International Journal of Geo-Information

MDPI

Article

Data Autodiscovery—The Role of the OPD

Adrian J. M. Cox *, Andrew J. Milsted and Christopher J. Gutteridge

Research & Innovation Services, University of Southampton, University Road, Highfield, Southampton SO17 BJ, UK; A.Milsted@soton.ac.uk (A.J.M.); cjg@ecs.soton.ac.uk (C.J.G.)

* Correspondence: A.J.Cox@soton.ac.uk

Academic Editors: Constanze Curdt, Christian Willmes, Georg Bareth and Wolfgang Kainz

Received: 29 January 2016; Accepted: 18 September 2016; Published: 22 September 2016

Abstract: The importance of open data and the benefits it can offer have received recognition on the international stage with the signing of the G8 Open Data Charter in June 2013. The charter has an early focus on 14 high value areas, including transport and education, where governments have greater influence. In the UK, we have seen the funding of the Open Data Institute (ODI) with a remit to support small and medium sized enterprises (SMEs) in identifying benefits from using open data, whereas, within HE, open data discussion is in its infancy although is acknowledged as a sector challenge by the Russell Group of universities. There is an evident need for the academic community to influence the adoption of applications using linked open data techniques in data management and service delivery. This article introduces the concept of "data autodiscovery", highlighting the role of the Organisation Profile Document (OPD) and its contribution to the early success of the UK National Equipment Portal, equipment.data, along with discussing the need for greater dialogue in linked and open data standards development.

Keywords: open data; linked data; autodiscovery; RDF; data profile; schema mapping; data structures; web crawler; organisation profile document; OPD

1. Introduction

Establishing harmonised vocabularies and interoperability in the data management landscape is becoming an increasing requirement with the need to deliver increased efficiency in management and reporting along with greater value added benefits to the wider community from the published data. The aim of "Linked Data" in this area of web and Internet science is described as "enabling people to share structured data on the web as easily as they can share documents today" [1]. More recent application and wider discussion of Linked Data techniques is further advancing the value and application of Open Data. A more commonly used term is "Linked Open Data" (LOD) [2] which can be best defined as data structured using linked data techniques and published using methods making it as accessible for machines as it is humans. If LOD is to generate its optimum value it should not only be published in accordance with the "five stars of open data" [3] but should also be easily machine discoverable.

In a landscape with over 1300 providers, the UK Higher Education (HE) sector has hundreds of reporting obligations, many statutory, often leading to duplication in data capture and management, and in many instances, the use of a range of different information systems generates interoperability challenges. For many institutions each new report, requires a considerable element of manual input in the compilation of a new dataset. Although initiatives such as the Higher Education Data and Information Improvement Programme (HEDIIP) [4], funded by the Higher Education Funding Council for England (HEFCE), aim to promote a new data landscape, there is still a need to identify and manage new standards underpinning development. Such initiatives come with fresh challenges and questions, for example, which technology and standards will deliver the most effective and efficient

approach? It could be argued that with appropriate standards, linked open data based systems can deliver flexible cross-institution, cross-sector infrastructure that will enable greater value from data aggregation and reporting.

Throughout 2014–2015 Universities UK (UUK), working with the Open Data Institute (ODI) [5], hosted a series of workshops aimed at improving the understanding of open data across HE and promoting greater application of linked open data approaches within institutional data management. Attending a number of these workshops the equipment.data team noted the challenges faced, both in establishing sufficient knowledge of systems and perception of value in publishing using an open data approach.

The launch of the UK National Equipment Portal, equipment.data [6], in April 2013, introduced the application of linked open data technology in the delivery of a web based data autodiscovery service. Essential to this process is the publishing of an Organisation Profile Document (OPD) [7]. The OPD is a machine readable Resource Description Framework (RDF) document embedded in an institution's website containing the organisation's full name, homepage, logo, dataset location, license and contact information for open access datasets. Unlike the process of data discovery in many current data aggregation systems, e.g., CKAN, the OPD removes the need for manual capture of data locations by the aggregator. The ability to autodiscover data locations should also compliment the future development of data aggregation services using proprietary systems further enhancing their data discovery process. The OPD is an essential component of the autodiscovery process used by National Research Equipment Portal, equipment.data. Its development has the backing of UKRI (formerly RCUK) as its preferred medium for national equipment.data sharing with the service now endorsed as strategically significant by HEFCE, recognising the infrastructure's potential contribution to a future more "open" Research Excellence Framework (REF).

Equipment.data has demonstrated a linked open data infrastructure can be implemented at a sector-wide scale and in the process has established the foundation components for wider data sharing. The use of linked open data in the data management landscape is growing enabling new approaches to data capture from a range of formats (CSV, Excel, JSON, RDF Documents) and publishing patterns (APIs, data catalogues), webpage-embedded data, .xls and JSON exports from bespoke system Application Programming Interfaces (APIs).

2. How Did equipment.data Evolve?

The development of equipment.data was funded by EPSRC in response to the need to improve visibility and utilisation of UK HE research equipment following the Wakeham Review of Efficiency in HE [8]. A simple and easily reproducible tool allows discovery and aggregation of UK research equipment databases into a single searchable portal. The development of equipment.data builds on a partnership between a number of UK universities, primarily building on outcomes of the UNIQUIP Project, Defining standards for the publication of research facilities and equipment data [9].

Applying the process used by equipment.data for the aggregation of published research equipment (Figure 1), autodiscovery of data requires just four key components:

1. An authoritative list of organisation homepages from whom you wish to capture the data, e.g., .ac.uk list from the JANET HE Network [10], used by equipment.data service, or via the use of a bridge identifier system such as ISNI [11] which provides a method of linking to a dataset of selected web domains. The website opd.data.ac.uk also contains a list of HE OPDs listing the web homepages.
2. Auto-discovery and aggregation software hosted by the requesting organisation. The tools enabling equipment.data to discover and aggregate data are published in a code repository [12].
3. A requirement that all organisations publishing data host an OPD enabling machine/autodiscovery of their data profiles, and

4. The data is managed to a required standardised profile, e.g., UNIQUIP, ORCID, Research outputs metadata profile developed by the RDA Metadata Standards Working Group [13] and ideally managed through a standards organisation, e.g., The Consortia Advancing Standards in Research Administration Information (CASRAI) [14] or World Wide Web Consortium (W3C) [15].

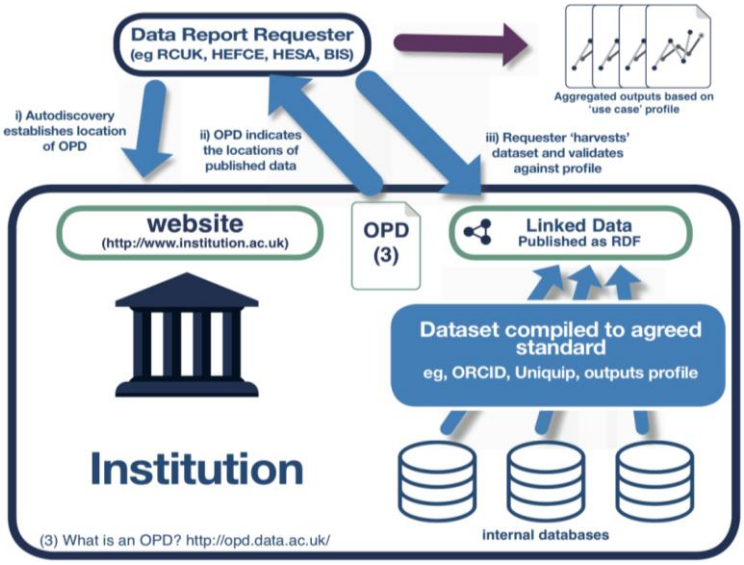

Figure 1. Simple model for data discovery using the OPD. Source: Abstract from poster submitted to Research Data Alliance 5th Plenary, San Diego [16]. Copyright: Adrian J. M. Cox, University of Southampton, 2015.

3. The Importance of the OPD

The OPD, including the associated embedded link in the home page, is the key enabler to the process of data autodiscovery, enabling machine discovery of data locations rather than current manually entered methods of dataset discovery used in many data aggregations, e.g., requiring notification of (OAI-PMH) end-point locations to a data aggregator. The OPD provides a formal, machine-readable, managed description of the organisation, stating what is published and the location/s of the data (the catalogue of datasets). It provides essential organisational information that will verify who it is, e.g., the organisation ID, official name, organisation type, official logo and geographical location. A fundamental feature is the trust that can be placed in the data found via the OPD, similar to that from finding information on the top level web pages of an organisation website.

For the equipment.data project to reach its goal of a fully sustainable system, it required a method of updating sources as efficiently as possible with minimal or no human intervention. To encourage adoption of a sustainable method of contribution the service established a compliance rating system [17] with gold, silver and bronze ratings to indicate to what level each contributing institution's data input is sustainable (Figure 2).

By publishing a fully autodiscoverable "Gold" compliant OPD any changes to data, which can include an institution altering its logo, to moving its data source from one system to another, would be reflected on the OPD. The ideal situation for data discovery services is that all institutions will be operating to the gold compliance rating, using a fully autodiscoverable OPD, therefore no human intervention is required from either the contributing institution or the discovery service in updating information as it will be automatically identified by the OPD.

As wider use of the OPD increases the challenge will be establishing appropriate ownership and governance of the OPD within organisations. It may be logical for this to be the marketing and communications department who typically will be responsible for an organisation's website (home

page) and could therefore manage the OPD content and/or link to the OPD. Due to the focus on research equipment data, the equipment.data service team mainly worked with staff from research support offices and IT departments. However, as more links to structured datasets are established and the use of the OPD extended, ownership could become the responsibility of marketing and communications or IT departments. To enable decisions around governance the sector will require greater confidence in this emerging technology. To support this requirement the aim is to establish a W3C Community Group engaging the sector in future development of the OPD [18].

	Bronze	Silver	Gold
Data is on the internet and in an acceptable format.	✔	✔	✔
Description of dataset is provided by a remotely hosted OPD		✔	✔
The OPD is discovered via autodiscovery.			✔
The OPD/dataset has a recognised and supported open licence (e.g., CCO, ODCA or OGL)			✔

Figure 2. Compliance rating applied to data discovery.

4. Structure of an OPD

The OPD uses RDF to describe the organisation in a machine readable form referencing many well established standard terms and vocabularies. The Core information uses OpenOrg, Dublin Core, W3C standards and FOAF RDF vocabulary. In doing so the OPD avoids defining new terminology requiring management and adoption within a new or existing standard. It is anticipated that any datasets listed on an OPD would be published to an agreed profile/standard, e.g., Research outputs locations meeting the OAI-PMH standard.

An OPD is split into two distinct sections; the first being the basic structure [19], the "core" information, describing the organisation, the second is an extendable component describing the datasets the organisation is publishing. Essentially the second component is a "catalogue" of discoverable open datasets available in defined data/application profiles. It will provide the dataset locations, e.g., URLs, a contact for each dataset and the license applicable to its re-use.

The recommended minimum data within the core OPD information includes the organisation URI, parent or sub-organisations, geographical location and primary contact information. The document is typically in the Turtle format which allows an RDF document to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes (Figure 3).

The method used to enable autodiscovery of the OPD requires a link in the organisation homepage header (Figure 4).

This link in the html header provides the location of the OPD enabling discovery programmes, "web crawlers", to interrogate the OPD and harvest data meeting the criteria set in their query. What the OPD provides web crawlers is a machine discoverable authoritative catalogue of LOD i.e., data and locations for data in defined "data profiles", e.g., The UNIQUIP Data Publishing Specification used by equipment.data, therefore making data discovery significantly more efficient and fundamentally adding value to the data enabling standardised datasets to be easily aggregated.

If a change to an organisation's html home page header is not possible the discovery programme has been developed so that the .well-known [20] method can be used. This method uses a specific URL from the organisation's homepage to link to the profile document, e.g., if the homepage is http://www.example.ac.uk then http://www.example.ac.uk/.well-known/openorg should serve (or redirect to) the OPD.

```
@prefix owl:      <http://www.w3.org/2002/07/owl#>.
@prefix foaf:     <http://xmlns.com/foaf/0.1/>.
@prefix oo:       <http://purl.org/openorg/>.
@prefix dcterms:  <http://purl.org/dc/terms/>.
@prefix geo:      <http://www.w3.org/2003/01/geo/wgs84_pos#>.
@prefix skos:     <http://www.w3.org/2004/02/skos/core#>.
@prefix org:      <http://www.w3.org/ns/org#>.
@prefix xtypes:   <http://purl.org/xtypes/>.
@prefix lyou:     <http://purl.org/linkingyou/>.
@prefix vcard:    <http://www.w3.org/2006/vcard/ns#>.

<> a oo:OrganizationProfileDocument ;
      dcterms:license <http://creativecommons.org/publicdomain/zero/1.0/>
;
      foaf:primaryTopic <http://id.example.ac.uk/> .

<http://id.example.ac.uk/>
      a org:FormalOrganization ;
      skos:prefLabel "The University of Example" ;
      skos:hiddenLabel "Example" ;
      skos:hiddenLabel "Example U" ;
      vcard:sortLabel "Example, University of" ;
      vcard:tel <tel:+441234567890> ;
      foaf:logo <http://www.example.ac.uk/example-logo.png> ;
      foaf:homepage <http://www.example.ac.uk/> ;
      owl:sameAs <http://id.learning-provider.data.ac.uk/ukprn/12345678> ;
      owl:sameAs <http://dbpedia.org/resource/University_of_Example> .
```

Figure 3. OPD Basic core structure.

```
<link rel="openorg" href="http://www.example.ac.uk/profile.ttl" />
```

Figure 4. Home page html header link.

5. The Need for Consensus Driven Managed Data Profiles

Beyond the aggregation of equipment data and the institutional URI structure "Linking you" [21] extending open publishing and aggregation of data in structured forms will require mechanisms for managing and agreeing data profiles. These managed data profiles will be required if meaningful data aggregations are to be achieved. There is evidently further potential to advance the adoption of data autodiscovery, exploiting the current growing UK HE infrastructure in the aggregation of other datasets where there is consensus and/or an agreed profile, e.g., research outputs metadata through OAI-PMH, offering significant improvements to the discoverability and accessibility of research data. However, to do so there is a further barrier—the need to agree the semantics and structure of other datasets. Bizer and Berners-Lee acknowledged [22] that the development of standardised languages

providing detailed "schema mapping" and "data fusion" i.e., enabling the aggregation of such datasets is an issue. Surprisingly this has continued to be a challenge, largely due to the complexity of ownership and governance. The organisation CASRAI has set out to become coordinating authority for the management of such schema mappings—defining them as "data profiles", that when agreed are registered in an online dictionary of research administration information. Furthermore the future creation of large data profiles, or aggregations for reporting purposes, will require ownership by appropriate organisations prepared to resource their governance from establishing community consensus through to management by standards bodies.

These profiles will define the fields used to describe the content of datasets and/or part off a fuller dataset, i.e., the metadata enabling identification of an entry within the dataset. The UK HE is piloting a response to this challenge through the Jisc-funded CASRAI UK Pilot [23]. Like the community developed data profile the UNIQUIP data publishing specification CASRAI will provide the community with a managed "dictionary" of dataset terms. Longer term international adoption of standards enablers such as CASRAI will provide mechanisms for structured datasets to be established and discovered. This concept is discussed by Baker and Cox in the short article "Buttons to Beacons" [24].

6. Rethinking Workflows

The core infrastructure for data discovery and sharing is now largely defined in UK HE with the adoption of the OPD. The simplicity of this infrastructure will allow easy scalability providing there is established governance enabling managed development and appropriate standards to be applied to other data profiles as they are established.

Once published an OPD should not require any significant ongoing maintenance other than corrections to data locations or dataset contact information. There are likely to be many stakeholders with a responsibility for maintaining data locations on the OPD which raises some significant questions relating to maintenance and workflow:

- Who is responsible for the OPD Home page link and hosting of the OPD?
- Are appropriate procedures in place identifying stakeholder responsibilities?
- Is there sufficient understanding of publishing issues, e.g., risk and licensing?

The wider use of the OPD as a standard for data autodiscovery will place further focus on the requirement to establish appropriate ownership and governance of the document and related datasets. As noted earlier it may be logical for this to be your marketing and communications department who typically will be responsible for your website (home page) and could therefore ensure the link to the OPD is maintained and explained in the website build documentation. In UK HE governance of an OPD within an organisation could reasonably reside with the organisation's research data management "front of house", e.g., Library. It may also be practical to define management and maintenance in the Data Management Planning Strategy or Policy of the institution.

For those publishing data or considering publishing data there is a need to understand the workflow associated with that data. Who is responsible for the data? Do they understand the additional use and license to be applied? Do they need to—why not publish if there is no risk? Does the data map to an agreed profile? If it does this will enable greater value in its application in future use, e.g., in analytics. (Figure 5), below, illustrates the typical workflow and possible routes to publishing research equipment data, enabling discovery by the equipment.data service. For this simple dataset, it is evident that there are a number of stakeholders, including procurement, finance, research support offices and those responsible for the institutional website.

Many system vendors will aim to ensure their system complies with or is aligned to sector standards and will engage the sector as required to ensure such, for example Elsevier engage with the UK HE through the Pure User Group, who actively support development focussing on research outputs metadata and the equipment profile. Providing a system has a similar level of support for

open publishing it is very likely staff should only need to ensure data is of appropriate quality for their use and risks associated with publishing are considered.

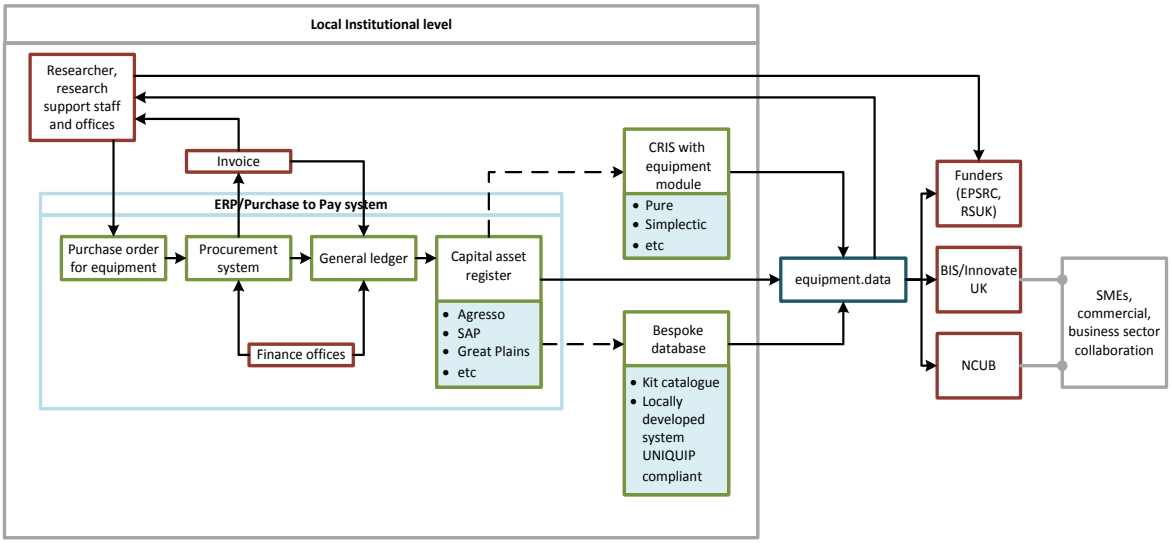

Figure 5. Equipment data workflow—Publishing equipment data in HE.

7. Conclusions

The OPD has proven to be successful in a small subject area with the early success of the equipment.data project demonstrating a data aggregation service can be established using an infrastructure built around the OPD. At the time of preparing this article there are 36 organisations in UK Higher Education publishing OPDs, 31 fully autodiscoverable, i.e., achieving the equipment.data Gold compliance rating. Noting this success and the sustainability it offers for data harvesting the National Centre for Universities and Business (NCUB) are importing data from the full OPD data aggregation api in the development of an industry focused intelligent information search and brokerage tool. However, it is noted that the sector's engagement with equipment.data has been driven to a large extent by a "policy" obligation, with success in its adoption noted as being the "low technical entry requirement" for contribution and a confidence that technical support was available through the implementation project. Such factors would need to be considered along with measurement of potential uptake in any future data discovery development using the OPD infrastructure.

For the HE sector and indeed wider industry to have confidence in the adoption of such open data technologies the requisite skills will be required within organisations. In 2009 Siorpaes and Simperl [25] (p. 33) noted, "interacting with semantic technologies today requires specific skills and expertise which are not part of the mainstream IT knowledge portfolio". This is still the case with only a handful of UK universities delivering open data specific taught modules and the ODI focussing on industry specific open data training. It is therefore possibly not surprising there is currently limited academic discussion about standardised approaches to data discovery, making it challenging to introduce an approach to data autodiscovery such as the OPD. Discussion could be encouraged through the adoption of the OPD as a W3C community group which may also advance the wider discussion and adoption of LOD and data discovery.

Although not noted, it is widely acknowledged by data managers that it is impossible to assess all potential users of openly published data as some users will be interested in a broad range of data, therefore data outside a defined data profile could be of interest. The decision therefore may be to publish all information fields within a given dataset, i.e., both structured to an agreed data profile and those outside the profile. We are already seeing systems developed to aggregate data in structured profiles able to validate the data and simply extract data within the specified profile and

ignore the fields outside the profile. This engagement and development trend is likely to continue and compliment the growing use of LOD based systems encouraging greater consideration of the workflow supporting them.

Service owners considering data aggregation, either for reporting purposes or delivering a search capability, will require an awareness of the data management challenges organisational workflows can present, including maintaining data quality, ownership and access. However, data discovered via an OPD will demonstrate to data aggregators the published data conforms to a standard data profile, has a person responsible for the data and specifies the license applied to the data, therefore, demonstrating a level of integrity in the data management process. The securing of an international standard for the OPD, e.g., becoming a W3C community group and/or standard, alongside registration in the CASRAI dictionary, will undoubtedly provide greater confidence for future adopters of the technology.

It is too early to assess in any measurable way the impact of open publishing and data reuse, the second edition of the "Open Data Barometer" [26] notes "While the "big tent" of open data, the well networked open data community, and the availability of shared guides, tools, and technologies, have all helped the open data concept to spread rapidly, there is no single best practice for delivering an open data initiative". These challenges are echoed by those discussed in this article. There are great opportunities for organisations to gain more value from the data they already create and curate, equipment.data already demonstrates the potential from the aggregation and re-use of institutionally published research equipment data. To exploit these opportunities more fully will require a greater awareness and application of open data concepts, such as quality, licensing and, fundamentally discoverability, where there is a very clear role for the OPD.

Acknowledgments: The development and delivery of the equipment.data service was funded by the Engineering & Physical Sciences Research Council (EPSRC) and Jisc.

Author Contributions: Christopher J. Gutteridge conceived and designed the programming for equipment.data and the OPD; Adrian J. M. Cox was responsible for service design and implementation and coordination of data analysis relating to service performance; Andrew J. Milsted provided programming development and support for analysis; Adrian J. M. Cox wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. How to Publish Linked Data on the Web. Available online: http://wifo5--03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/ (accessed on 5 July 2016).
2. Mitchell, E. Building Blocks of Linked Open Data in Libraries. 2013. Available online: https://journals.ala.org/ltr/article/view/4692/5584 (accessed on 5 July 2016).
3. Berners-Lee, T. The Five Starts of Open Data. Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 5 July 2016).
4. Higher Education Data and Information Improvement Programme (HEDIIP). "A New Landscape". Available online: http://www.hediip.ac.uk/about-hediip/ (accessed on 5 July 2016).
5. The Open Data Institute. Available online: http://opendatainstitute.org/ (accessed on 5 July 2016).
6. Equipment.data. The UK National Research Equipment Portal. Available online: http://equipment.data.ac.uk/ (accessed on 5 July 2016).
7. Organisation Profile Document (OPD). Available online: http://opd.data.ac.uk (accessed on 5 July 2016).
8. Wakeham, W. Wakeham Review—Financial Sustainability and Efficiency in Full Economic Costing of Research in UK Higher Education Institutions. 2010. Available online: http://www.rcuk.ac.uk/research/efficiency/efficiency2011/ (accessed on 5 July 2016).
9. The UNIQUIP Project. Available online: http://www.uniquip.ecs.soton.ac.uk/ (accessed on 5 July 2016).
10. JANET Network. Available online: https://www.jisc.ac.uk/janet (accessed on 5 July 2016).
11. MacEwan, A.; Angjelib, A.; Gatenby, J. The International Standard Name Identifier (ISNI): The Evolving Future of Name Authority Control. *Catal. Classifi. Q.* **2012**, *51*, 55–71. [CrossRef]

12. Github, Equipment.Data Autodiscovery Programming. Available online: https://github.com/data-ac-uk/equipment (accessed on 5 July 2016).
13. Research Data Alliance (RDA), Metadata Standards Directory Working Group. Available online: https://rd-alliance.org/groups/metadata-standards-directory-working-group.html (accessed on 5 July 2016).
14. Consortia Advancing Standards in Research Administration Information (CASRAI). Available online: http://casrai.org/about (accessed on 5 July 2016).
15. World Wide Web Consortium (W3C). Available online: http://www.w3.org/ (accessed on 5 July 2016).
16. Cox, A.; Milsted, A.; Gutteridge, C. Autodiscovery of Linked Open Data—The Need for Standards. In Proceedings of the Research Data Alliance 5th Plenary, San Diego, CA, USA, 8–11 March 2015.
17. Equipment.Data Data Discovery Compliance. Available online: http://equipment.data.ac.uk/compliance (accessed on 5 July 2016).
18. Organisation Profile Document W3C Community Group. Available online: https://www.w3.org/community/opd/ (accessed on 5 July 2016).
19. Documentation: Basic Structure. Available online: http://opd.data.ac.uk/docs/core (accessed on 5 July 2016).
20. Nottingham, M.; Hammer-Lahav, E. Defining Well-Known Uniform Resource Identifiers (URIs). 2010. Available online: http://tools.ietf.org/html/rfc5785?chocaid=397 (accessed on 5 July 2016).
21. The Linking You Toolkit. Available online: http://lncn.eu/toolkit (accessed on 5 July 2016).
22. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. 2009. Available online: http://www.igi-global.com/gateway/article/37496 (accessed on 5 July 2016).
23. Jisc. CASRAI UK Pilot Project. Available online: http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/researchinformation/casraipilot.aspx (accessed on 5 July 2016).
24. Equipment.Data Newsletters Archive. Available online: http://equipment.data.ac.uk/newsletters/issue4/beacons (accessed on 5 July 2016).
25. Siorpaes, K.; Simperl, E. Human Intelligence in the Process of Semantic Content Creation. *World Wild Web* **2010**, *13*, 33–59. [CrossRef]
26. The Web Foundation. *The Open Data Barometer*, 2nd ed.; The Web Foundation: Washington, DC, USA, 2015.