Semi-automated application profile generation for research data assets

João Rocha da Silva	joaorosilva@gmail.com
Cristina Ribeiro	mcr@fe.up.pt
João Correia Lopes	jlopes@fe.up.pt

29th November 2012

Contents

- Introduction
- Goal
- Proposed approach
- Prototype
- Conclusions & future work

Introduction

Introduction

- Research datasets are valuable
- They document research publications by ensuring
  - *Reproducibility* and *Verifiability*
- They need to be managed, which requires adequate *curation*

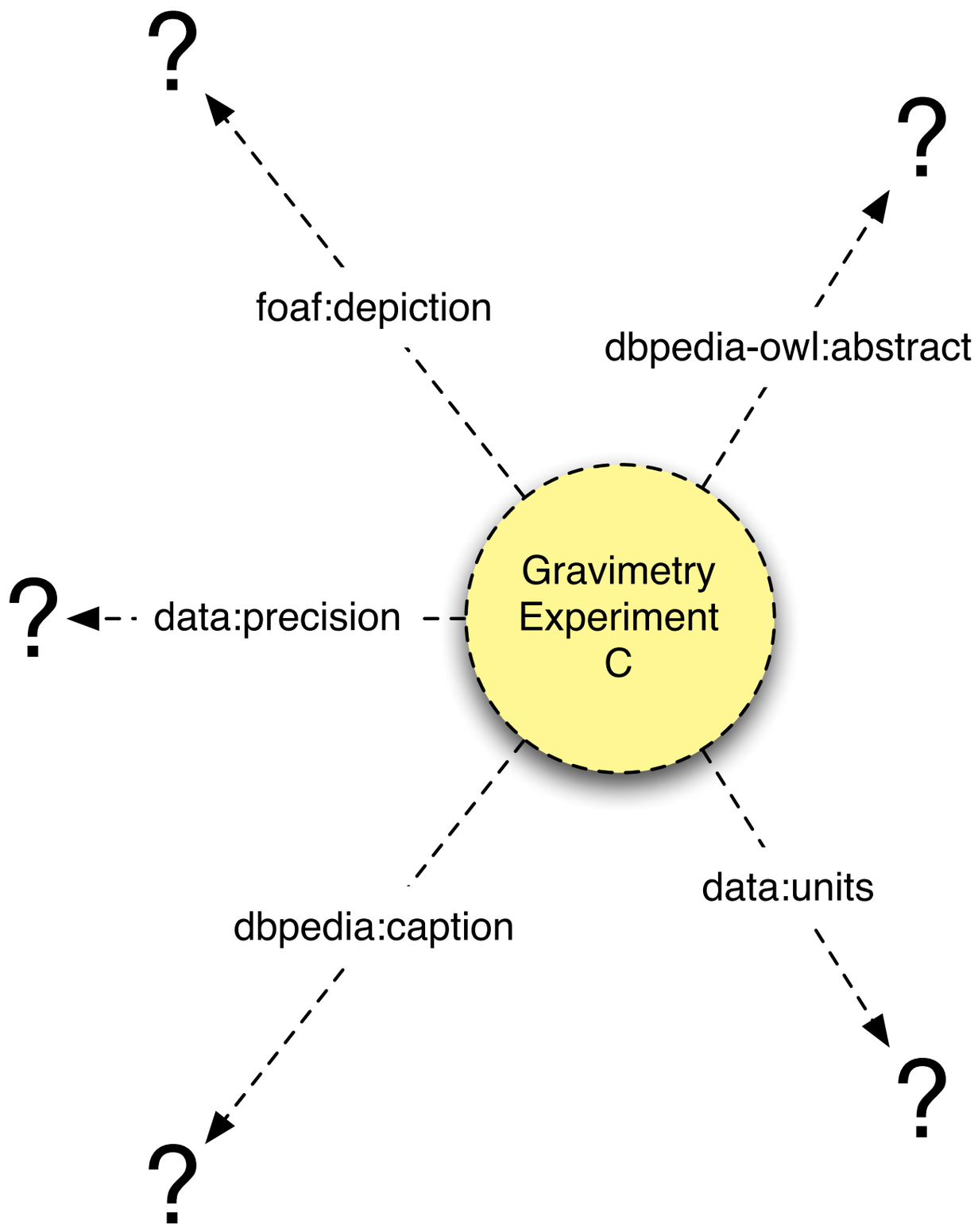
Introduction

Describing research data assets is not trivial

- There are many types of research datasets
- Each research domain may use different metadata profiles
- Researchers are not data management experts

Goal

- Assisting researchers in selecting the most appropriate descriptors for their research datasets, from domain-established metadata profiles

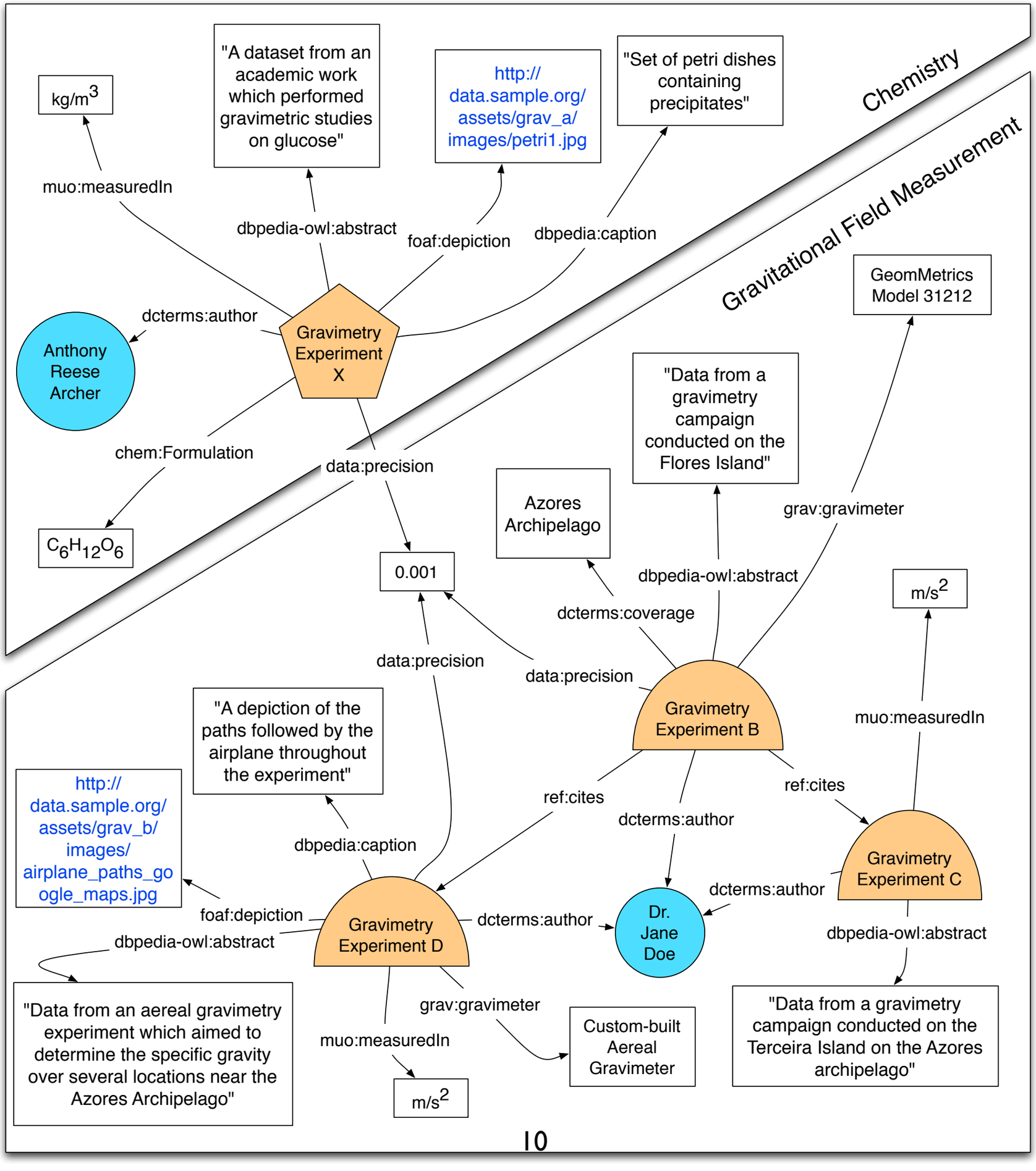

?
foaf:depiction
?
dbpedia-owl:abstract
?
data:precision
Gravimetry
Experiment
C
dbpedia:caption
data:units
?
?

Approach

Approach

1. Representation

- Representing research data assets in an interconnected graph (Linked Data)
- The vertexes of the graph represent the datasets themselves as well as related concepts
- The edges of the graph represent instances of metadata descriptors taken from domain-specific ontologies


Chemistry
Gravitational Field Measurement
kg/m3
"A dataset from an academic work which performed gravimetric studies on glucose"
http://data.sample.org/assets/grav_a/images/petri1.jpg
"Set of petri dishes containing precipitates"
muo:measuredIn
dbpedia-owl:abstract
foaf:depiction
dbpedia:caption
dcterms:author
Anthony Reese Archer
Gravimetry Experiment X
GeomMetrics Model 31212
chem:Formulation
data:precision
"Data from a gravimetry campaign conducted on the Flores Island"
C6H12O6
0.001
Azores Archipelago
grav:gravimeter
dbpedia-owl:abstract
dcterms:coverage
m/s2
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
dcterms:author
ref:cites
dbpedia:caption
Gravimetry Experiment C
foaf:depiction
dbpedia-owl:abstract
Gravimetry Experiment D
dcterms:author
Dr. Jane Doe
dcterms:author
dbpedia-owl:abstract
"Data from an aereal gravimetry experiment which aimed to determine the specific gravity over several locations near the Azores Archipelago"
grav:gravimeter
muo:measuredIn
Custom-built Aereal Gravimeter
"Data from a gravimetry campaign conducted on the Terceira Island on the Azores archipelago"
m/s2

Approach

2. Indexing

- Textual metadata elements (text Literals) are indexed
- E.g. for a given dataset, its `description` or `abstract` properties are indexed

Approach

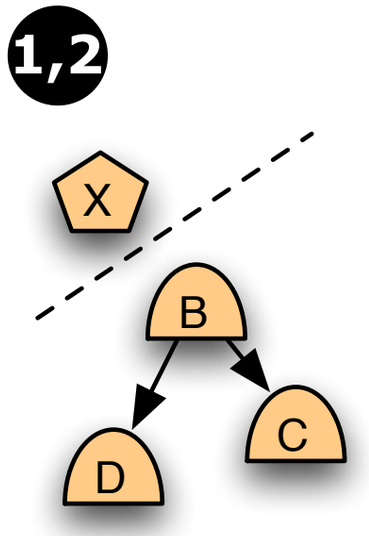
3. Retrieval

- A free-text query is run against the index
- A list of graph nodes is returned, complete with the *links* that connect them.
  - In ontology terms, a set of Class and Property instances

Approach

4. Descriptor recommendation

- Analysing the links between the nodes in the results page
- A first approach: ranking the descriptors by the number of times that they occur within the set of results
- Next: Calculate authority values and weigh the links according to the ranking of their linked nodes (HITS / PageRank)



3

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**

4

Simple Scoring

	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

Propagated Scoring

	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**

Approach

4. Descriptor recommendation (cont'd)

- Progressively expanding the result set to the immediate neighbours
- Re-rank the properties in the results set using the selected algorithms

Demo

DBpedia-based "search engine"

Conclusions & future work

- By analysing the topology of a graph consisting of datasets and related concepts we can recommend a set of descriptors for a new dataset
- The method is built on existing technology
  - graph databases
  - search indexes
  - link analysis algorithms

Conclusions & future work

- Ongoing
  - Improving Ranking : progressive expansion and algorithm variants
- Future work
  - Evaluating different variants of the techniques with real users
  - Including these techniques in a repository platform (e.g. DSpace)

Thank you

João Rocha da Silva	joaorosilva@gmail.com
Cristina Ribeiro	mcr@fe.up.pt
João Correia Lopes	jlopes@fe.up.pt