VIRTUAL ASTRONOMICAL OBSERVATORY

Data Curation and Preservation Activities in the VAO

Arnold Rots
SAO
Joseph Mazzarella
IPAC

The VAO is operated by the VAO, LLC.

Creating, Linking, and Preserving ADO

- This is about ***Astronomical Digital Objects***, meaning:
  - Publications – in a wide sense
  - Datasets – in a wide variety of places
  - Information on physical objects – as in NED and SIMBAD

- We will leave the creation of ADOs to the users
- In order to make them useful for research by the community, they need to be:
  - ***Linked***
  - ***Preserved***

Linking: Objective

- Link ADOs to each other
- Make objects and links accessible to tools that allow:
  - Searches
  - Discovery
  - Analysis

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 that knowledge base
  - 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.

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

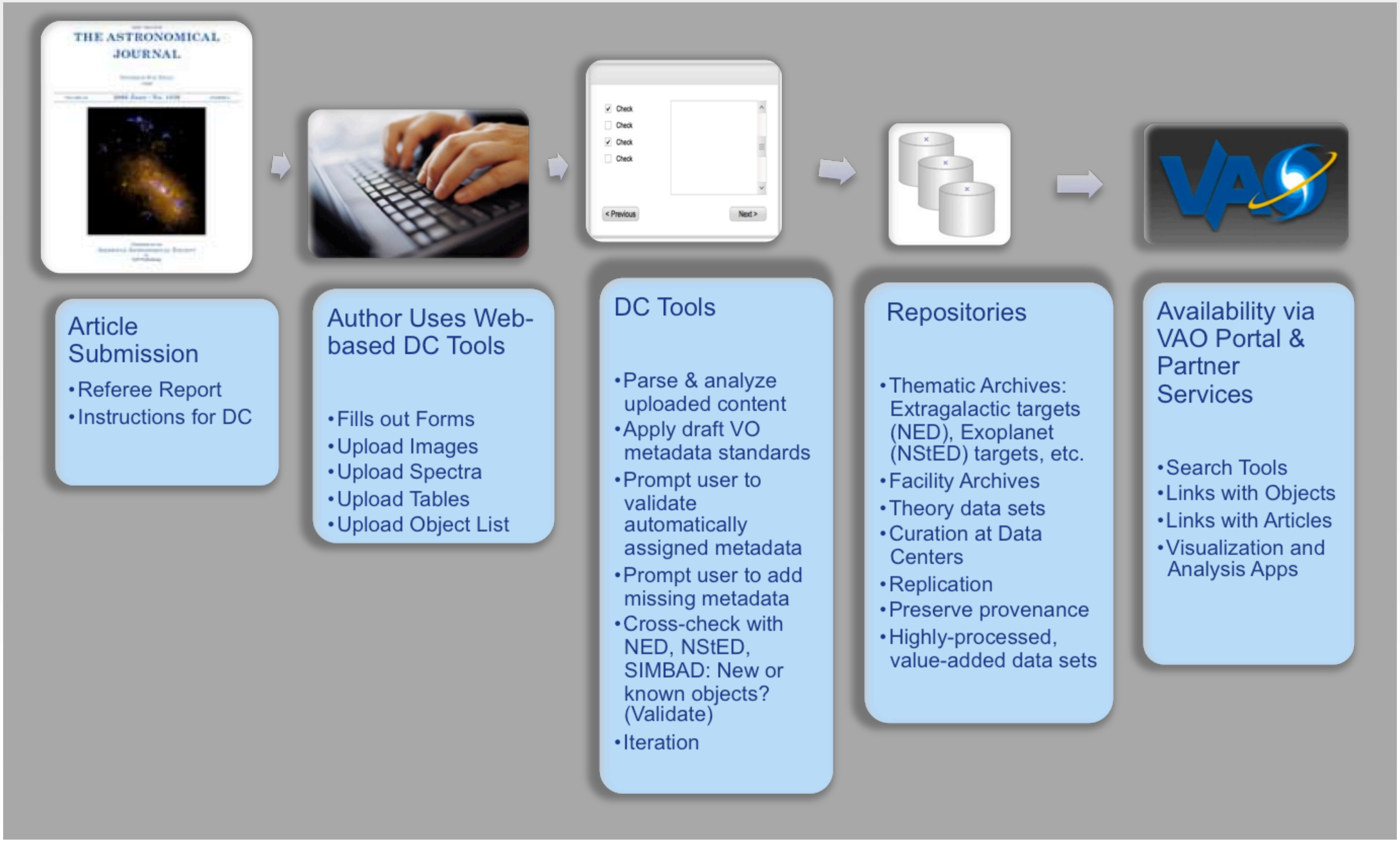
Problems with Data in the Literature

- 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

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

Data Capture Workflow (IPAC)



Article Submission

- Referee Report
- Instructions for DC

Author Uses Web-based DC Tools

- Fills out Forms
- Upload Images
- Upload Spectra
- Upload Tables
- Upload Object List

DC Tools

- Parse & analyze uploaded content
- Apply draft VO metadata standards
- Prompt user to validate automatically assigned metadata
- Prompt user to add missing metadata
- Cross-check with NED, NStED, SIMBAD: New or known objects? (Validate)
- Iteration

Repositories

- Thematic Archives: Extragalactic targets (NED), Exoplanet (NStED) targets, etc.
- Facility Archives
- Theory data sets
- Curation at Data Centers
- Replication
- Preserve provenance
- Highly-processed, value-added data sets

Availability via VAO Portal & Partner Services

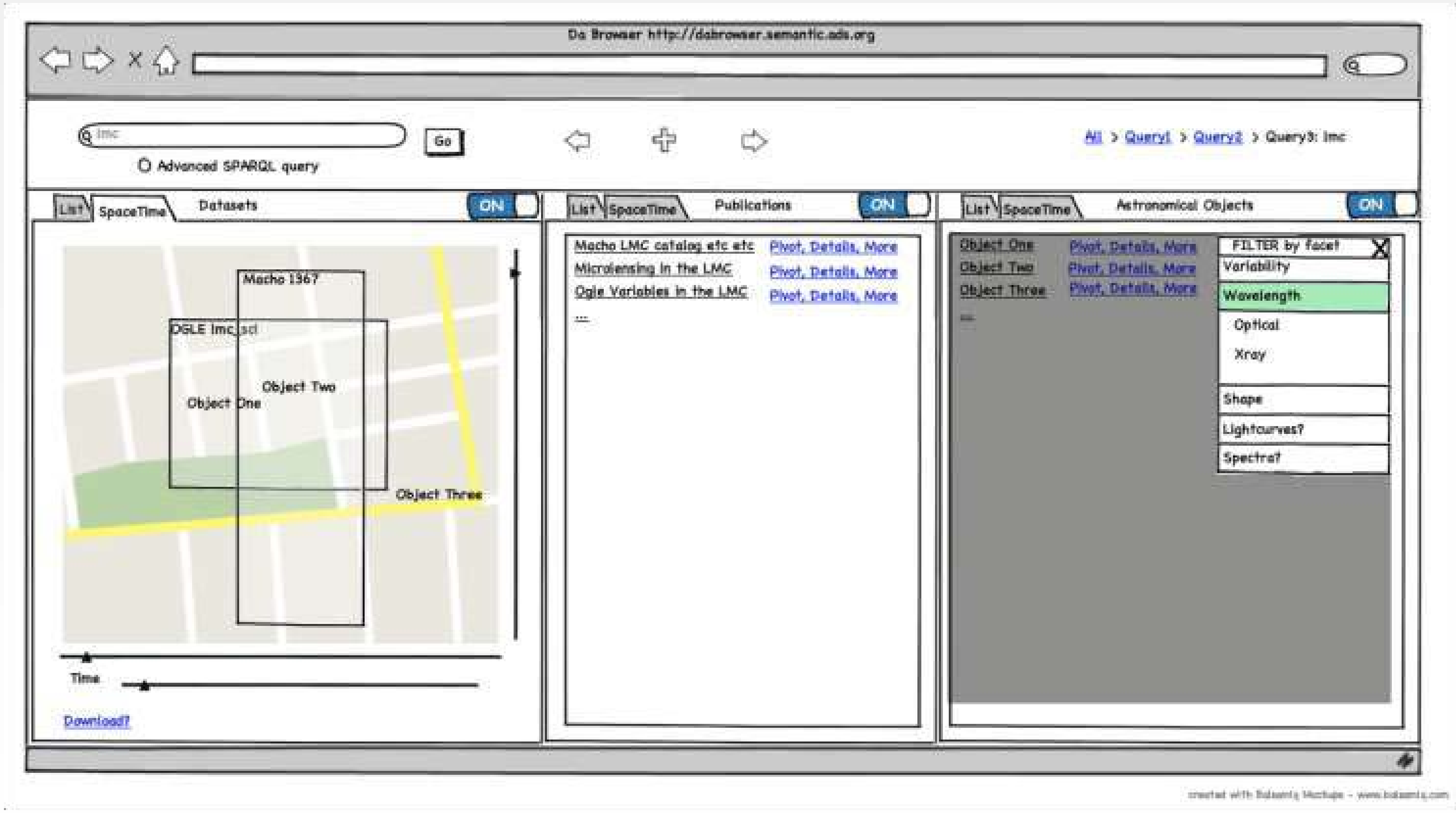
- Search Tools
- Links with Objects
- Links with Articles
- Visualization and Analysis Apps

Public Policy Mandate

- This ties in with a recent report by the U.S. National Academies entitled *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age:*
  - "Research data and other information integral to publicly reported results should be publicly accessible. Although legitimate reasons may exist for keeping some data private or delaying their release, the default assumption must e that research data and the information needed to interpret them will be publicly accessible in a timely manner to allow verification of findings and facilitate future discoveries."

Faceted Browsing Interface (SAO)

Datasets – Publications – Astronomical Objects



Macho LMC catalog etc etc	Pivot, Details, More
Microlensing in the LMC	Pivot, Details, More
Ogle Variables in the LMC	Pivot, Details, More
...	

Object One	Pivot, Details, More
Object Two	Pivot, Details, More
Object Three	Pivot, Details, More
...	

Start-up Activities for Preservation

- In a collaborative study and pilot project between SAO (ADS), IPAC (NED), AAS, ADEC members, and NSF Archives, we will:
  - Generate requirements for data publishing tool that can be integrated into a journal publishing workflow.
  - Create a prototype repository system addressing nomenclature, publishing, linking, and preservation of data sets (both user-supplied and archive-curated).
  - Collaborate with the Data Conservancy project and arXiv as that data capture project matures.

Initial Linkage Goals

- Topic-based literature search and faceted browsing of results through views on objects, object types, datasets, and wavelengths
  - Search literature by topic, refine results based on object types, data products and wavelengths.
  - Search sky by object name(s) or position and filter related literature by keyword(s).

Long-term Objective

- The long-term outcome of this work will be:
  - Comprehensive capture of unique and valuable data associated with journal publications utilizing assistance from authors at the time of publication, to include machine-readable object lists, tables, images and spectra where possible
  - Rapid and efficient integration of data and metadata into object databases and repositories for digital astronomical data objects
  - New search capabilities that utilize extensive metadata involving observation attributes, astrophysical object attributes and a rich set of topical keywords to facilitate powerful science queries

Planned Deliveries

- Data publishing tools:
  - Prototype in a year
  - Streamlined, operational system a few years down
- Develop a data storage plan
- Vocabulary and Ontology development:
  - Instrument knowledge base
  - ADS bibliographic model
  - NED/SIMBAD object ontology

Organizations and Personnel

- SAO:
  - Arnold Rots DC&P Lead
  - Alberto Accomazzi
  - Rahul Dave
  - Sherry Winkelman
  - Doug Burke
- IPAC
  - Joe Mazzarella DC&P Deputy Lead
  - Olga Pevunova
  - Marion Schmitz
  - Rick Ebert