Semi-automated application profile generation for research data assets

João Rocha da Silva[*]
Faculdade de Engenharia da Universidade do Porto/INESC TEC, Portugal,
Cristina Ribeiro and João Correia Lopes
DEI — Faculdade de Engenharia da Universidade do Porto / INESC TEC, Portugal
{pro11004,mcr,jlopes}@fe.up.pt

Abstract. Selecting the right set of descriptors for the annotation of a specific dataset can be a hard problem in research data management. Considering a dataset in an arbitrary domain, an application profile is complex to build because of the abundance of metadata standards, ontologies and other descriptor sources available for different domains. We propose to partially automate the process of data description by generating application profile recommendations based on a research data asset knowledge base. Our approach builds on existing technologies for exploring linked data and results in a process which can be tightly coupled with the research workflow, giving researchers more control over the description of their data. Preliminary experiments show that we can build on state-of-the-art technologies for search indexes, graph databases and triple stores to explore existing sources of linked data for our profile generation.

1 Introduction

Research data assets are very diverse, creating the need for a myriad of metadata descriptors adequate for each research domain. However, this diversity of metadata specifications makes the description of the datasets increasingly complex for researchers. While frequently unaware of the intricate complexities of data curation, they are a critical part of the data management workflow since they are both producers and consumers of research data [16]. Also, they are among the most benefited stakeholders in this process since it has been shown that linking properly described base data to research papers can lead to an increase in paper citation rate [20]. Current approaches at research data management usually call for a data curator [17,15], which is normally an expert on data management but not necessarily knowledgeable about the specific research topic of the data assets and, while this approach successfully yields high-quality curated data, considerable effort is required in this process [21]. Also, this curator-researcher interaction can only be carried out at specific moments throughout the research

[*] Supported by Ph.D. grant SFRH/BD/77092/2011, provided by the FCT (Fundação para a Ciência e Tecnologia).

workflow—in most cases at very late stages—while it should be carried out continuously, starting as early as possible [11].

In this paper, we propose a semi-automated data asset description model to make dataset description easier for researchers, acting as an automated assistant throughout the data curation process. The benefits include an off-load on data curators, while enabling the data management process to start earlier in the research activity schedule. Bringing data management to an earlier stage of the process may help researchers since annotated datasets are easier to share with similar research groups, reducing the traditional exchange of emails or other direct contacts whenever data exchange is necessary, for example. Our work aims at semi-automatically selecting the most appropriate metadata descriptors for a specific dataset. Recent developments in the field of the semantic web, linked data and graph databases make it technically viable to take the topology of a fairly large knowledge base (or a subset thereof) into account when making the suggestions; we will discuss some of these possibilities.

2 Linking research data assets

Linked data aims to establish connections between uniquely identified resources using unambiguous descriptors. Our current goal is to take advantage of the characteristics of linked data to recommend a set of descriptors that a researcher can use to annotate a specific dataset, without having to study complex domain-specific metadata standards. If we instead allow researchers more freedom to describe their datasets at an early stage (albeit roughly) and have a curator perform a validation step later, more datasets may actually enter the management process. Ad-hoc or duplicate metadata descriptor may be a problem, because it is unrealistic to believe that researchers would possess detailed knowledge about all descriptors in their domain, but we expect to mitigate this issue through our approach.

We regard all the datasets managed by a repository as part of a graph in which a dataset is a vertex, and the edges are the metadata descriptors (this model bears much resemblance to existing linked data knowledge bases, such as DBpedia). There have been recent moves towards transparent linking of resources through instances of properties taken from existing *semantic metadata* standards [1]. Semantic metadata consists in metadata descriptors that carry very specific semantics [8], something that is especially valuable for describing research datasets due to the specificity and diversity of notions involved in the research activities that make up the context of data production.

Figure 1 shows an example of a small subset of this “dataset knowledge base” containing 4 datasets. One is from the domain of chemistry (shown as ⬠), and three are from the domain of gravitational field analysis (shown as ⌓). This example highlights the inherent ambiguity between terms in these domains. Gravimetric analysis (in the chemistry domain) is sometimes referred to as *gravimetry*, and defined as “ the quantitative measurement of an analyte by weighing a pure, solid form of the analyte.”[19]. *Gravimetry* is a term which is also used in gravitational field analysis and measurement as "the measurement and analysis of the Earth's gravity field and its space and time variations" [4]. When searching for datasets by specifying a textual query, a system using pure text indexing might return all the datasets shown in the figure (albeit with different ranking scores). This creates the need to take into account other factors for disambiguating the concepts and contribute to the score, such as the topology of the local knowledge base subset in the *vicinity* of those datasets. The knowledge base depicted in Figure 1 uses 10 existing or purpose-created descriptors (or properties, in ontology vocabulary) from 5 different ontologies, shown in Table 1.

Descriptor	**Source ontology**	**Meaning**
`chem:formulation`	A purpose-built ontology for chemical experiments	The chemical formulation of the composition analysed in a chemical experiment
`dcterms:creator`	Dublin Core RDF Schema	"An entity primarily responsible for making the resource" [5]
`dbpedia-owl:abstract` / `dbpedia:caption`	DBpedia ontology	A short textual description of the resource / The caption of the image depicting the resource (indicated by the `foaf:depiction` property instance)
`muo:measuredIn`	Measurement Units Ontology[18]	"The unit used in the measurement of a particular quality value"
`data:precision`	A purpose-built ontology for specifying numerical parameters	The numerical precision of a measurement of a particular value
`foaf:depiction`	Friend Of A Friend Vocabulary Specification	"The depiction property is a relationship between a thing and an Image that depicts it." [3]
`grav:gravimeter`	A purpose-built ontology for gravitational field intensity measurement experiments	The identification of the gravimeter (device) used for the measurement of the intensity of a magnetic field
`dcterms:coverage`	Dublin Core RDF Schema [5]	"The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant."
`ref:cites`	A purpose-built ontology for dataset cross-referencing	"The subject resource makes a partial/whole reference to the object resource"

Table 1. Overview on the metadata descriptors used in the knowledge base depicted in Figure 1

3 Semi-automated application profile generation

Our goal is to explore diverse linked data sources and use the links between resources or *property instances* to select a set of properties considered more relevant for the annotation of a target research dataset. Figure 2 illustrates this challenge—the new dataset is the circle, which needs to be connected to the rest of the knowledge base via a set of property instances, shown as the dashed arrows. The ends of each arrow (highlighted by the question mark) are to be filled in by the researcher with either literals (in the case of datatype properties) or other resources in the knowledge base (e.g. datasets that are related to the one being added). The selection is intended as a recommendation: a descriptor may not be filled in by a researcher if he/she does not agree with the suggestion.

Link prediction is defined in social network analysis as "the task of inferring links in a graph G_{t+1} based on the observation of a graph G_t." [14], and is

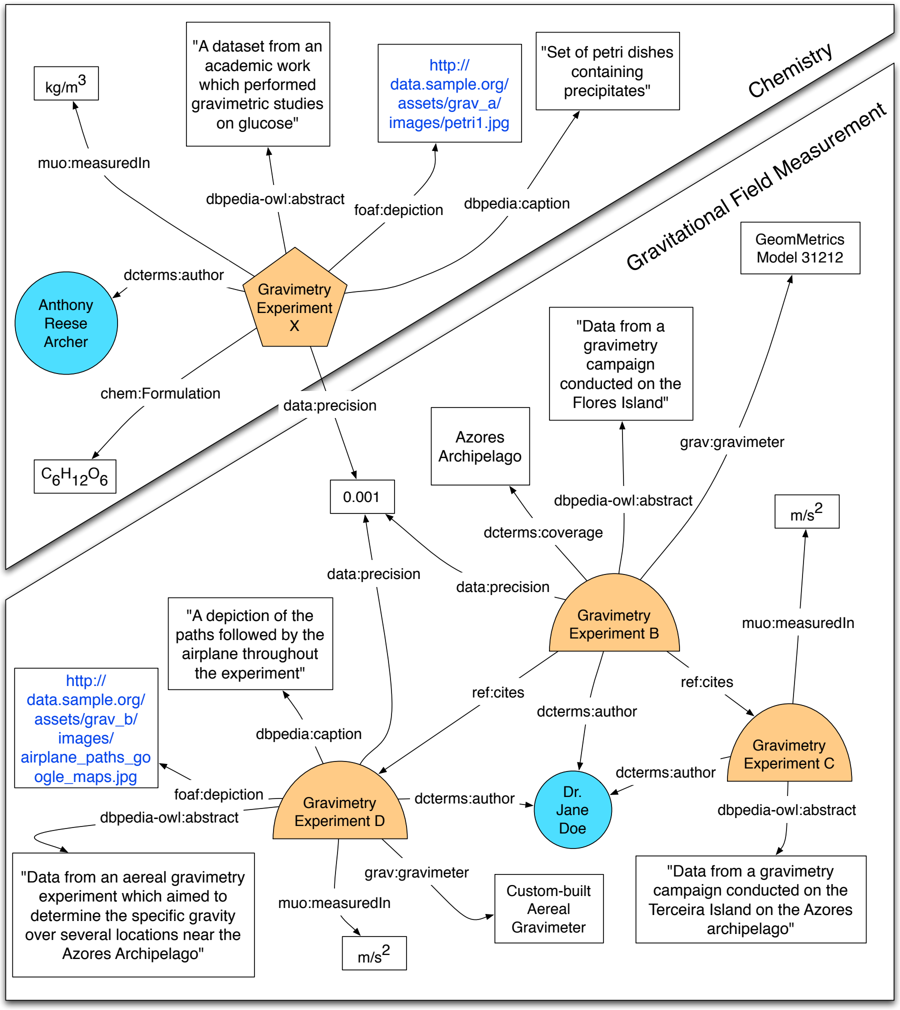

Chemistry
Gravitational Field Measurement
kg/m^3
muo:measuredIn
"A dataset from an academic work which performed gravimetric studies on glucose"
dbpedia-owl:abstract
http://data.sample.org/assets/grav_a/images/petri1.jpg
foaf:depiction
"Set of petri dishes containing precipitates"
dbpedia:caption
dcterms:author
Anthony Reese Archer
Gravimetry Experiment X
chem:Formulation
$C_6H_{12}O_6$
data:precision
0.001
GeomMetrics Model 31212
"Data from a gravimetry campaign conducted on the Flores Island"
grav:gravimeter
Azores Archipelago
dbpedia-owl:abstract
dcterms:coverage
m/s^2
data:precision
data:precision
Gravimetry Experiment B
muo:measuredIn
"A depiction of the paths followed by the airplane throughout the experiment"
http://data.sample.org/assets/grav_b/images/airplane_paths_google_maps.jpg
ref:cites
ref:cites
dcterms:author
dbpedia:caption
Gravimetry Experiment C
dcterms:author
foaf:depiction
Gravimetry Experiment D
dcterms:author
Dr. Jane Doe
dbpedia-owl:abstract
dbpedia-owl:abstract
grav:gravimeter
"Data from an aereal gravimetry experiment which aimed to determine the specific gravity over several locations near the Azores Archipelago"
muo:measuredIn
Custom-built Aereal Gravimeter
"Data from a gravimetry campaign conducted on the Terceira Island on the Azores archipelago"
m/s^2

Fig. 1. A set of interconnected research datasets

aimed at predicting relationships between people within a network [13] over time. Current link prediction approaches can combine nodes' content (*aggregated features*) [9] with *topological features* [10,6,14].

We approach property recommendation as a sequence of the following four steps, the first of which relies on a large textual index to narrow down the exploration to a subgraph, and the others explore the topology and content of the subgraph to provide a set of recommended properties:

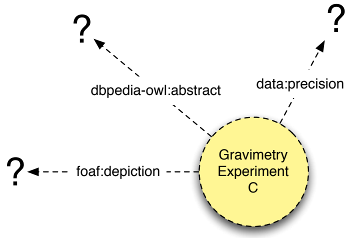


Fig. 2. A new dataset being added to the dataset knowledge, and its recommended properties

1. **Setting the scope of the base data used in the recommendation**
   Performing an analysis over the whole graph at runtime is technically inviable, so a subset of the knowledge base must be selected to serve as the basis for the recommendation. This sub-setting operation can be performed based on a series of *keywords* explicitly collected from the researcher or obtained from any indirect process. The keywords are ran as a query against a free-text index built over the datatype properties (string literals) of the resources. The most highly-ranked resources are likely to be relevant in the user's research domain, and will therefore constitute the initial sub-graph over which the next steps in this process can operate.
2. **Expanding the connections between the nodes within the selected scope**
   The selection on the first step is based on the textual content of the resources on the graph's vertexes. Complementary information comes from the graph topology, and we can expand connections on the subgraph to enrich it through the properties connecting the resources in the graph. This may bring into focus resources which are not directly linked via textual connections and may also reduce the number of disconnected graphs obtained from the search results. A challenge for this approach is, however, selecting the most effective depth of expansion—all connections by 1, 2 or more levels may be expanded, depending on the available processing power, for example. A more selective alternative is to expand only those connections between the resources that possess the highest hub/authority values within each increasingly large subgraph (using the hub/authority definition of the HITS algorithm [12]).
3. **Scoring the nodes according to their relative importance**
   The third step takes the expanded set of resources and scores them according to their connections to other resources, i.e. their position in the knowledge base's topology. Existing link analysis algorithm such as HITS or PageRank can determine the most influential nodes, based on their inbound and outbound connections. Links can be considered as unqualified, giving equal

weights to different properties, or different *a priori* weights can be provided according to the features of the resource and property on each link.

4. **Scoring the properties that relate the nodes according to their importance**
   After knowing which are the most prominent nodes contained in the subgraph of results, we may take their inbound and outbound properties and calculate their scores by combining the frequency with which they appear within the result subgraph, as well as the hub/authority value of the resources that they connect.

3.1 Experimental work

Our approach relies on a large knowledge base, containing a diverse set of resources connected by instances of many different properties. Since building a research data asset knowledge base using linked data is very time-consuming, we have decided to perform some experiments using data from DBpedia, an existing linked data knowledge base. Bizer et. al. specified the DBpedia project as a community-driven initiative designed to extract information from Wikipedia and make the information publicly available on the Web [2]. DBpedia offers several datasets that specify sets of resources and links between those resources, using concepts taken from ontologies such as FOAF (Friend of a Friend) [3], FreeBase [7] or DBpedia's own ontology. The information is generic, and thus not specifically targeted at interconnecting research datasets. The structure of the knowledge base is similar however, making these datasets an interesting workbench for initial experimentation of our approach.

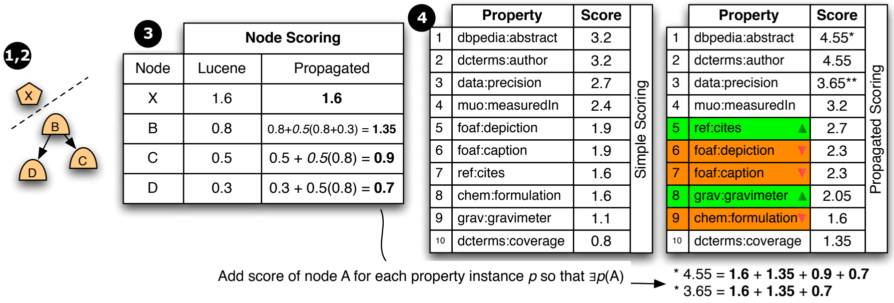
The knowledge base is separated into several datasets usable on their own, but that can be combined to provide additional knowledge and connections between resources. For example, there is a dataset which contains all the *extended abstracts* of the resources present in the knowledge base, which is the text portion that is normally found in the initial section of a resource's corresponding Wikipedia page. We have used this dataset in our small experiment since it contains a sizeable amount of running text. To establish the links between the resources, we have also included the *page links* dataset, which contains the representation of all the links between resources in Wikipedia.

Our prototype currently offers two search modes: a *simple* search mode scores resources according to their extended abstracts' and URI's textual content using the Apache Lucene library, and a *propagated* search, where the scores of resources within the results list that are connected via wikipedia links are boosted. The boosting is performed by simply adding the score of all *neighbours* multiplied by 0.5 (an arbitrarily selected "dampening" parameter) to the score of the resource itself. The boosting is symmetric, meaning that if resources A and $\{B_0, ..B_n\}$ are contained in the results list and there is a link between A and $B_i : i \in 0..n$, regardless of the link's direction,

$$score_{propagated}(A) = score_{simple}(A) + 0.5 * \sum_{i=0}^{n} score_{simple}(B_i), \text{ and}$$

$$score_{propagated}(B) = score_{simple}(B) + 0.5 * \sum_{i=0}^{n} score_{simple}(A_i)$$

The contribution of each of the “neighbours” to the final score of the resource is shown in the results list. At the present time, the application provides a basic ranking system for the resources (steps 1 to 3 the process outlined above), and is a first step in selecting the most appropriate technology stack for building a system that incorporates all four steps.



	Node Scoring	
Node	Lucene	Propagated
X	1.6	**1.6**
B	0.8	0.8+*0.5*(0.8+0.3) = **1.35**
C	0.5	0.5 + *0.5*(0.8) = **0.9**
D	0.3	0.3 + 0.5(0.8) = **0.7**

	Property	Score
1	dbpedia:abstract	3.2
2	dcterms:author	3.2
3	data:precision	2.7
4	muo:measuredIn	2.4
5	foaf:depiction	1.9
6	foaf:caption	1.9
7	ref:cites	1.6
8	chem:formulation	1.6
9	grav:gravimeter	1.1
10	dcterms:coverage	0.8

	Property	Score
1	dbpedia:abstract	4.55*
2	dcterms:author	4.55
3	data:precision	3.65**
4	muo:measuredIn	3.2
5	ref:cites ▲	2.7
6	foaf:depiction ▼	2.3
7	foaf:caption ▼	2.3
8	grav:gravimeter ▲	2.05
9	chem:formulation ▼	1.6
10	dcterms:coverage	1.35

Add score of node A for each property instance *p* so that ∃*p*(A) →

* 4.55 = **1.6** + **1.35** + **0.9** + **0.7**
* 3.65 = **1.6** + **1.35** + **0.7**

Fig. 3. Ranking the properties in the query results subgraph

Figure 3 highlights the last step in the four-step recommendation process. We can see the impact on the ranking of adding a simple topological signal, which is the existence of a property connecting two resources. In step 4 of the figure (and also of our method), the properties are ranked and the top-n properties are presented to the end-user. For a knowledge base G, two resources A and B and a property $P \in G$, the propagated score of P is calculated as $score_{propagated}(P) = \sum score_{propagated}(A)$, for all p such that p is an instance of property P and $p(A) = B$.

We have also written a multi-threaded loader program that adds the dataset triples to the knowledge base and indexes each resource’s extended abstract; it scales rather well, until it becomes bound by I/O speed. Knowledge base triples are currently stored in a fully de-normalized PostgreSQL database with only three tables to make querying as fast as possible, but it has become clear from our experiments that a different solution will have to be used, because the system becomes very slow (even on indexed text columns) due to the large number of vertexes (3.550.567, 2.66GB uncompressed) and even more due to the huge number of edges (145.877.010, 36.68GB uncompressed, wikipedia page links only). To cope with this kind of numbers we are currently experimenting with OrientDB and Jena, which are open-source storage layers especially designed to handle very large numbers of triples. A long-standing solution, Neo4j, was left out because it is made available under a rather restrictive Commercial/GPL license instead of the Open Source License Apache 2.0, which is less usage-restrictive.

4 Conclusions and future work

In this paper we have presented our goal of partially automating the selection of metadata profiles for the description of research datasets. By linking research data assets in a graph-like knowledge base, we can gather evidence of the importance of candidate descriptors from their use.

The proposed approach uses the contents of each node in the graph as well as its topology to perform a ranking of the properties within the knowledge base using state-of-the art link analysis algorithms, in order to rank and suggest properties that are considered adequate for the description of a dataset from a specific domain.

A small prototype for performing basic manipulation over a subset of DBpedia, a linked data knowledge base, has been built. It provides a first insight on the complexities of handling the massive amounts of information contained in these triple-based knowledge bases, as well as a preliminary study for the implementation of our proposed four stage recommendation model. Future work perspectives include an iterative development of the prototype to turn it into a testbed for gathering feedback from real users on the quality of recommendations—using generic resources at first and then real research data assets from existing research groups at U.Porto.

References

1. H. S. Al-Khalifa and H. C. Davis. The evolution of metadata from standards to semantics in E-learning applications. *Proceedings of the seventeenth conference on Hypertext and hypermedia - HYPERTEXT '06*, page 69, 2006.
2. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, Sept. 2009.
3. D. Brickley and L. Miller. FOAF Vocabulary Specification 0.98, 2010.
4. E. Calais. Gravity and the figure of the Earth. `http://web.ics.purdue.edu/~ecalais/teaching/eas450/Gravity1.pdf`, 2012.
5. Dublin Core Metadata Initiative. DCMI Metadata Terms. `http://dublincore.org/documents/dcmi-terms/#terms-creator`, 2012.
6. M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link Prediction in Social Networks Using Computationally Efficient Topological Features. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 73–80, Oct. 2011.
7. Google Freebase. Freebase Documentation. `http://wiki.freebase.com/wiki/Main_Page`, 2012.
8. K. Haase. Context for semantic metadata. *Proceedings of the 12th annual ACM international*, pages 204–211, 2004.
9. M. A. Hasan, V. Chaoji, and S. Salem. Link prediction using supervised learning. *SDM'06: Workshop on Link*, 2006.

10. Z. Huang. Link Prediction Based on Graph Topology : The Predictive Value of the Generalized Clustering Coefficient. 2006.
11. S. Jones, S. Ross, and R. Ruusalepp. Data Audit Framework Methodology, 2009.
12. J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
13. D. LibenNowell. The link prediction problem for social networks. *CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management*, (November 2003):556–559, 2004.
14. R. N. Lichtenwalter, N. Dame, and N. V. Chawla. Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction. (1019):1019–1028, 2012.
15. L. Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships. Technical report, 2007.
16. L. Martinez-Uribe and S. Macdonald. User engagement in research data curation. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, ECDL'09, pages 309–314, Berlin, Heidelberg, 2009. Springer-Verlag.
17. P. A. A. i. D. Media. Digital preservation strategies. *Workbook on Digital Private Papers*, pages 222–246, 2008.
18. Morfeo Project. Measurement Units Ontology. `http://forge.morfeo-project.org/wiki_en/index.php/Units_of_measurement_ontology`, 2008.
19. Oracle ThinkQuest. Information Internet : Chemistry Gravimetry. `http://library.thinkquest.org/10679/chemistry/gravimet.html`, 2012.
20. H. A. Piwowar, R. B. Day, and D. S. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3), 2007.
21. A. Treloar and R. Wilkinson. Rethinking Metadata Creation and Management in a Data-Driven Research World. *2008 IEEE Fourth International Conference on eScience*, pages 782–789, Dec. 2008.