Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 191 – 198

Procedia Computer Science

CRIS 2014

From Open Data to Data-Intensive Science through CERIF

Keith G Jeffery[a]*, Anne Asserson[b], Nikos Houssos[c], Valerie Brasse[d], Brigitte Jörg[e]

[a] Keith G Jeffery Consultants, Shrivenham, SN6 8AH, U,

[b] University of Bergen, Bergen, 5009, Norway

[c] Hellenic Documentation Centre, Athens, GR-11635, Greece

[d] IS4RI, Strasbourg, France

[e] JeiBee Ltd, London, UK

Abstract

OGD (Open Government Data) is provided from government departments for transparency and to stimulate a market in ICT services for industry and citizens. Research datasets from publicly funded research commonly are associated with the open scholarly publications movement. However, the former world commonly is derived from the latter with generalisation and summarisation. There is advantage in a user of OGD being able to 'drill down' to the underlying research datasets. OGD encourages cross-domain research because the summarized data from different domains is more easily relatable. Bridging across the two worlds requires rich metadata; CERIF (Common European research Information Format) has proved itself to be ideally suited to this requirement. Utilising the research datasets is data-intensive science, a component of e-Research. Data-intensive science also requires access to an e-infrastructure. Virtualisation of this e-infrastructure optimizes this.

Keywords: rich contextual metadata; CERIF; e-Science; open data; e-infrastructure; research data

* Corresponding author: Keith G Jeffery, Tel.: +44 7768 446088
E-mail address: keith.jeffery@keithgjefferyconsultants.co.uk

doi:10.1016/j.procs.2014.06.032

1. The Concept of Open Data

There is increasingly a move towards open data. Open data has two different – but related – kinds: it may be OGD (Open Government Data) or it may be data (including publications) generated by government funding of activities in research through grants to research projects at universities or research institutions.

1.1 Open Government Data

OGD typically is collected from citizens through various forms such as tax returns or census, by surveys done by the department or by various means under external contracts. It is made available for reasons of transparency and to promote a marketplace in added-value services. It is not unusual for the former to be informed by - and to be a summarisation of - the latter since government departments may well commission academic research or utilise the results of government-funded research.

Starting with a desire of governments to appear more transparent, OGD (Open Government Data) has become a trend in Western countries. In fact the major motivation is that by making available datasets collected by government departments with taxpayer funding commercial companies – especially SMEs (Small and Medium-Sized Enterprises) - will be encouraged to provide commercial services utilising this open data and adding value for the end-user. Well-known examples are 'apps' on smartphones for finding parking spaces in London or finding the arrival time of the next metro train. However, the census/demographic information, climate datasets, transport datasets, anonymised health datasets, education datasets are all valuable for commercial companies in planning products and their market and for ICT companies in creating new products and services utilising the datasets. In a substantial number of cases, these OGD datasets are summarised / derived from more detailed research datasets. These research datasets are generated commonly by research projects funded publicly. Thus we can distinguish OGD and publicly funded research datasets yet appreciate the relationship between them.

OGD has based itself on W3C (World Wide Web Consortium) standards LOD (Linked Open Data)[1] which indicates relationships between datasets typically using RDF (Resource Description Framework)[2] a representation using triples of subject-relationship-object. The metadata associated with a dataset can be represented in RDF – as is used in CKAN (Comprehensive Knowledge Archive Network)[3] - and even the records of the dataset itself can be represented as RDF triples. The lexical strings in the triples representing subject, object and relationship need to be understood and so typically SW (Semantic Web) technologies – such as ontologies in OWL (Web Ontology Language)[4] or SKOS (Simple Knowledge Organisation System)[5] are used.

1.2 Research data

Increasingly datasets produced in projects publicly funded are being made available openly. This may be due to the desire of researchers to have their work scrutinized, acknowledged and cited but more commonly it is being mandated by funding organisations – partly due to association with the open access to research publications movement. The variety and complexity of research datasets is huge; it is commonly very difficult for a researcher in one domain of research to understand a dataset from another. It is here that the association of the dataset with scholarly publications (white) or technical reports (grey) becomes vitally important. Furthermore, research datasets may only be utilised effectively with their associated software.

2. The Jungle of Open Data

There is a jungle of open data available with many different kinds of data accessible yet commonly lacking information to assist in their use. Here we include data of all kinds - not just traditional structured datasets – but encompassing also documents, video, audio and multimedia. In fact OGD consists more of pdf documents than of

structured datasets. Detailed research results (publicly-funded research data) are incomprehensible or unusable without additional information on precision, accuracy, instruments / equipment used and the purpose of the experiment or observation. For assessment of quality and relevance of research datasets it is also useful to know about the research context: project (including title, abstract, keywords), persons, organisations, funding, related scholarly publications and presentations at events. If the dataset is produced by experiment or observation then facility and equipment should be recorded. Even summary government department data such as crime statistics by region or education 'league tables' require explanations about the collection methods and hints on how they should be interpreted. Using open data should come with a 'health warning' else serious consequential misinterpretations may be made.

Moreover, open data is available in a variety of formats. It is striking that the majority of European government open data is in pdf format (and so readable but not understandable or processable by computers). Much of the remainder is in Excel format only usable in or through a proprietary spreadsheet environment. Some exists in the more general csv (comma separated value) format which can be read into other compatible systems as well as spreadsheets. Much government-funded research data made publicly available is in scholarly publications, again commonly in pdf format even if tables and graphs (which could be processed further) are involved. Of the remainder a vast variety of formats are used with little or no commonality. For these reasons it is commonly necessary to associate software with the data in order to be able to deal with it in an intelligent manner and not just reproduce it.

To add to the complexity OGD or research datasets have declared or implied licence conditions (legalistics). These may restrict the use of the data or make demands that the user of the data has to fulfil. An outline approach to this problem has been presented[6] .

3. Metadata

The key to managing open data is metadata. Metadata is required for discovery, for context (understanding the dataset and assessing quality and relevance) and for detailed utilisation (computer processing using domain specific formats) *Fig. 1*. The three-level architecture for metadata has contextual metadata as the middle layer. From this contextual layer one can generate discovery level metadata (so providing congruence with the contextual layer) and also one can navigate to detailed metadata related to an individual dataset or research domain. This three-layer structure also allows the association of LOD/SW OGD representations associated with the discovery layer to be linked to research datasets associated with the contextual and detailed layers. Metadata is, itself, data for some purposes and so the distinction between data and metadata is lost.

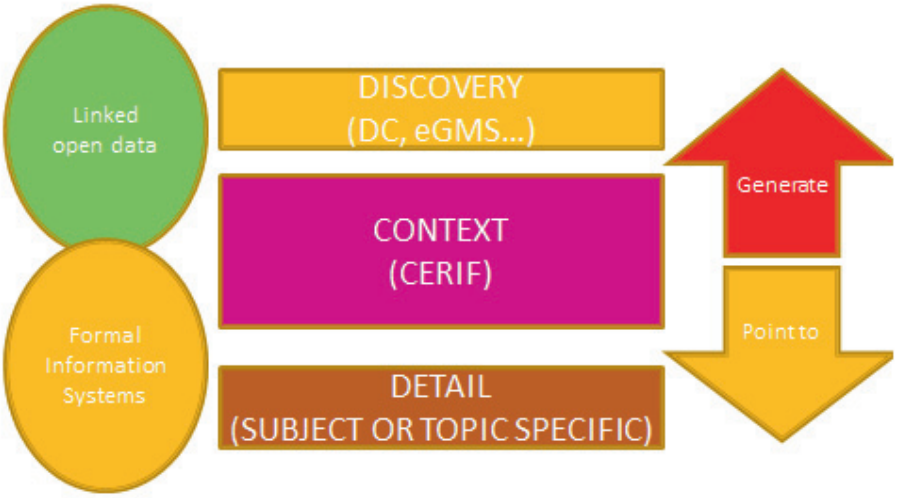

Fig. 1. 3-layer model of metadata

Discovery metadata is used by humans (query and selection from a list) or computers to find data that may be of interest. However, the conventional discovery metadata formats such as DC (Dublin Core)[7], CKAN, eGMS (e-Government Metadata Standard)[8], ADMS (Asset Description Metadata Schema)[9], INSPIRE (an EU standard for geospatial data)[10], DCAT (Data Catalog Vocabulary)[11] and others have limited expressive capability even for discovery (and certainly for other purposes). They have the advantage of simplicity and terseness. They have the serious disadvantage of lacking referential and functional integrity[12] and so contravene basic principles of information science. Nonetheless, because of their popularity, i.e. wide-spread use, and the need for interoperation at discovery level, they require to be supported.

Contextual metadata is a superset of discovery metadata and defines applicable relevant information for associated underlying datasets. These may include projects, persons, organisations, funding, related publications (grey, technical publications as well as white peer-reviewed formal publications), other datasets, software, patents and services, facilities and/or equipment used for generating the data. It can also describe legalistics, rights and obligations associated with utilisation of the data. Most importantly, contextual metadata describes not only the entities or objects related to the dataset but also the relationships between them and the temporal validity of the relationship. This is important for provenance (tracing the history) and versioning of datasets. Contextual metadata is the lowest or most detailed level of metadata that is common across all – or most of – the variety of data available. CERIF (Common European Research Information Format)[13] is the most complete contextual metadata 'standard'[†] with formal syntax and declared semantics.

Detailed metadata is specific to the data it describes. At the lowest level it is a schema (in the database sense or the document structure sense) which is specific to a particular set of data or to a research domain producing multiple datasets in a standard format and provides the link between the dataset and computer processing. It may also include detailed information on accuracy, precision and other parameters required by software to process the data. Detailed instructions on processing the data are usually found in grey literature publications (technical reports) related to the dataset through the relationships in the contextual layer.

---

[1] Technically CERIF is an EU Recommendation to Member States

4. Open Government Data plus Research Data

We now have the basis for understanding how to provide an end-user with appropriate access to - and utilisation of - open data. We need to bring together OGD (commonly summary data, mainly pdf documents, limited discovery metadata) with the results of government funded research (detailed data in many formats with detailed and contextual metadata). The ENGAGE project has adopted the 3-layer architectural model for metadata with discovery, contextual and detailed levels. The end-user has the choice to browse across discovery metadata so allowing serendipitous discovery of new knowledge or to query more precisely contextual metadata when searching for specific targeted knowledge. Moreover contextual metadata can be used to enhance that associated with OGD datasets improving relevance and recall as well as quality as seen by the end-user. The 3-layer model of metadata – relying on CERIF as the core - bridges across OGD and research datasets Fig. 2 thus bringing together the two kinds of open data.

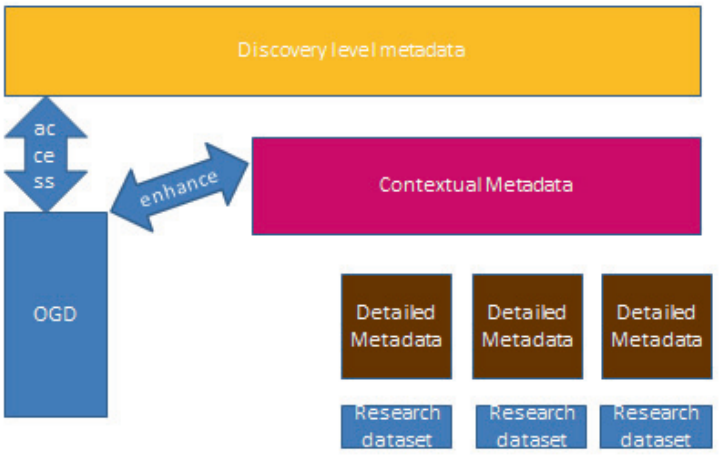

Fig. 2. Bridging OGD and Research Data

5. The e-Infrastructure Requirement

The provision of access to – and download of – open data is only one step in the provision of an e-infrastructure for the research domain. ENGAGE – although a significant advance over comparable portals – lacks processing capability and the variety of dataset formats precludes easy interoperation. Typically in ENGAGE the end-user has to do a lot of manual work to match and map datasets, invoke software and produce reports or visualisations – commonly with several data format interconversions between the steps. This demonstrates a need for a complete e-infrastructure to support the end-user in utilising this information. Such a proposal[14] was made to UK government by resulting in the UK e-Science research programme (and subsequently that of the EC) and a refined version was presented to the CRIS community[15] indicating the merging of research management and e-infrastructure research outputs. The use of metadata may be extended beyond modelling data and information to provide a model of the services (software processes commonly in workflows), the utilisation of resources (computers, data stores, communications and detectors/sensors) and the end-user (characteristics, authorities, responsibilities and

preferences) Fig. 3. This is virtualisation which simplifies the view of the data processing world. The essential point of virtualisation is that the user neither knows nor cares how or where the computer processing of the data is done provided that service level demands are met.

This implies rich metadata describing the various components of the e-infrastructure that can be assembled (data, software, computing and other resources) to satisfy the end-user request. Furthermore, this implies autonomic (self-managing) middleware services to perform this assembly, to manage and monitor resources and to assure trust, security and privacy as appropriate including rights (usually controlled by licences) management.

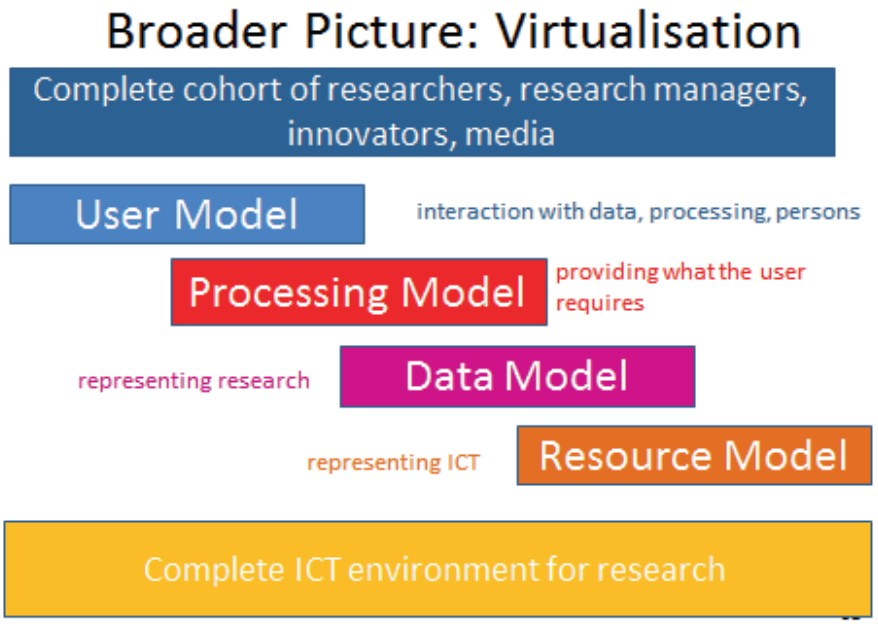

Fig. 3. The e-Infrastructure Components

As an example, this approach is key to EPOS-PP (European Plate Observing System – Preparatory Project (http://www.epos-eu.org/) where the EPOS catalog uses CERIF to relate all the e-infrastructure objects. These include not only research datasets (at various levels of abstraction through refinement and processing) but also software, users and computing resources. In addition the catalog describes arrays of sensors/detectors and their processing systems, equipment used in rock mechanics laboratories, satellites with their observing equipment and many other components of EPOS. Of course all these components are related to other entities – among others - organisations, persons, funding and publications.

6. CERIF for Data Intensive Science

The importance of metadata – and specifically CERIF - can now be appreciated; it provides the necessary information model to characterize not only data / information but also all the components of the e-infrastructure such that intelligent computer middleware can utilise the metadata (models) to put together a package to respond to

the needs of the human end-user and provide that which is required. This is the concept of e-Research (commonly called e-Science) whether used by the researcher in the laboratory, the educator, the citizen scientist or the interested layman.

In the domain of data-intensive science – a part of e-Research - CERIF can model datasets but also – importantly – relate the dataset to software and to persons, organisations, projects, publications, funding etc. Better, CERIF can also manage versions of datasets and their contextual relationships thus providing either provenance (the history that brings the dataset to its present state) or interoperable linkages based on common entities and attributes – or both.

Data intensive science is commonly associated with high throughput and/or high performance computing. Here the partitioning / replication of datasets and the dynamic re-distribution of services (software components) across the available computing, data storage and networking resources is of paramount importance. This is commonly known as scheduling with or without load-balancing. The approach was demonstrated in GRID Computing projects in the years 2000-2010 and is part of the CLOUD Computing approach.

In fact CLOUD Computing goes further in providing dynamic elastic scale up or down (increased / decreased power from increasing / decreasing active processors in one cluster of servers) and/or scale out or in (transfer part of the workload to other clusters of servers or return to using fewer clusters). This approach can produce both cost-savings and energy consumption savings. However, initiating or closing down processors in a cluster, or shipping VMs (Virtual Machines) from one cluster to another has an overhead and latency.

The choice of appropriate hardware resources to be utilised may well depend as much on data locality as processing characteristics. The latency of shipping a large dataset over the network commonly dictates the use of mobile code - moving the software to the data. However, the processing facility associated with the data storage may not be optimal for the requirement and may have restrictions – from security through commercial to licence agreements – precluding hosting mobile code related to datasets. The alternative may then be to ship the data to the processing facility where more efficient throughput is available. The balance between improved throughput and transmission latency has to be evaluated. Moreover, real-time data collection from equipment / instrumentation / detectors commonly is involved in data-intensive science requiring data streaming management technology: the locality of the detectors/instrumentation collecting the data becomes a factor. To add complexity, all this has to be done under constraints including performance, elapsed time, privacy, security and trust.

The EC part-funded project PaaSage (http://www.paasage.eu/) is tackling optimisation of applications in a multi-CLOUD environment. A simple use case is for an application running on an in-house CLOUD to require extra resources to remain within SLA (Service Level Agreement) constraints. The requirement is to scale out to one or more public CLOUD platforms. The PaaSage approach is to use middleware (profiler, reasoner, adapter) to configure an ‘envelope’ around the application allowing dynamic deployment across heterogeneous CLOUD platforms while respecting the constraints due to SLA (mainly elapsed time and cost), legalistics (e.g. no personal data outside EU), data locality, platform availability etc.) CERIF is being used within the Metadata Database of PaaSage as it can model effectively the parameters needed for this dynamic (re-)scheduling. Together with appropriate middleware, CERIF provides the models for a virtualised, autonomic e-Research environment suporting data-intensive science.

7. Acknowledgements

The authors acknowledge the contributions of colleagues in the ENGAGE project part-supported by EC Contract 283700. Keith Jeffery wishes to acknowledge the work of colleagues on the PaaSage project part-supported by EC Grant agreement no: 317715 and on the EPOS-PP project part-supported by EC Grant Agreement no. 262229

References

1.http://www.w3.org/wiki/LinkedData retrieved 12-06-2013
2. http://www.w3.org/RDF/ retrieved 18-02-2014
3. http://ckan.org/features/metadata/ retrieved 08-06-2013
4. http://www.w3.org/2001/sw/wiki/OWL retrieved 18-02-2014
5. http://www.w3.org/2004/02/skos/ retrieved 12-06-2013
6. Bunakov V, Jeffery KG. Licence Management for Public Sector Information. In: Parycek P, Edelmann N, editors. *Proceedings Conference for E-Demoracy and Open Government (CeDEM)* 2013 pp 292-302
7. http://dublincore.org/
8. http://www.esd.org.uk/standards/egms/ retrieved 18-02-2014
9. https://joinup.ec.europa.eu/asset/adms/home retrieved 18-02-2014
10. http://inspire.jrc.ec.europa.eu/index.cfm/pageid/101 retrieved 18-02-2014
11. http://www.w3.org/TR/vocab-dcat
12. Jeffery KG, Asserson A, Houssos N, Jörg B. A 3-layer model for Metadata In: Greenberg J, Ball A, Jeffery K, Qin J, Kakela R. editors. *CAMP-4-DATA Workshop; Proceedings International Conference on Dublin Core and Metadata Applications*, Lisbon September 2013
13. http://cordis.europa.eu/cerif/
14. Jeffery, KG. Knowledge, Information and Data (*Internal Technical Report proposing to UK government the e-Science programme*) https://epubs.stfc.ac.uk/work/28736 retrieved 20140221
15. Jeffery K . CRIS in 2020. In: Dvorak J, Jeffery KG, editors. *Proceedings 11th International Conference on Current Research Information Systems (CRIS2012)*, Prague, 2012