Data Curation and Preservation

Opportunities for IVOA

- Develop guidelines for defining curation processing steps
    - Documenting collections
        - Processing and validation of data going into the collection
        - Metadata definition should cover this – seeing characterization, calibration, conditions of the observation, instrument set up,
        - How maintain the links to the separate calibration database (in the registry or in the data set. Need link pointing to the calibration file.
    - Making the collections available
    - Development of consensus on processing steps for given data type
        - Data centers build archives that are viable for the long term
        - Recommendation from IAU for preservation – need to review a copy
    - Feedback to the digital library community

- Management of links to software
    - Migration of software to new operating systems
    - Ultimate solution is archive that is independent of the original softare.
    - Software may be available to read. Source code is part of archive, but may not be executable
    - Implies that have defined all of the preservation processes and applied them
    - Need data encoding format that is self-describing – DFDL

Preservation for projects and missions defined above

How preserve data from individuals – build Dspace instance for astronomy?

- Preservation of images published in ApJ – opportunity for IVOA
- Repository for publishing images provided by IVOA?
- CDS is managing catalogs from papers – apply curation processes for ingesting the table
    - Standards for catalogs
    - Check metadata, numbers, consistency, description of the physical organization and meaning of columns
    - Do not enter anything without checking consistency
    - Need statement of processing steps. Processing done by librarians. Description checked by trained astronomer
- Implies standard description for tables for publications

There are some large-scale catalogs that have not been published. Active effort by CDS to capture catalogs. Astronomy and Astrophysics sends catalogs to CDS.

Scale of such catalogs should be billions of records:

- Journals provide more than 100 records per table
- Largest catalog is 1.5 billion (1,500,000,000)
- In aggregate have about 5-10 billion across 4000 catalogs

Have no similar system for digital data referenced in a journal (images, data cubes,
Scale for supporting image collections

- Expect 100 terabytes – quite feasible
- Challenge is a change in policy that the digital products will be archived

- Challenge is how provide adequate curation for seeing, calibration, processing, links to the original image
    - Extraction of proper metadata
    - Characterization of the original digital representation of data
    - Association of image with a journal article
    - Equivalent to publication standards for images
        - ApJ, A and A,
        - Funding agency requirements
        - Images have to be published as FITS files
        - Standard for the FITS header implementation

Groups interested in preservation

- CDS catalogs
- ADS
- NVO image collections
    - Form long term partnership with publishers (AAS, A and A, ….)
    - Integrate publishing with deposition in the archive

How validate the data quality

- Can standard processing be used to compare quality in image to the "standard" image at that wavelength
- Is there a way to improve the quality of the images or annotation of the images to make the image better for publication?
    - How validate FITS header (review, consistency, controlled vocabulary)
        - Could require conform to IVOA data model, UCDs
        - Do we have tools that provide the checks today? Not yer.
        - Minimum set of parameters, validated by software tools
    - How negotiate such standards with the publishers

Resolution by Ray Norris, 24 July 2003, adopted by IAU

- Recognising need full access to data,
- Considering technology exists, virtual observatory increasing access
- Recommends publicly funded data be made accessible after proprietary period, data not be subject to intellectual property rights, available for science usage, encouragement from funding agencies for deposition

- http://www.atnf.csiro.au/people/rnorris/WGAD/Resolution.htm

Interest Group orgnization

- Participants
    - Reagan Moore
    - Francoise Genova
    - Robert Hanisch
    - Pepi Fabbiano
    - Bob Mann
    - Ray Plante?
    - Gunther Eichhorn
    - Mike Watson (policy)
    - Michael Kurtz
    - Peter Quinn?

- Activities:
- Curation prolicies
    - Catalogs – experience from CDS for publishing tables
    - Images from individuals – need a candidate policy
        - Policies for images would be negotiated with publishers
        - IVOA role may be the creation of validation tools for FITS
        - IVOA projects role to provide storage for image
    - Need to coordinate with:
        - UCD working group for validating vocabulary
        - Data model working group for validating FITS
        - DAL link for metadata capture – minimum set of metadata defined by SIA and SSA protocols or equivalent

Assert that the validation tools will correctly check for correspondence to IVOA standards for long term preservation
Assert that the tools will solicit information needed for repair and creation of valid data sets

Community acceptance is critical, Depends on:

- Minimal metadata requirements
- Integration with manuscript submission process
- Integration in a real service
- User friendly submission

Support for preservation repository for images

- Expectations on scale of such a facility – 100 Terabytes
- Expectation on lifetime of the facility – 100 years
- Expectation on interactions with publishers – view of the preservation facility as a resource by the publishers
- Annotation support for researchers, provided by Fedora
    - How manage annotations, filter relevance, summarize
    - Support annotations on other collections?

Support for collections at risk

- IAU group for historical spectral data
- How provide support without impacting current projects
    - Ability to gain support from new sources
    - Space plasma physics data center is an example – CTPP
        - Took many years, case-by-case analysis
- How develop expertise needed for such curation
- How is appraisal done to decide collection worth preservation

Support for hardcopy

- Digitization of images
    - Appraisal
    - Standards for digitization process (resolution)
    - Standards for characterizing data quality
- Current effort at Harvard collection (some NSF support)
    - Scale is 300,000 plates
- Vatican has material
- Bulgaria assembling repository for scanned plates

Correlation between images and catalog records:

- Is an image correctly preserved if all objects have been extracted into records in a catalog - astrometry
- Requires major calibration effort (being tried for Harvard plates) – photometry
- Correspondence between the pixels and catalog records

Provenance metadata – should be part of Data Model, also in Registry metadata

- Quality of data
  - Calibration
  - Seeing
- Origin (processing steps used to create the digital entity)
  - Proper credit
- Addition of provenance in registry – lower priority
  - Service, resource, collection provenance, properties of the aggregation
  - Quick simple entry with standard vocabulary – select between options
- Digital library standards
  - Dublin Core
  - METS profile

Priority:

1. Policies for capturing digital data (images, graphical data based on digital representation) in an approved form
   - Robert Hanisch
   - Pepi Fabbiano
2. Standard metadata capture for catalog entries (CDS) – document describing practice
   - CDS tools for metadata validation correctly done
3. Preservation repository example (mechanism includes physical storage space, software to manage storage, digital library interface)
   - Tools for validating submissions
     - Pepi Fabbiano
   - Interactions with publishers
     - Robert Hanisch
     - Francoise Genova
   - Current tools available from existing repositories
     - FITS validation tool – very general compliance to standard
     - Semantic validation is needed
     - Compliance for access through SIA or SSA protocol – should be part of data model group, or data access layer
   - IVOA compliant repository
     - Reagan Moore
4. Interactions with existing preservation groups
   - IAU working group for digitization and preservation of photographic plates
     - Elizabeth Griffin,
     - Pisgah Astronomical Research Institute archive
     - Support for digitization of plates – what metadata do they generate
   - Digital Library preservation efforts
     - Fedora / DSpace, ….

Need to contact the following for their area of highest interest.

- Bob Mann
- Ray Plante
- Gunther Eichhorn
- Mike Watson (policy)
- Michael Kurtz
- Peter Quinn