Astronomy 3.0 Style

Alberto Accomazzi
NASA Astrophysics Data System

17 Feb 2010
Library and Information Services VI, Pune, India
Overview

• The Scientific Lifecycle model
• Resource metadata in Astronomy
• What is Web 3.0 and how it can help
• A proposal for interlinking Astronomy resources
• Ongoing work and future plans
The Big Picture

- Scientific research requires repeatability
- The lifecycle of a research project should be documented by capturing all artifacts
- Data, Processes, Results need to be properly described, accessible, and linked together
- Provenance information should be attached to metadata throughout the process
plans are produced. The calibration stage can include equipment calibrations both in the laboratory and in the field, as equipment often is recalibrated to reflect field conditions. Artifacts such as lists of equipment taken into the field and the condition of that equipment may be produced at the planning stage or may be documented more fully during and after data collection. In the data capture phase, records on the initial placement of sensors, movement of sensors, and decisions made in the field may be produced. This array of contextual information about a field study can be essential documentation for interpreting results and for planning subsequent field research. To account for this set of scientific artifacts, Figure 3 integrates the life cycle of environmental sensing research (of Figure 2) with the larger range of scientific products identified in our study of this scientific community.

Figure 3. The integrated scientific life cycle of embedded networked sensor research.

Pepe et al, JASIST (2009)
Plans are produced. The calibration stage can include equipment calibrations both in the laboratory and in the field, as equipment often is recalibrated to reflect field conditions. Artifacts such as lists of equipment taken into the field and the condition of that equipment may be produced at the planning stage or may be documented more fully during and after data collection. In the data capture phase, records on the initial placement of sensors, movement of sensors, and decisions made in the field may be produced. This array of contextual information about a field study can be essential documentation for interpreting results and for planning subsequent field research. To account for this set of scientific artifacts, Figure 3 integrates the life cycle of environmental sensing research (of Figure 2) with the larger range of scientific products identified in our study of this scientific community.

Figure 3. The integrated scientific life cycle of embedded networked sensor research.

Pepe et al, JASIST (2009)
Why do we care?

- attribution: links allow one to follow the evolution of research activities
- aggregation: links allow one to find related resources, ancillary data, versioning
- preservation: knowledge about the origin of metadata as important as the data
- discovery: the presence (or absence) of links carries semantic meaning
The case of Astronomy

- Bibliographic Data & Metadata: Publishers, ADS
- Observational Data & Metadata: Archives, Observatories, Surveys
- Object Metadata: NED, SIMBAD, Vizier
- VO to “tie things together”
The Web of Astronomy
Links

• Established very early in the game thanks to agreement on identifiers (pre-web!)
• Dense, curated by individual projects on behalf of the community
• Links across domains (papers, data, objects)
• Provides many of the connections required by the science life-cycle model
Optical follow-up of new Small Magellanic Cloud wing Be/X-ray binaries

Title: Optical follow-up of new Small Magellanic Cloud wing Be/X-ray binaries
Affiliation: AA(School of Physics and Astronomy, Southampton University, Highfield, Southampton SO17 1BJ), AB(School of Physics and Astronomy, Southampton University, Highfield, Southampton SO17 1BJ), AC(School of Physics and Astronomy, Southampton University, Highfield, Southampton SO17 1BJ), AD(School of Physics and Astronomy, Southampton University, Highfield, Southampton SO17 1BJ), AE(South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa), AF(Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA), AG(Universities Space Research Association, X-ray Astrophysics Laboratory, Mail Code 662, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA), AH(South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa), AI(South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa)
<table>
<thead>
<tr>
<th>Data Related Directly to Object Names</th>
<th>Site/Service</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Query SIMBAD by primary NED object name</strong> -- <strong>MESSIER 031</strong></td>
<td>SIMBAD (CDS, Strasbourg, France)</td>
</tr>
<tr>
<td>Revised New General Catalogue -- NGC 0224</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>Uppsala General Catalog -- UGC 00454</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>Original Zwicky Catalog -- CGCG 535-017</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>Query UZC Spectral Archive (60 arcsec search radius)</td>
<td>Updated Zwicky Catalog Data (Harvard/SAO)</td>
</tr>
<tr>
<td>Morphological Catalog of Galaxies -- MCG +07-02-016</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>The Third Bologna Survey -- B3 0040+409</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>2MASS Extended Source Images (JHKs) -- 2MASX J00424433+4116074</td>
<td>NASA/IPAC Infrared Science Archive (IRSA)</td>
</tr>
<tr>
<td>IRAS Point Source Catalog -- IRAS 00400+4059</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>IRAS Faint Source Catalog -- IRAS F00400+4059</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>Catalogue of Principal Galaxies -- PGC 002557</td>
<td>VizieR Catalog Query (U.S. mirror, CfA/Harvard)</td>
</tr>
<tr>
<td>Retrieve mean data from LEDA -- PGC 002557</td>
<td>The Lyon/Meudon Extragalactic Database (LEDA)</td>
</tr>
<tr>
<td><strong>Query GALEX (NUV/FUV) Mission Archive (6' search radius)</strong> -- <strong>MESSIER 031</strong></td>
<td>GALEX Mission Data Archive at MAST</td>
</tr>
</tbody>
</table>

**General Archive Resources -- All queries centered at 00h42m44.3s, +41d16m09s (J2000)**

<table>
<thead>
<tr>
<th>Site/Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Optical and UV Mission Archives (Default search radius)</td>
</tr>
<tr>
<td>Query High Energy Mission Archives (Default search radius)</td>
</tr>
<tr>
<td>Explore resources with DataScope (15' search radius)</td>
</tr>
<tr>
<td>Visualize Coverage Map with IMPReSS</td>
</tr>
<tr>
<td>Retrieve 2MASS Atlas Images</td>
</tr>
<tr>
<td>Retrieve IRAS ISSA Images</td>
</tr>
<tr>
<td>1-D Coadd of IRAS Scans (ADDSCAN/SCANPI)</td>
</tr>
<tr>
<td>Retrieve NVSS Image</td>
</tr>
</tbody>
</table>
Some numbers for ADS

• citations: 40M (across databases)
• readership: 18M (90-day window)
• fulltext: 5M (journals, arXiv, ADS)
• astronomical objects: 240K (SIMBAD + NED)
• data products: 130K
• bibliographic groups: 200K
Bibliographic groups

- CfA
- ESO
- ROSAT
- HST
- IUE
- CFHT
- Chandra
- ISO
- PhysEd
- XMM
- USNO
- NRAO
- Gemini
- Keck
- Spitzer
Web 3.0

- The “Web of Data” or “Semantic Web”
- Describes resources in a formal way
- Machine-readable content throughout
- Based on Linked Open Data principles
- “Read, Write, Execute”
Linked Data principles

• resources are named via HTTP URIs
• metadata is open and in a standard format
• links between resources are typed
• built on the architecture of the web, no APIs
Links to Datasets

- have been based on reciprocal links between URLs curated by collaborators
- do not identify resources as unique URIs
- do not make use standard vocabularies to describe data or the types of links between them
- are not actionable by applications
How to get there

• grow the astronomy data cloud from the bottom-up

• identify our resources (data), name them uniquely, expose their metadata

• expose relationships between resources (links)

• what should be included: observing proposals, observational metadata, instrumental metadata, papers, catalogs, objects
Publication Metadata

• Authors
• Keywords
• Affiliations
• Citations
• Links to data, objects
• Readership
Observational Metadata

- Unique Dataset Identifiers
- Position, footprint and date
- Objects observed
- Wavelengths
- Instruments
- People, Proposals, Collaborations
A VAO Proposal

• Identify, collect and expose metadata for Datasets (Chandra, MAST, TSC, etc.)
• Create typed links between Data products, Bibliographic and Object databases
• In the process, create knowledge base about Instruments and Services
• Incorporate all of the above in a “metadata store,” exposing resources using semantics
Sample Applications

• A view of bibliographies with facets based on objects, wavelengths, keywords, etc

• A recommender / alert service for data products, objects

• A UI integrating views of bibliographies, objects, and datasets interacting together

• An application computing metrics on data, objects based on citation and use of papers
Views on Papers

Tuesday, February 16, 2010
Recommendations?

Collaborations

My data

Recent data

Hottest data
APOD Application
Ongoing Efforts

- IVOA / ADEC dataset identifiers
- IVOA data models, vocabularies for keywords and UCDs, ontology of object types
- MSR / WWT work on mashups
- ADS work with VAO, text mining, faceted views
The Role of Libraries

- Astronomy librarians have been at the intersection of archives and journals
- They are experts in metadata curation and search, maintain bibliography-data links
- They have been collecting/generating much of this metadata already
- They understand the need for (and limitations of) metrics
A call to action

• Tell us what bibliographic and observational metadata you are collecting and curating
• Use our upcoming fulltext search on current literature to collect/annotate papers
• Drink the Web 3.0 coolaid
• We have freed our data, now let’s free your metadata!