VIRTUAL ASTRONOMICAL OBSERVATORY

Research Data: Creating, Linking, Preserving

Arnold Rots, SAO

The VAO is operated by the VAO, LLC.

Creating: Research Environments

- Of particular interest: collaboration tool boxes
- There are a number of environments, like:
  - Google groups, projects
  - Islandora
- Choice depends on needs (e.g., distributed collaborative group analyzing a shared dataset) and comfort level (personal preferences)

Linking: Objective

- Link astronomical digital objects to each other
- Make objects and links accessible to tools that allow:
  - Searches
  - Discovery
  - Analysis
- Astronomical digital objects are, in this context:
  - Publications – in a wide sense
  - Datasets – in a wide variety of places
  - Information on physical objects – as in NED and SIMBAD

Linking: Dataset Identifiers

- Use of dataset identifiers needs to be firmly established and expanded
- This will allow expansion of the semantic linking between datasets:
  - Papers
  - Other published materials
  - Observational datasets
  - User-contributed data
  - Theoretical data
  - Simulation data
- Users to be encouraged to embed dataset identifiers whenever and wherever feasible and appropriate

Semantic Linking and Tools

- Build a semantic knowledge store based on the harvesting of dataset identifiers and of key information contained in available datasets
  - This will provide the infrastructure needed for developing semantically enabled applications

- Build an interface enabling a seamless search on publications, objects and datasets based on the knowledge base described above
  - Such an interface will allow users to drill-down or expand a view of any of the three domains (objects, datasets and literature) based on the connections between them.

Problems with Data in the Literature

- Data are not published following rules or standards
- Data are often “published” in personal web sites with URLs in journal articles (sometimes footnotes)
- IAU rules for source nomenclature are often not followed
- Data “behind the plots” are not part of traditional publications
- e-journals keep changing formatting/markup

Preservation of Astronomical Objects

- Objects that are currently available (reliably):
  - Observational datasets in existing datacenter and observatory archives
  - Publications in, or accessible through, the ADS
  - Database repositories like NED and SIMBAD
  - Existing trusted repositories for user-contributed datasets and published materials
- What's needed:
  - Repositories for processed data and data used in publications
  - An architecture for integrating existing and future repositories of contributed objects
  - Encourage the development of links and provide tools for that

Preservation

- Trusted repository functions:
  - Storage
  - Proper metadata
  - Reliable access
  - Curation
  - Authentication
- Preservation metadata cover:
  - Authenticity
  - Original arrangement
  - Integrity
  - Chain of custody and history
  - Trustworthiness

IVOA Services for Preservation

- Data Management Plans:
  - Publish requirements, template
- Linking tools
- Repositories:
  - Integration of distributed repositories
  - Encouraging their use

  - I would like to suggest that iRODS might be an option for integrated trusted repositories:
    - integrated Rule-Oriented Data System
    - Developed by DICE (Data Intensive Cyber Environments) groups at UNC & UCSD, led by Reagan Moore
    - A more mature successor to SRB (Storage Resource Broker)

iRODS

- Data grid management system
- Download and install in 30 minutes under BSD license
- Provides preservation and curation services through a rule-based system:
  - authentication, integrity, chain of custody, trustworthiness, …
- Platform-independent
- Supports various configurations:
  - Master/slave grids, central archives, chained grids, deep archives
- NARA, LSST, SHAMAN, North Carolina Digital Reposit.
- Draft of paper presented at US Digital Data Preservation workshop:
  - http://ddp.nist.gov/workshop/papers/02_02_NIST-irods.doc