Data Curation & Preservation IG Summary

Arnold Rots for Alberto Accomazzi

Spring 2012 IVOA Interop Meeting
Urbana-Champaign, IL

Monday 21	4:00-5:30
Tuesday 22	11:30-12:30

Unified Astronomy Thesaurus: Motivation

- Growing interest among IVOA members, publishers to create an interoperable Thesaurus
    - IOPP, AIP each developed Physics and Astronomy Thesauri
    - IVOA updated IAU93 Thesaurus into a draft IVOAT
    - SIMBAD, VOTheory, ADS all using/developing frameworks to define and relate concepts
- N. Gray submitted a JISC Proposal, which while not funded, provides a blueprint:
    - identify & resolve incompatibilities between (astronomy parts of) existing thesauri
    - ‘productise’ the resulting thesaurus (need ‘full’ + ‘core’ division?)
    - develop a stable maintenance process, mediating between publishers’ and community’s needs and expertise (technical and procedural problems)
    - identify and prototype exemplar applications

Unified Astronomy Thesaurus: Use Cases

- A. Accomazzi discussed ADS use cases:
  - Text Mining: look for concepts in literature, expose to users
  - Enhanced Searching: find all possible instances of a concept via synonyms, hyponyms
  - Document Classification: assign consistent keywords to documents across the entire corpus
  - Faceted Views: filter search results, focus or broaden concepts, results
- S. Derriere discussed VO use cases for Semantic Applications:
  - Registry: standardize metadata used in Resource descriptions
  - Annotations, Tagging: use RDFa, microformats to embed semantics in descriptive metadata about VO resources
  - SKUA: framework to create and share annotations

Unified Astronomy Thesaurus: Plan

- Without JISC support, a plan B needs to be thought out, and community participation will be essential.
- Cleanup: ADS working with publishers, Library community to reconcile the astronomy portions of IOPP & AIP Thesauri with IVOAT. Help needed!
- Process: we want this to be a community effort but will need clarity on editorial responsibilities, policies
- Maintenance: essential to keep Thesaurus updated via to include terms from search logs, text mining. Help needed!
- Platform: still looking for a web-based, distributed platform supporting community input and curatiorial roles. Suggestions welcome!

Long-Term URIs

- N. Gray presented an update on his long-term URI note: http://www.astro.gla.ac.uk/users/norman/ivoa/long-term-uris.html
- Motivation: SemWeb and Linked Data technologies rely on URIs being stable, and should be dereferenceable
- DOIs work ok for individual resources, but lack of “scribbability” and hierarchy make them difficult to use for structured knowledge bases such as Vocabularies
- Proposal: piggy-back on existing persistent URL schemes, e.g.: http://purl.org/astronomy
- VOTheory already using the scheme above for their vocabularies, should this become an IVOA standard or best practice?

Metadata and Registries

- A. Rots presented a proposal calling for an enhanced level of resource metadata to support discovery and provenance tracking
- Arnold also suggested the creation of a data product registry recording individual data products metadata and persistent identifiers assigned to them
- A brief discussion made it clear that this issue needs to be debated within the Registry WG and with input from the VAO data discovery team and other projects
- The general consensus was that the current resource metadata is adequate for homogeneous collection, but is not satisfactory for heterogeneous data products (e.g. Vizier tables)

Libraries and Curation

- R. Wagner presented data curation efforts within UCSD’s libraries
  - Library manages Digital Asset Management System, supports finding agencies mandates (e.g. NSF data management plan)
  - UCSD effort leverages UC Digitial Library Infrastructure for DOIs, archiving
  - Need help? Talk to your librarian!

The Astronomy Dataverse

- A. Muench presented the Astronomy Dataverse, a web-based repository supporting publication, citation and discovery of research datasets
  - Collaboration between the CfA “Seamless Astronomy” team and the Dataverse Network team at Harvard
  - Goals: allow publishing of data, simplify citing, prevent “leakage” of unlinked resources
  - Supports branding, versioning, minting of persistent IDs
  - In the works: VAO plugin to Dataverse, connection to VOspace, ADS