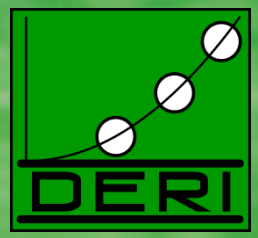


dcat: An RDF vocabulary for interoperability of data catalogues

Richard Cyganiak, Fadi Maali, Vassilios Peristeras

Agenda

- Why catalogue interoperability is important
- A survey of data catalogues
- Introducing the *dcat* vocabulary
- First experiments with integrated catalogue data
- Where to take this next?

Government data catalogues

- Now more than 30 catalogues online
- National
  - U.S., UK, Australia, New Zealand
- State level
  - New South Wales, California, Massachusetts, Maine
- Regional and local
  - New York, San Francisco, London, Vancouver, Kent County
- Both official and private initiatives

Catalogue websites do not unlock the full potential of the collected metadata.

Beyond catalogue websites

- Querying across catalogs
  - Overlapping regional coverage – U.S., California, SF
  - Supra-national catalogs – data.gov.eu?
- New user interfaces
  - Faceted browsing
  - Specialized UI for geographical/statistical/tabular sub-sections of a catalogue
  - Social annotation
- Bulk processing of datasets
  - Search indexes that inspect dataset contents
  - Update notifications

Current state of interoperability

- Most major catalogues do expose their contents in a structured format!
  - CSV
  - Atom feeds
  - RDFa
- But using this data is difficult
  - Different formats for each catalogue
  - Different metadata fields in each
  - Metadata fields poorly documented
  - Contents of metadata fields are inconsistent or do not match documentation

A survey of data catalogues

- In-depth review of seven catalogues
  - data.gov, data.gov.uk, data.gov.nz, data.australia.gov.uk, datasf.org, data.london.gov.uk, statcentral.ie
- Looking at *metadata*, not *into* the datasets

	data.gov	data.gov.uk	data.govt.nz	data.australia.gov.au	datasf.org	data.london.gov.uk	statcentral.ie
Size	1320	2879	251	69	132	189	227
Machine-Readability	Dataset	RDFa + dataset	Feeds	RDFa	--	--	--

Metadata structure

	General												Categorization		Dataset Access			Other		
	title	description	publisher	frequency	release date	update date	temporal coverage	geographic coverage	license	data dictionary	granularity	metadata update	theme	tags/keywords	dataset URL	format	size	references & citations	quality characteristics	data collection
data.gov	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	Y
data.gov.uk	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	-	Y	Y	Y	Y	-	Y	Y	-
data.govt.nz	Y	Y	Y	-	Y	-	-	-	Y	-	-	-	Y	Y	Y	Y	-	-	-	-
data.australia.gov.au	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
datasf.org	Y	Y	Y	Y	Y	-	Y	-	-	Y	-	-	Y	Y	Y	Y	-	-	-	-
data.london.gov.uk	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	Y	Y	Y	Y	-	Y	-	-
statcentral.ie	Y	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	-	Y	Y	Y

Consistency and availability

	Metadata consistency							Metadata availability				
	geographic coverage	temporal coverage	frequency	release date	update date	theme	tags	geographic coverage	release date	license	frequency	tags
data.gov	–	–	–	–	–	+	–	95	79		99	100
data.gov.uk	+	+	–	±	±	+	–	99	52	100	52	94
data.govt.nz				+		+	–		100	98		100
data.australia.gov.au	–	±	±	+	+	+	±	81	70	93	8	68
datasf.org		–	+	–		+	–		100		38	100
data.london.gov.uk	+		+	–	+	+	–	93	95	91	94	62
statcentral.ie	+	–	–	+	+	+	–		100		100	100

Direct download links

- Download links
  - Can go straight to the data (Excel, CSV, …)
  - Or to a splash page or license page
- % of direct links
  - data.london.gov.uk: 100%
  - data.gov: 95%
  - datasf.org: 10%
  - data.gov.uk: 7%

The *dcat* vocabulary

- Intended as interoperability standard
- Vocabulary expressed in RDF Schema
- http://vocab.deri.ie/dcat#
  - Vocabulary namespace
- http://vocab.deri.ie/dcat-overview
  - Misc information

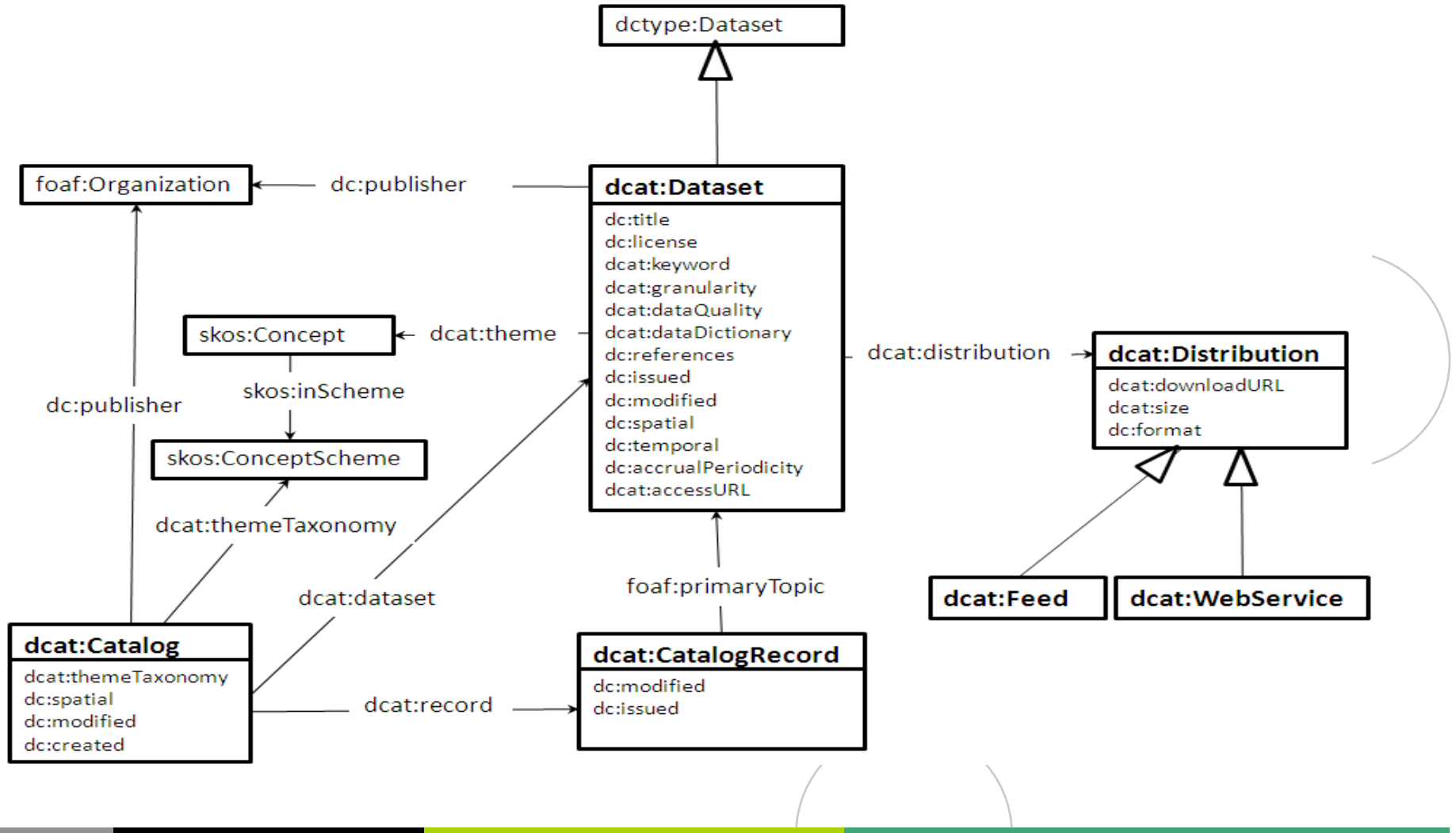
Design notes

- Hepp's Law: An integration ontology must not introduce distinctions that are finer than the distinctions made in the data to be integrated.
- Focus on the metadata fields that's available in all/ most catalogues
- Require no data cleansing before catalogue can be published in dcat
- Re-use Dublin Core, SKOS, FOAF whenever possible

Concepts

- dcat:Catalog
- dcat:Dataset
- dcat:CatalogRecord
- dcat:Distribution
  - subclasses dcat:Feed, dcat:WebService
- skos:Concept, skos:ConceptScheme
- foaf:Organization

Vocabulary overview



Initial experiments

- Set up a D2R Server over four catalogues
  - US, AU, SF, London
  - http://lab.linkeddata.deri.ie/govcat/
  - SPARQL interface: http://lab.linkeddata.deri.ie/govcat/snorql/
  - Links to Geonames, DBpedia

SPARQL across datasets


```
SELECT ?title ?url
WHERE {
    ?dataset a dcat:Dataset;
        dc:title ?title;
        dcat:theme :education;
        dcat:distribution ?distribution.
    ?distribution dcat:downloadURL ?url;
        dc:format ?format;
        dcat:size ?size.
    ?size dcat:bytes ?bytes.
    FILTER (?bytes<1048576 && ?format="text/xml").
}
```

SPARQL query with external data


```
SELECT ?title
WHERE {
    :data.gov dcat:dataset ?dataset.
    ?dataset dc:title ?title;
          dc:publisher ?agency.
    ?agency dbpedia:budget ?budget.
    FILTER (?budget>50000000000)
}
```

Benefits of the *dcat* standard

- Embedded metadata in catalogue web pages increases findability
- Enables decentralised publishing
- Enables federated search
- Will enable one-click download and installation of data packages
- Serves as manifest file for digital preservation
- Applications can be built once and work with multiple catalogues

Where next?

- Get feedback on the vocabulary, improve where necessary
- Write up a Guide to using *dcat*
- Explore how to use it with voiD, SDMX+RDF
- Get more catalogues to expose *dcat* format

- So far, everything happened in DERI, but we want to open this up. Where?