Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Data Intensive Science: Shades of Grey

Keith G Jeffery[a] *, Anne Asserson [b]

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

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

Abstract

The vast majority of research output is grey; white (peer reviewed scholarly publications) forms a minor proportion. Historically, grey material was generated and used within an organisation. However, in recent years some research-relevant grey material is being made openly available. Among grey outputs, research datasets are the largest proportion by volume and increasingly these are being made openly available. It is necessary for users of grey material to have some indication of reliability (quality, context, availability) so that they can judge whether the grey material is appropriate to their requirements. Rich metadata provides a mechanism for such evaluation. CERIF (Common European Research Information Format) provides such a rich metadata environment. Furthermore, CERIF allows generation of discovery level metadata (such as DC (Dublin Core), DCAT (Data Catalog Vocabulary), CKAN (Comprehensive Knowledge Archive Network) for simple retrieval or browsing and provides navigation to more detailed and specific metadata about the grey object. CERIF provides a bridge over research datasets and open government data. CERIF thus forms the lowest common level of metadata across grey (and white) objects.

Peer-review under responsibility of euroCRIS.

Keywords: metadata; syntax; semantics; grey literature; datasets; contextual metadata; discovery metadata;

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

1. Introduction

1.1. Data Intensive Science

The concept of data-intensive science has gained prominence recently with the publication of 'The Fourth Paradigm' in honour of Jim Gray[1]. The idea is not new but recent technological advances have made it both

Peer-review under responsibility of euroCRIS.

affordable and achievable. Conventionally research has involved hypothesis followed by observation or experiment and publication leading to discussion and confirmation or refutation. In some cases the hypothesis came as a result of observation (the work of Charles Darwin on Galapagos finches is a case in point). Data-intensive science takes this one step further and – using datasets from observation or experiment – looks for patterns (so-called data mining) and thus generates a hypothesis using induction over the dataset. Of course the patterns may have no causal origin and the skill in data-intensive science is to discover the patterns that do relate to causality. Once discovered the causality of the pattern may be tested by experiment.

This re-use of research datasets allows a kind of peer review of the datasets upon which grey or white publications may be based since the datasets are re-examined. Data collection is expensive, and data curation (including adding appropriate metadata and information concerning the research method or process utilised) usually more so. Re-use of datasets to extract the maximum science from them is economically sensible and morally justifiable. However there are dangers: a researcher (or citizen scientist) unfamiliar with a particular research domain may well draw erroneous conclusions from analysing datasets. It is for this reason that research datasets made openly available should have attached all relevant additional information to improve the utilisation experience.

1.2. *Grey*

Grey literature (or, more generally, grey objects) is helpfully (!) defined as that which is not white. White is defined as scholarly publications peer-reviewed and published in commercial or semi-commercial journals or proceedings. Grey literature originally referred to internal technical reports (i.e. written documents) within an organisation but over the last 20 years has come to include all grey objects ranging from datasets and video/audio recordings through software to newsletters and brochures – the latter being so called ephemera. Increasingly grey material is made available online and as such is more readily available and discoverable. Here we use grey to refer to the plethora of grey objects and we are concerned with grey in the domain of research. In this context the major grey component is research datasets and research datasets are the primary material used in data intensive science. Hence we relate data intensive science to grey.

1.3. Research Information

White literature is a tiny proportion of the information available in the domain of research. Grey is very much the dominant component both in number of object instances (items) and volumes of data. In the past grey material typically remained within the confines of an organisation since it provided a record of the ‘know how’ of that organisation and was - in many cases – commercially sensitive. Examples include a technical report on how to synthesise a particular chemical compound in a pharmaceutical company or a dataset recording symptoms of patients having been treated with a particular drug in clinical trials. In such examples it is clear that reliability (as well as security and privacy) is important; grey material produced in one department of an organisation and used in another required some sort of reliability assurance for its correct interpretation and utilisation. `

While the quality of white information is (arguably) assured by peer review, the quality of all grey has not yet been assured and indeed there are no generally accepted methods for such assurance. However, that is not to say that grey material is intrinsically less reliable than white. Internal technical reports are usually subject to strict internal peer review within an organisation but not external peer review. In some ways within-organisation review (i.e. members of one department reviewing the output of another) can be more critical than academic peer review. Patents are subject to a very strict peer review which is external to the originating organisation but within the secure and confidential procedures of a patent office. Again review is rigorous with checking of any related patents or applications (analogous to a research reviewer checking the earlier literature) for replication of the idea including possible plagiarism. PhD and Masters’ theses are subject to peer review and in the case of a thesis composed of a group of published papers to double peer review. Indeed, it can be argued that these kinds of grey are subjected to stricter quality controls than white.

However, the quality of the major (by volume) component of grey is not assured by a formal process; this is the domain of research datasets. In the domain of grey research information, research datasets range from documents through images, video and audio to complex digital representations of artefacts and structured research datasets. There are examples – particularly in the medical domain - of datasets associated with a white publication being analysed by another researcher and leading to a further publication supporting or opposing the hypothesis in the original white publication. At this point there is an interesting question whether the research dataset – now subjected to vigorous peer review by at least one other researcher - has crossed over from grey to white. Somewhat perversely, some researchers have developed software to 'scrape' a dataset from tables or graphs in white scholarly publications; this process seems intrinsically less reliable than accessing the original dataset used in the production of the publication but, of course, may be used to get around licensing / access limitations on the original dataset if they exist. Similar products exist for 'scraping' web pages; Scraperwiki is perhaps the best-known.

As we have argued elsewhere[2,3] grey is commonly the basis for innovation, wealth creation and improvement in the quality of life and thus has a value different from conventional academic (white) outputs and should be evaluated differently. However, now that (some) grey material is escaping from the confines of one organisation to a wider audience, and since grey is such an important component of research information, it is necessary to provide some mechanism for assuring reliability of the information. In the absence of formal external peer-review type procedures the best mechanism is to provide as much information as possible about the research dataset – in the form of rich metadata – and let the end user decide if the dataset is reliable (that is relevant, appropriate and of sufficient quality) for the purpose.

1.4. Open Government Data

There is increasingly a move towards open data. 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. 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.

2. Reliable Information

Reliable information – and specifically its utilisation in the research domain - depends on three aspects: quality, context and availability. All three depend critically on the provision of appropriate, high-quality metadata**.**

2.1 Quality

Quality consists of several aspects: data integrity defined by schema and constraints (which may operate over textual data (XML schema) as well as structured numeric data; accuracy and precision defined by the detailed domain-specific (even experiment or observation-specific) metadata; dealing with incomplete and inconsistent information; assuring recording of temporal validity and mechanisms for independent validation including quality rating (as used in amazon.com to rate purchased products or booking.com to rate hotels). Such an evaluation is, loosely, related to scholarly peer review. The independent evaluation is, of course, subjective but the other aspects are objective and with appropriate recording can be used to assure quality at least in the sense of whether the data within the dataset are of sufficient precision and accuracy for the intended purpose and whether the dataset is sufficiently complete and consistent for the kind of analysis envisaged.

2.2. Context

Context concerns related information in entities and attributes that give confidence that the information (dataset) of interest is understood within the research environment. It provides the end-user with information concerning the environment of the research dataset such as the persons in the research team, the equipment used, the project(s) within which the work was done and the related white publications and thus allows the end-user to judge the quality and relevance of the dataset for her work. Most researchers know the competing teams in their domain of interest and the characteristics of particular research facilities or specific pieces of equipment and can thus judge the reliability of the outputs from such a team. Additionally, it is necessary to record the provenance of a dataset; commonly one dataset may be derived from one or more pre-existing datasets and the derivation may involve transformations, corrections, selections / summarisation. These processes that have acted on the dataset need to be recorded so that – if necessary – an end-user can go back to an earlier version more suitable for their purpose.

2.3 Availability

2.3.1 Persistence

Persistence requires digital preservation; it is necessary to have a system for media migration. Who can now read a 7 inch floppy disk? Who can now read a 3420 IBM tape? With the migrated dataset the associated metadata must have formal syntax and declared semantics in order to be both read and understood centuries hence. It is not sufficient just to record the record format used within the dataset (i.e. attribute names and data types); as much as possible of the context in which it was collected should be recorded and preserved. However not only the dataset(s) (and their associated metadata) need preservation; commonly much of the 'science' is in the software (which records the analysis technique or simulation technique as algorithms) so any related software also needs preserving. Again there is a migration problem – who now can read Mercury autocode? We have to deal with changing languages, compilers / interpreters, changing operating environment (sequential, parallel, distributed, data dependencies) and so the most promising route to preservation is to utilise software specifications. In the ideal case such specifications allow accurate regeneration of the software in the current operating environment so that it has the same functional characteristics as the original. Of course the non-functional characteristics may well change – not least performance which usually is much improved on more modern hardware.

2.3.2 Access

Access in the research domain implies open (no barriers) and toll-free (no fee subject to conditions, licences). For grey material to be open then it must be discoverable; the richer the metadata the greater the chance of discovery and – more importantly – the greater the likelihood of discovering dataset(s) which are relevant (retrieval precision) and finding all such datasets (retrieval recall). In fact embargo periods are relevant for research datasets giving the principal investigator or research team sufficient time for data analysis and prior publication. This implies that the dataset(s) may not be open until a certain date or only open in limited way to specific persons in specific roles until a certain date. Similarly some grey material may be available under strict conditions to preserve commercial confidentiality. Recording such restrictions requires sophisticated metadata; it is analogous to conventional access control. In particular it requires minimally temporal relationships associated with role-based relationships between persons and datasets. There may be conditions attached to dataset access; typical is conventional acknowledgement as is usual in the research domain but there may be further conditions such as reporting any errors to the originator or even transferring ownership of derived data to the original dataset owner.

The question of toll-free access remains problematic. As publishers of white literature progressively store also associated datasets in their systems – and either charge subscription fees for access (thus limiting access to those users licensed by the subscription conditions) or author fees to publish and thereafter offer free access – it is clear that access to the dataset(s) in the world of commercial publishing is not toll free. If a white literature scholarly paper is published under the subscription or author pays regimes and the author makes the associated dataset(s) available in a local institutional repository there is open, toll-free access and the advantage of expertise about the dataset co-located with the dataset itself but as a disadvantage the cost of maintaining the dataset openly available. As the number and size of research datasets increases this cost could become significant; the major cost is not in the digital storage but in ensuring availability. Furthermore, for large datasets the network latency is large and

increasingly there are demands to run software at the locality of the dataset. This implies free access to computing resources at the site of the repository of datasets. Some organisations provide datacentres for datasets produced in projects they have funded; in UK NERC (Natural Environment Research Council) is an example. In such professional datacentres the datasets are curated in a way similar to specimens in a museum and thus their reliability is greater and their value to researchers increased. However, the paramount consideration should be that the metadata concerning the dataset provides all the required discovery information for finding the dataset, the contextual information for understanding the dataset and evaluating its relevance and reliability and the detailed, specific metadata allowing software to process the dataset appropriately.

3. Rich Metadata

Rich metadata allows better discovery and utilisation of datasets. Rich metadata has formal syntax (structure) and declared semantics (meaning); these two aspects are related. Consider a simple search using a term such as 'Green', 'Brown' or 'Black' (or for that matter, 'Grey'). All terms are colours and so could retrieve datasets with such a term in the title or abstract but equally all these terms are common English family names and could be found in the attribute for person. This lack of declared semantics (indicating whether the lexical terms are colours or names) leads to poor relevance (precision) in the query. The use of terms in specific fields (formal syntax) improves greatly the situation but such a query relies on rich metadata having been input.

We have argued that rich metadata is required to assure reliability (including quality) of grey and to understand the grey material in context. Furthermore, rich metadata - and a query screen allowing search terms to be placed in specific fields – improves greatly precision (relevance) of the query. Whereas white research information, and open government (summary) data can be discovered with relatively poor metadata (although with some lack of recall and precision) it is necessary to have rich metadata for grey research information to be discovered, evaluated for relevance and quality and then utilised. In fact the utilisation of white literature and open government data would also benefit from rich metadata but it is less necessary than for grey material since there is some assurance of quality.

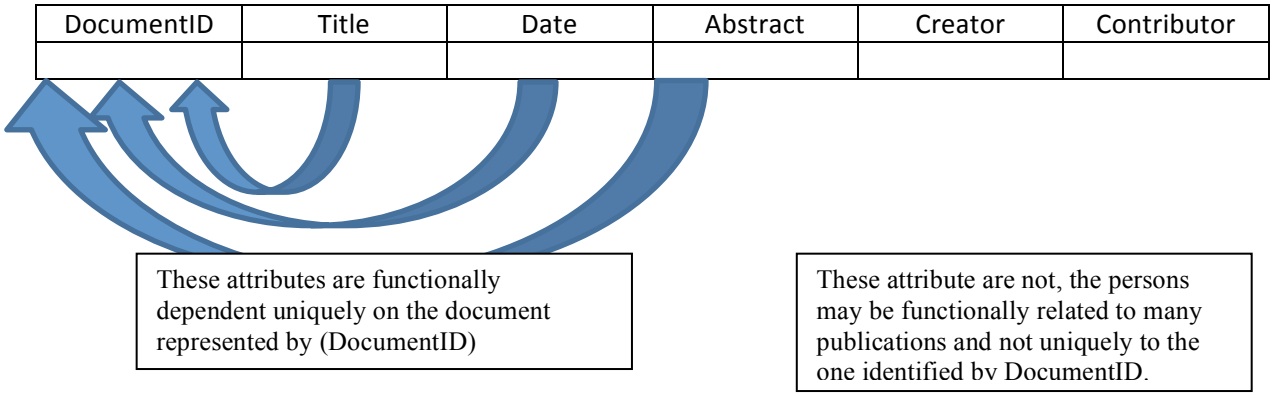

Fig. 1. Referential Integrity Problem

It has been asserted previously[4] that the metadata standards used commonly in the library domain and for description of web resources (such as DC[5] and CKAN[6]) are incapable of providing the required richness. Not only do they have insufficient relevant elements or fields to cover the required properties of the research dataset needed for discovery and use but their 'flat' structure means that they violate functional and referential integrity. As an example, consider a metadata record for a publication. Such metadata in DC has the name of a person as an element or attribute (confusingly as <creator> or <contributor>). Information theory states that each attribute has to be functionally and referentially dependent on the identity of the instance of the object being described if it is to have

integrity. In the case of this record describing a document or publication, clearly the person identified by the name does not exist only as a creator or contributor of a particular publication (i.e. related functionally the unique identifier of the record referring to the publication) but exists independently and probably participates in projects, is employed by an institution etc. The person as creator or contributor in this case is not exclusively functionally dependent on the unique identifier (of the document) Fig.1.

The correct structure – to represent the real world - is a relationship between the person and the publication with the person in the role of author – and for a temporal duration. Similarly, the person would have a relationship to one or more projects, organisations and other entities Fig.2. A similar argument can be made for the person(s) responsible for a research dataset. The use of flat metadata standards for research datasets is thus inappropriate.

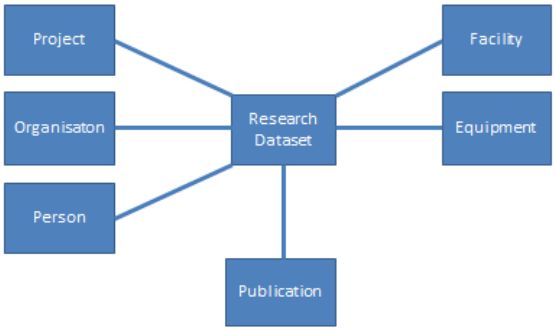

Fig.2. Major Entities Related to Research Datasets

It should be noted that the entities concerned may have multiple roles. For example, Organisation may be that owning/operating the facility/equipment from which the dataset was generated, or that employing the person (researcher), that providing the datacenter where the dataset is stored or the funding organization for the project within the context of which the dataset was generated. Of course for each entity one instance can refer to another instance (e.g. one dataset can have a relationship to another).

To overcome the aforementioned information violation CERIF 2000[7] was designed. The earlier CERIF91 was also 'flat' and suffered from the information integrity violation problem. The OpenAIRE community has now moved from flat metadata to CERIF for exactly this reason. More recently the W3C community has moved in this direction with the development of RDF (Resource Description Framework – a system of triples of subject-role-object). OGD - which is grey although some is subjected to internal peer review and increasingly portals offer the end-user the ability to provide a 'star rating' for a dataset - is intended to be based on RDF[8] and the LOD[9] (Linked Open Data) concept. At present <4% of the open government datasets are in RDF. A slightly larger number have metadata in RDF using CKAN; however while CKAN is an improvement over DC it lacks many of the elements or attributes needed to describe research datasets and so is insufficiently rich for the purpose. Furthermore, CKAN cannot record the complex relationships between the entities in the research domain. Typically OGD portals provide only a list of datasets and clicking provides a landing page with metadata and the URL of the target dataset or a direct link to the dataset if on the same server system. The use of RDF makes it clumsy to express both role and temporal relationships. However, this problem is also an exciting opportunity for the euroCRIS community to provide another route to research information - i.e. the research content of OGD alongside grey research datasets - but still using CERIF.

3.1 CERIF

Fortunately CERIF was developed – a decade earlier than the LOD / RDF concept – with similar principles. CERIF has base entities related by linking entities so reflecting the triple structure in instances within those entities. CERIF has – above a formal syntax – declared semantics in the 'semantic layer' which is analogous to ontologies[10,11] and indeed can represent anything that can be represented in an ontology. CERIF does not break down records to the level of triples (with each attribute related to the record identity) but has attributes referentially and functionally dependent on the record instance identity represented as a tuple. This has advantages in performance and compact digital storage. The result is that CERIF can generate other metadata formats and can represent multiple ontologies, including crosswalking between them. It is thus a superset or rich representation over other metadata standards.
Furthermore, CERIF through its federated identifiers feature allows inclusion of the increasing number of so-called unique identifiers (any system trying to manage identifiers is error-prone) for persons, projects, organisations etc. Importantly, by relating the various IDs to other attributes and entities CERIF provides the ideal mechanism for disambiguation. CERIF internally uses UUIDs (Universal unique IDs) which are generated and therefore incur no management overhead.

However, not all is perfect. The effort of providing rich metadata is considerable – and that is why simple flat metadata schemes are more popular. We have demonstrated previously[12] that incremental collection of rich metadata utilising the natural workflow of research reduces considerably the burden if the underlying data model and processes are optimized for data re-use so avoiding re-input.

3.2 3-layer model

The ENGAGE project (www.engage.eu) relates research datasets to OGD and the project team members from euroCRIS developed and promoted a 3-layer metadata architecture (Fig. 3.) with CERIF as the contextual metadata in the middle of the sandwich, generating the upper discovery level metadata in DC, eGMS[13], CKAN or other formats (and by generating it ensuring congruency) while simultaneously pointing to more detailed metadata associated with a research dataset which is domain specific at the lower level. This links together the LOD/semantic web world using RDF and browsing (with some structured retrieval using SPARQL) and the world of information systems with structured query (SQL) and richer metadata. Thus CERIF forms the lowest common level of metadata across research datasets from multiple research domains. In this way ENGAGE brings together open government datasets – usually summary information and commonly in pdf rather than structured data and with rather poor metadata in DC, DCAT or CKAN – with research datasets described by rich metadata and specifically those research datasets from which the government summary information was derived. This provides another – and generally more open - access route into research datasets for policymakers, researchers and citizen scientists.

4. Conclusion

We assert three points: (1) in the context of data-intensive science the importance of grey; (2) the need for reliability mechanisms to ensure the quality and relevance of grey and (3) the need for rich metadata to support the usage of grey.

Grey objects provide the vast majority of instances and volumes of research information. Of this grey information, research datasets form overwhelmingly the largest part. There is a need for reliability (including quality) of grey so that it is used appropriately. Although there are no conventional peer review quality mechanisms in common usage for grey, we argue that the provision of rich metadata with research datasets (and other grey objects) provides the end-user with a basis for assessment of the relevance and quality for their purpose. Such metadata also provides a bridge to the other kind of open data - OGD - so providing the user with a convenient access to all the relevant information. The effort of providing rich metadata may be mitigated by the CERIF data model which encourages re-use of data and no re-input. What remains is the need for the research system (widest sense) to reward researchers who provide rich metadata so that their research outputs are re-used. This relates to research indicators

for outputs, outcomes and impacts and the allocation of research funding based on those indicators. It also relates to motivating researchers to look beyond white literature as research output and embrace grey – particularly datasets.

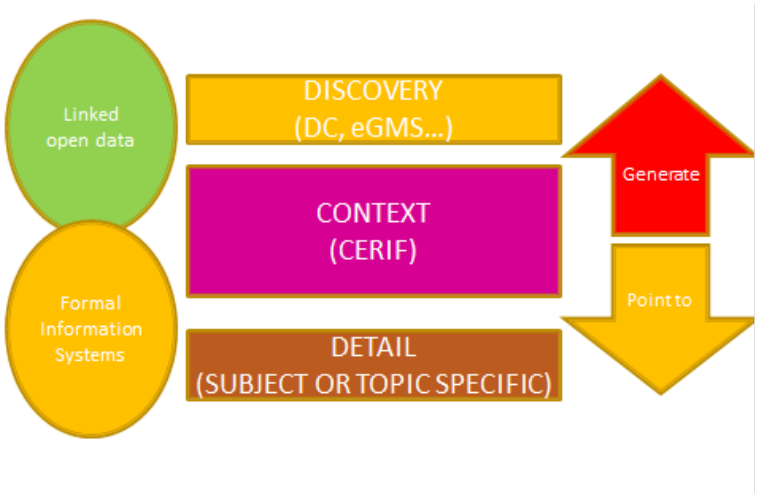

Fig. 3 The 3-layer Metadata Model

Acknowledgments

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 EPOS-PP project. EC Grant Agreement no. 262229

References

1. Hey T, Tansley S, Tolle K, The Fourth Paradigm. Microsoft Research 2009 http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf
2. Jeffery KG, Asserson A. Greyscape. In: Ferace D, Frantzen J, editors. *Proceedings Grey Literature 9 Conference Antwerp (GL9)* Textrelease, Amsterdam; 2007. p. 9-14
3. Jeffery KG, Asserson A. Grey in the Innovation Process. In: Farace D, Frantzen J, editors. *Proceedings International Conference on Grey Literature*. Rome 2012; p. 25-34
4. Jeffery KG, Asserson A, Houssos N, Jörg B. A 3-layer model for Metadata. In: Greenberg J, Ball A, Jeffery KG, Qin J, Koskela R, editors. *CAMP-4-DATA Workshop; Proceedings International Conference on Dublin Core and Metadata Applications.* Lisbon, 2013 http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/208
5. (DC) Dublin Core http://dublincore.org/
6. (CKAN) http://ckan.org/features/metadata/ retrieved 08-06-2013
7. (CERIF2000) http://cordis.europa.eu/cerif/
8. (RDF) http://www.w3.org/RDF/ retrieved 18-02-2014
9. (LOD) http://www.w3.org/wiki/LinkedData retrieved 12-06-2013
10. (OWL) http://www.w3.org/2001/sw/wiki/OWL retrieved 18-02-2014
11. (SKOS) http://www.w3.org/2004/02/skos/ retrieved 12-06-2013
12. Jeffery KG, Asserson A. Supporting the Research Process with a CRIS. In: Asserson A, Simons EJ, editors *Enabling Interaction and Quality: Beyond the Hanseatic League*, Leuven University Press, 2006 p. 121-130
13. (eGMS) http://www.esd.org.uk/standards/egms/ retrieved 18-02-201