Digital Public Space: Publishing Datasets

BBC Archive Development

Mo McRoberts, April 2012

1. Organise your data into sets.

BBC Archive Development

Implications

- Your data should ideally exist within a conceptual hierarchy (even if it's a single-level hierarchy).
- The aim is to make it easy for consumers to discover and use your data, which rich descriptions and links make possible.

- Express subsets as subsidiary resources, but keep the canonical item URIs at close to the top level as is reasonable.

- You might wish to think about organising these hierarchies around conceptual classes: e.g., `/articles`, `/books`, `/places`.

II. Use the Vocabulary of Interlinked Data (VoID) to describe those sets.

BBC Archive Development

Implications

- Publish documents at the root dataset URIs which describe the sets.
- Include information about URI patterns, endpoints, and links to example resources and subsets.
- The document is the dataset: e.g., `/items` is an instance of `void:Dataset`.

III. Make discovery easy.

BBC Archive Development

Implications

- If you can, publish a dataset description at your site root and at `/.well-known/void`.
- Within your sets, include descriptions of your text search and SPARQL endpoints, if you have them.
- Describe any data dumps that you make available.
- Arrange your sets so that clients can traverse them and retrieve their contents.

- If you're able to, include links to each of the items within the set using `rdfs:seeAlso`.

- To paginate, link to the first, next, previous and last pages in the set using the XHV terms (e.g., `xhv:first`).

- Order your data by most-recently-modified first to prioritise updates when consumers iterate the set.

- If you have data dumps available, link to them with `void:DataDump`.
- Include a description of the dump (the target of that link) detailing the creation/ modification dates and MIME type of the dump resource.

- Provide dumps as `.zip` files or gzipped single-file RDF/XML (amongst other formats).
- Within `.zip` files, put an RDF/XML file named `index.rdf` at the root which describes the resources in the dump using relative paths.

- Where subsets are organised around classes, describe them using `void:classPartition` and `void:class` if you can.
- Otherwise, use `void:subset` to reference them.
- In subsets, link back to the parent using `void:inDataSet`.

- If possible, use the Semantic Web extensions to your Sitemap to describe the datasets (alongside your VoID descriptions).

IV. Describe your items.

BBC Archive Development

Implications

- If possible, include information about each of the items in the sets which contain them.
- There's little sense in including *all* of the information about something — consider what you would typically present in a browsing interface.

- Where you include depictions of items, try to describe those image resources — the MIME types, and dimensions (using `exif:imageWidth` and `exif:imageHeight`).

- Rights matter! Include copyright and licensing information in the dataset descriptions.
- Publish rights information for both the data in the documents and (where applicable) the things described by those documents.
- The DMCI Metadata Terms schema includes predicates to aid this, and for many sets the Creative Commons ontology may also be useful.

V. SPARQL and data dumps are nice-to-have.

BBC Archive Development

Implications

- The primary aim is building a web of linked and linkable data.
- Don’t assume all consumers will want to only use your data, nor ingest it all into their own triple-stores in order to process or run queries upon it.

- SPARQL, search endpoints and data dumps are really useful features which enable a variety of interesting applications and they're worth providing if you can — but not at the cost of data you can link to.

Resources

BBC Archive Development

- http://vocab.deri.ie/void
  - Vocabulary of Interlinked Datasets (VoID)
- http://vocab.deri.ie/void/autodiscovery
  - VoID Autodiscovery via a RFC5785 `.well-known` resource.
- http://purl.org/NET/mediatypes
  - Linked data for MIME types (for use with `dct:format`)

- http://dublincore.org/documents/dcmi-terms/
  - DCMI Metadata Terms
- http://www.w3.org/2003/12/exif/
  - Exif RDF Schema

- http://dublincore.org/documents/dcmi-terms/
  - DCMI Metadata Terms

- http://www.w3.org/2003/01/geo/
  - Basic geo (WGS84 lat/long) Vocabulary

- http://creativecommons.org/ns
  - Creative Commons Rights Expression Language
- http://sw.deri.org/2007/07/sitemapextension/
  - Semantic Web Crawling: A Sitemap Extension