From Public Sector Information Catalogue to Productive Data: Defining a National Information Infrastructure

Johanna Walker[1], Jacqui Taylor[2], Leslie Carr [1]

[1] University of Southampton, Southampton, UK
{j.c.walker, lac} @soton.ac.uk,
[2]Flying Binary, London, UK
jacqui.taylor@fliyingbinary.com

Abstract. This position paper briefly reviews some of the limitations of open government data in the UK and describes a structure for the National Information Infrastructure (NII) as a Web Observatory for public sector information. Based around a three-part data structure of Core Reference Data, Subject Data and Thematic Data, set within a specified framework of tools and services and supported by sound governance, it will enable improved delivery and support effective use of government open data and beyond.

1 Introduction

In order to facilitate the posited benefits of open government data, it is vital to provide, 'well structured and curated results from the *web of data* for the purpose of deriving insights' (Brown, Hall and Harris, 2013). Despite this, in many cases government catalogues have been populated without a complete and strategic understanding of what datasets government holds and consequently publication has often been ad hoc (Cabinet Office 2013, Shakespeare 2013). Massive government data catalogues may be largely populated with a single type of dataset. Erickson et al (2013) found that of over 400,000 open US government datasets only 1% were not geolocation data. Government data repositories may also suffer from minimal interoperability between their own and other data sets (European Commission 2011a, Plu and Scharffe 2012).

1.1 Data Catalogue Issues in the UK

Despite ranking highly in various open data assessments such as the Open Data Index, the curation of coherent open government data in the UK has its own specific challenges. There is no central list that identifies all public sector bodies making it challenging to locate the associated datasets (ODUG 2015). There is no coherent classification of public sector organisations and services, making them difficult to describe. Essentially, such lists currently require compiling on a search by search basis. A "canonical source" is required (Brown, Hall and Harris 2013).

Two other issues have been identified by ODUG (2015)[1]. The first is that current solutions fail to define a comprehensive picture of how datasets relate to each other. Secondly, the gap between the repositories and data owners reduces the effectiveness of governance, resulting in the deterioration of existing open data over time.

2 The National Information Infrastructure

These issues can all be addressed with a National Information Infrastructure (NII). This has been envisaged in different ways, but can usefully be conceptualized as a Web Observatory. This comprises the Data, an assemblage of data assets created by public sector organisations in the course of delivering public services, and the Framework, tools and services to maintain data and make it accessible. This creates an agreed architecture for the publication of key datasets and protect them for use by all types of organisations and by citizens. This process will become more robust, consistent and reliable, and individual datasets will become interoperable, offering new ways of analysing the work of government. The NII will set publication standards for key datasets, creating a benchmark for future releases. The structure outlined below builds on a basis of Core Reference Data (Shakespeare, 2013) - universal data to connect datasets, through Subject Data - identifiers applicable to multiple datasets but not all, to Thematic Data describing specific areas such as Health or Construction). This tri-part hierarchy not only reduces the 'vastness and incongruence' of the data but enables better use and analysis (Gloria et al 2013). It also provides the 'connection beyond the query' (Brown, Hall and Harris 2013).

.

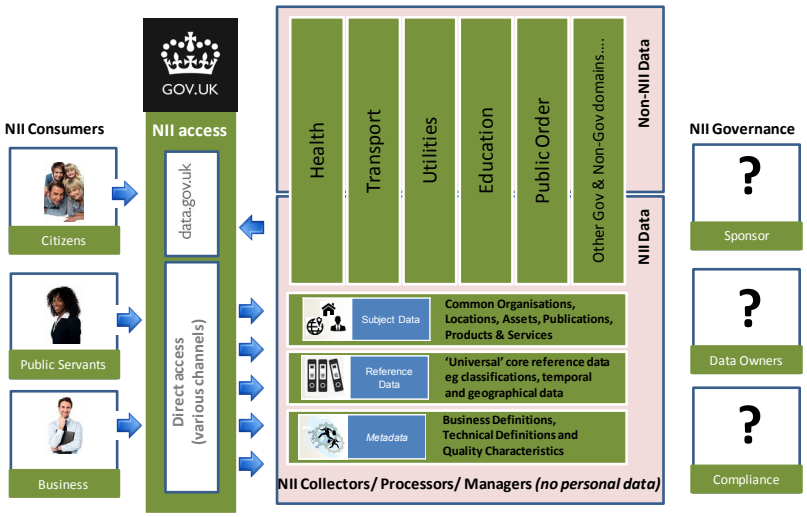

Fig.1. Architecture of National Information Infrastructure

[1] The Open Data User Group (2012-2015) was an independent body that advises the UK Cabinet Office Transparency Team on dataset release.

The NII data itself contains only those datasets that are key and useful to data owners, collectors, publishers and consumers. The NII framework specifies metadata, governance over schemas, release schedules and data quality and delivery services. Although carefully specified, the NII has the flexibility to expand over time and should also begin to include performance data (statistics) or transaction/event data, ideally captured in real time. Unlike many Web Observatories, an NII calls for a strong political and public sector commitment and independent oversight. Accordingly, the issue of governance will have greater importance than in other observatories which may only require support from a single source.

2.1 The Horizontal Data – Linking and Enabling

Enabling the successful exploitation of data are the foundations of the Core Reference Data, Subject Data and metadata (Fig.1). Open Core Reference Data such as classifications (e.g. types of places, organisations, products and assets), Open National Address Data and Open Geospatial (mapping) will allow the cross-referencing and analysis of multiple datasets, that are currently siloed, on a non-personal basis. Subject Data is a less powerful set of identifiers common to several thematic areas. This will enable analysis of tightly-focused, cross-thematic questions. Establishing these horizontal datasets is the priority activity for the creation of the NII.

2.2 The Vertical Data – Sectors in Depth

These demand-led collections of data introduce a level of flexibility to the NII. Each thematic area requires analysis to determine how its data should be defined and structured to best fit into the overall schema. This means detailed examination of the key infrastructure and the services delivered in each sector, identification of the most important components and designing how this data should be integrated to support, but not perform, analysis. Thematic data also allows for interaction with other datasets in the same theme and extends the usefulness of the data beyond the PSI and the UK. Related data such as Event/Transaction data or Performance Information might include datasets from multiple sources; PSl, other organisational or private data. It might at first appear that thematic organisation simply replicates the siloes of the current situation. However, within the NII, linkage between themes is facilitated through the Core Data, whereas currently they are difficult, or impossible, to link.

3 Conclusion

The value of an NII comprising Core Reference Data was accepted by the previous UK government (Cabinet Office 2013). However, for an NII to meet the needs of the public sector, citizens, organisations, academia and more, a greater depth of datasets in a variety of sectors and tools and services must form an essential part. To provide this, there must be an underlying principle of ‘publish by default’. There is also more

work to be done on charging, licensing and curation barriers, as well as on how best to integrate local government data and the need for an overall governance mechanism.

Acknowledgments. The Open Data User Group (UK) first defined the National Information Infrastructure as outlined in this paper.

References

1. Brown, I., Hall, W., Harris, L.J.: From Search to Observation. In: WWW 2013 Companion, Rio de Janiero, Brazil. May 13–17 (2013)
2. Brown, I., Hall, W., Harris, L.J.: Towards a Taxonomy for Web Observatories. In: WWW 2014 Companion. Seoul, Korea. April 7-11 (2014)
3. Cabinet Office Policy Paper: National Information Infrastructure: First Interation. October. London. UK. (2013)
4. Erickson, J.S., Viswanathan, A., Shinavier, Y., Hendler, J.A.: Open Government Data: A Data Analytics Approach. In: IEEE Intelligent Systems Vol: 28 (5) pp.19-23 (2013)
5. European Commission: Open Data. An Engine For Innovation, Growth And Transparent Governance. Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions. Brussels. (2011)
6. Gloria, M.K., McGuinness, D.L., Luciano, J., Zhang, Q.: Explorations in Web Science: Instruments for Web Observatories. In: WWW 2013 Companion, Rio de Janiero, Brazil. May 13–17 (2013)
7. Open Data User Group: The National Information Infrastructure: Why, What and How. 3rd http://data.gov.uk/blog/national-information-infrastructure-nii. (2015)
8. Plu, J., Scharffe, F.: Publishing and Linking Transport Data on the Web. In: Proceedings of the First International Workshop On Open Data, WOD-2012. Nantes, France. May (2012)
9. Shakespeare, S.: An Independent Review of Public Sector Information, London. UK (2013)