Facilitating NCAR Data Discovery by Connecting Related Resources

Toni Rosati
arosati@ucar.edu
SIParCS intern

Matt Mayemik
mayemik@ucar.edu
Mentor
Outline

- Purpose
- Background
  - DOI
  - Semantic Web/Web 3.0
- Types of Metadata Markups
  - RDF
  - Microdata
- Case Studies
  - NARCCAP
  - VisLab
  - ACADIS
- Evaluation Criteria
- Conclusions
Purpose

- Linking datasets, creators, and users
- Discoverability
  - increase the impact of funded research
Digital Object Identifiers (DOI)

- Officially names a resource
- NCAR uses the “DataCite” service to acquire DOIs
- DOI is essentially a stable, permanent URL
Semantic Web

- Web 1.0 – read static websites
- Web 2.0 – read websites and contribute some content (blogs, wikis)
- Web 3.0 – intelligent and connected
Resource Description Framework (RDF)

- Stand-alone depiction of metadata
- Triples that use URIs
- The Open Archives Initiative-Object Reuse and Exchange (OAI-ORE) is one particular implementation of RDF.
Microdata

- Schema.org is one implementation
- Announced in mid-2011
- Backed by Google, Yahoo!, and Bing
- Embedding into web pages

```html
<div itemscope itemtype="http://schema.org/Person">
  <span itemprop="name">Anakin Skywalker</span>
  <img src="http://www.imdb.com/media/rm906279680/ch0244573" itemprop="image" />
  <span itemprop="jobTitle">Fallen Jedi</span>
  <span itemprop="parent">Luke Skywalker</span>
</div>
```
Case Studies

- The North American Regional Climate Change Assessment Program (NARCCAP) dataset and related publications
- The Palmer Drought Severity Index (PSDI) animation and image files from NCAR’s Visualization Lab (VisLab)
- The multidisciplinary data types and formats from the Advanced Cooperative Arctic Data and Information Service (ACADIS)
NARCCAP - RDF

- DOI created through “DataCite”
- Metadata was already XML standardized, so could fit into OAI-ORE RDF structure

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:datacite="http://purl.org/spar/datacite/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterm="http://purl.org/dc/terms/"
```
NARCCAP - Microdata

- Access to edit the HTML
- Couldn’t use the “Data Cite” metadata file already created (except as a resource)
- Vocabularies needed to be remapped to the new schema
VisLab - RDF

- Metadata was scattered
  - needed to be hand-created
- Does not match directly to DataCite
- Needed to manually tweak a lot of the coding to link all the resources
VisLab - Microdata

- Access to edit the HTML
- Metadata just needed to be marked up according to mapping I already did.
ACADIS - RDF

- Lots of metadata already there
- Metadata not mapped to OAI-ORE standards
ACADIS - Microdata

- Fairly straightforward process to mark up the HTML
- But these keywords are not in Schema.org vocabulary
Evaluation Criteria

- Ease of use
- Available vocabularies and standards
- Relation to data citation standards
Ease of Use

- **RDF**
  - Requires a triple store
  - Can be overwhelming
  - Based on XML

- **Microdata**
  - Any webmaster can use schema.org
  - Can be implemented on any webpage
  - Based on HTML
Available vocabularies and standards

- Microdata is lacking in specific vocabularies for academia and datasets.
- RDF has unlimited vocabularies
  - including standards for locations and datasets
- Vocabularies can be registered
- Customization
Relation to data citation tools and methods

- Data citation tools (DataCite) use XML
- Data citation metadata doesn’t map well to Microdata
- RDF is more scalable to link resources than Schema.org
- Needs more than just search engine results
Conclusions

- The official naming of a dataset with a DOI allowed each markup schema to link resources better.
- Requires access to edit HTML.
- Enterprise vs piecemeal.
- Vocabularies.
- More than just discoverability.
- Both formats benefit NCAR, especially if there is a central triple store connecting all resources.
Questions?