CRIS2016

Title: equipment.data – Delivering a data autodiscovery infrastructure

Track II: Communication, sharing and profiling research
Track III: Advances in research information technology

Authors: A J Cox, C J Gutteridge

University of Southampton, University Road, Highfield Campus, Southampton, SO17 1BJ

Corresponding author: A J Cox, Project Manager, equipment.data

Email: A.J.Cox@soton.ac.uk

Keywords: autodiscovery, rdf, data profile, linked open data, organisation profile document

Introduction

The development of the National Research Equipment Portal, equipment.data[1], was funded by EPSRC (Jisc from April 2015) in response to the need to improve visibility and utilisation of UK research equipment. The aim of equipment.data was to deliver a sustainable solution for the aggregation and displaying of published research equipment data from across UK HE in order to promote greater conversation around the utilisation of existing research infrastructure. Its development has the backing of RCUK as its preferred medium for national equipment data sharing with the service now endorsed as strategically significant by HEFCE.

Launched in April 2013, equipment.data introduced the concept of linked open data technologies enabling data autodiscovery to provide a service. Essential to this process is the publishing of an Organisation Profile Document (OPD)[2]. 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.

In addition to the service equipment.data provides it has also demonstrated a linked open data infrastructure can be implemented and with it established the foundation components for wider data sharing. The application of linked open data in data management is growing enabling new approaches particularly in the development of standardised data profiles. This allows data to be captured 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)). Equipment.data is the first linked open data driven service in UK HE demonstrating a simple sustainable system can deliver a service which is not only extendable but with introduction and application of the OPD re-usable.

[1] http://equipment.data.ac.uk
[2] http://opd.data.ac.uk

A new way to discover published data

For the equipment.data project to reach its goal of a fully sustainable system, it required a method of updating data sources as efficiently as possible with minimal or no human intervention. To encourage adoption of a sustainable method of contribution, i.e. using the OPD, the service established a compliance rating system[3] with gold, silver and bronze ratings to indicate to what level each contributing institution's data input is sustainable. Fig. 1.

	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)			✔

Fig. 1. Compliance rating applied to data discovery

The OPD, including the associated embedded link in the home page, is the principal enabler to the process of data autodiscovery; enabling machine discovery, it describes the organisation, and states 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.

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.

The "core" OPD information includes the organisation URI, parent or sub-organisations, location, primary contact information and dataset ts the organisation is publishing. 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. Fig. 2.

```
@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#>.
```

[3] http://equipment.data.ac.uk/compliance

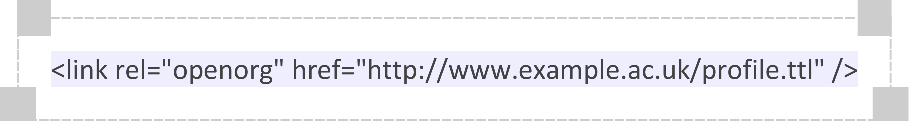
```
@prefix org:     <http://www.w3.org/ns/org#>.
@prefix xtypes:  <http://purl.org/xtypes/>.
@prefix lyou:    <http://purl.org/linkingyou/>.
@prefix vcard:   <http://www.w3.org/2006/vcard/ns#>.

<> 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> .
```

Fig. 2. OPD Basic Structure

The method used to enable autodiscovery of the OPD requires a link in the organisation homepage header fig. 3.



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

Fig. 3. Home page html header link

This link in the html header provides the location of the OPD enabling discovery programmes, "web crawlers", e.g. "dinas"[4] used by equipment.data 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. the 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 isn't possible the discovery programme has been developed so that the .well-known[5] method can be used. This 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.

[4] https://github.com/data-ac-uk/equipment

[5] http://tools.ietf.org/html/rfc5785?chocaid=397

By publishing a fully autodiscoverable 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 publishing fully autodiscoverable OPDs, 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. Data made discoverable via an OPD will demonstrate to data aggregators the data is published to a standard data profile, has a person responsible for the data and will specify the license applied to the data, therefore demonstrating a level of integrity in the data management process.

As wider use of the OPD increases the challenge will be to establish appropriate ownership and governance of the document and referenced datasets. This might be the marketing and communications department who typically will be responsible for an institution's website (home page) and could therefore ensure the link to the OPD is maintained and explained in the website build documentation. In UK HE the governance of an OPD could reasonably reside with an institution'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 open data or considering publishing 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.

Fig. 4. 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. As further standardised datasets e.g. research outputs, are considered, contributors to the OPD could include institution's Library who will be responsible for their repository and RDM related aspects.

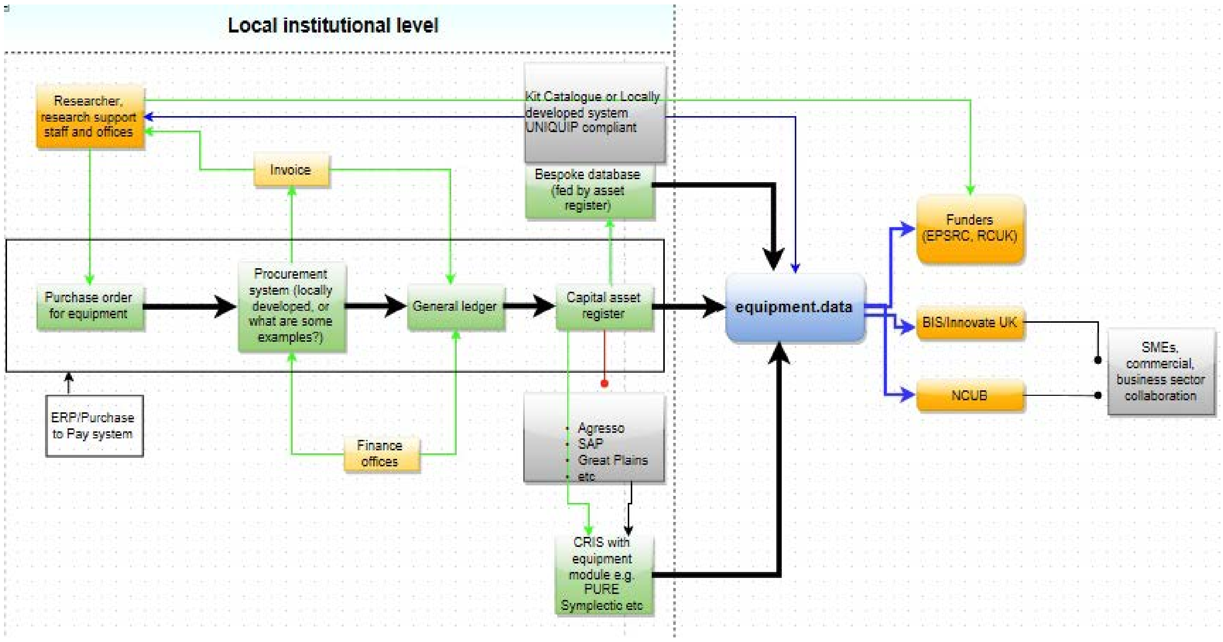

Local institutional level
Researcher, research support staff and offices
Invoice
Kit Catalogue or Locally developed system UNIQUIP compliant
Bespoke database (fed by asset register)
Funders (EPSRC, RCUK)
Purchase order for equipment
Procurement system (locally developed, or what are some examples?)
General ledger
Capital asset register
equipment.data
BIS/Innovate UK
NCUB
SMEs, commercial, business sector collaboration
ERP/Purchase to Pay system
Finance offices
Agresso
SAP
Great Plains
etc
CRIS with equipment module e.g. PURE Symplectic etc

Fig. 4. Equipment data workflow- Publishing equipment data in HE

The website opd.data.ac.uk currently discovers 31 published OPDs which include a range of standardised data and information including core institutional information, equipment and facilities data, research outputs OIA-PMH end points and “Linking you” web pages. This data aggregation provides a comprehensive data catalogue describing the “Who?”, “what?” and “Where?” for open published data and business information describing UK HE institutions. This data aggregation is then made available in .csv and JSON formats and is currently used in the delivery of the Intelligent Brokerage Tool, “Konfer”, the National Centre for Universities and Business (NCUB)[6] data services to be launched April 2016.

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, where the OPD could offer significant improvements to the discoverability and accessibility of research data through initiatives such as the Jisc funded UK Research Data Discovery Service (RDDS) project[7].

Conclusions and recommendations:

The success of equipment.data as national research equipment portal and the use of the opd.data.ac.uk data aggreagtions has demonstrated that linked open data based web services can offer very efficient ways of discovering standardised data. However, the discovery of future datasets are currently constrained by the need for consensus on standardised vocabulary if like for like datasets are to aggregated and re-published or used in comparative analytics.

Beyond the aggregation of equipment data and the institutional URI structure “Linking you”[6], if organisations are to extend the open publishing of data enabling aggregation in any meaningful structured form, mechanisms for managing and agreeing data profiles will be required.

[6] http://www.ncub.co.uk/

[7] https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

With organisations such as CASRAI[8] now established to provide a coordinating authority for the management of data profiles the HE sector, and indeed industry, are well placed to advance linked open data discovery.

These profiles will define the fields used to describe the content of datasets and/or part of 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[9]. 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 in the short article "Buttons to Beacons"[10].

As highlighted earlier, the challenge as wider use of OPDs increases 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), or the Library, who will be responsible for institutional Research Data Management (RDM). However, to date, due to the focus on research equipment data, the equipment.data service team have mainly worked with staff from research support offices and IT departments. As more links to structured datasets are established, and OPDs have a wider use, ownership and governance will need to be more firmly established. To enable such decisions it is likely the sector will require greater confidence in this emerging technology, therefore the aim is to establish a W3C Community Group[11] engaging the sector in future development of the OPD. It is hoped the future securing of an international standard for the OPD, e.g. W3C, will provide greater confidence for future adopters of the technology beyond HE. The sector itself should be encouraged to explore opportunities for the OPD in wider standard practices e.g. within RDM, where a Digital Curation Centre (DCC) led project has already explored opportunities within the Jisc Research Data Spring[12].

There are many emerging opportunities for institutions to gain more value from the data they already create and curate. To exploit these opportunities more fully will require a greater awareness and application of linked open data concepts such as quality, structuring, licensing and, fundamentally discoverability, where there is a very clear role for the OPD.

March 2016

[8] http://casrai.org/Main_Page

[9] http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/researchinformation/casraipilot.aspx

[10] http://www.data.ac.uk/newsletter/july2014/beacons 2014-11-05

[11] https://www.w3.org/community/opd/

[12] https://www.jisc.ac.uk/rd/projects/research-data-spring