10**

Services and resources	Score	Skills and training	Score
Shortage of technical resources to collect data	3	Inadequate technical expertise to produce data in a usable format	0
Finding data; 100's of datasets, in danger of becoming a data dump, needs improved standards for cataloguing and describing data, also linking relevant datasets	??	Lack of training to go about finding data that is relevant for the purpose required	5

Reliability of data feeds and keeping them updated; old data is gone off	3	Fears of criticism (by govt organisations) from the public and inadequate support and training to field and reply to these concerns	4
Poor service design and management	2	No minimum skill-set is defined with which the archive is comfortable to use	1
Information spread out over multiple organisations, lack of one portal	2	Inability to interpret data might be seen as a permanent problem	0
Poor information management	1	Lack of educational material to acquire minimum skill-set.	1
Demand for cleaned data and data control prior to being released in an OpenData portal	??	Lack of skills/education to utilise open data	1
Difficulty finding potential data dump rather than standards for cataloguing & linking data	2	Lack on freely available software that users can download, i.e. Arc GIS vs Q GIS	0
Metadata problems	1	Lack of skills / education to utilise open data	??
Available open dataset are not relevant or 'speaking' to people's interests	6	Lack of accredited open data training courses	0
Guaranteeing / cleaning personal data from data sources (historical/ legacy)	1	Citizens may need to be computer literate to gain access to data	0
		Lack of understanding of technical requirements for publishing open data within organisation	??
Total	**21**	**Total**	**12**

Communication / Accessibility	Score	Government / Organisational	Score
Data on screen may be displayed in a technical way or use unfamiliar technical language	??	Failure by government departments to advertise that data is available to the public	0
Citizens may not always have up to date browsers on their computers	??	Resistance to releasing / publishing data in open format	0
Minimal publicity about data available leading to lack of awareness of its existence	??	Failure to understand the organisational benefits of releasing open data	3
Data is in a dense form and requires design input to make it accessible	2	Fear of how transparency via open data might affect organisation	8
Lack of information about the circumstances of data production	??	A perception that Open data is simply 'something that Governments do' and not private sector industry	0
Lack of user-friendly file-formats	2	Lack of belief that Governmental Open data is reliable data	1
Lack of user-friendly interface	??	Fear of loss of data ownership once data is released in open format	1

Lack of engaging activities/information for those users who arrive to a page without a clear goal	??	(Perception of) Inadequate organisational legal frameworks to permits data to be released in open format	0
Lack of examples available for smart use of open data	2	Lack of in house knowledge and skills to publish data in open format	4
Poor quality of data, right formats to the right audience e.g. spreadsheets for 'tourists' and feeds/API for data 'miners'.	4	Unwillingness to change current data reporting practices	0
Useability; need preview, mapping, visualisation, multiple data layering	??	Lack of development of wish list of potential open datasets that can be used to address societal challenges in Ireland. What data do we need from Government?	??
Lack of access to necessary software / hardware to utilise Open data	1	Fear of causing panic or data being misread	1
Limited usability of Open data, preview, visualisation, layering	1	Lack of open data from HSE such as health facilities in Ireland	??
Lack of sufficient broadband / bandwidth to successfully interact with Open Data	0	Inadequate institutional capacity to provide OD services, to develop standards, to provide expertise	0
Lack of data abstraction (info graphics/ data stories)	3		
Level of openness and licences for use in commercial remit			
Total	**15**	**Total**	**18**

Cost	Score	Privacy and security	Score
The cost of accessing data may be prohibitive	4	Personal information accessed by public can lead to data protection infringement	3
Inadequate finances to fund the sustained collection and sharing of open data		National security issues as a result of the release of sensitive information	0
Difficult to make money from open data (furthering resistance of government)	2	Culture of secrecy	1
Open data is low priority for government to implement, due to uncertain ability to generate revenue.	0	Highly selective groups allowed access to certain types of data	1
Lack of understanding of actual real cost of publishing open data within organisations. Open data does not mean 'free' and comes with significant cost		Dilution of information available to the public	0
Total	**6**	**Total**	**5**

APPENDIX 3: 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
Personalisa-tion	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 personaliz-ation	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