Data Discovery:

Drilling down into Compound Data Objects

- Granularity and Persistent Identifiers

Arnold Rots
SAO/ CXC

Discoverability

- Scope: higher level data products
- Scattered over many repositories
- Requires
  - Persistent Identifiers (PID)
  - Registry infrastructure
  - Metadata standards
  - Metadata extraction tools
  - Provenance information
- But this is only part of the problem

Complex Cases

- Versioning
  - Purists will insist that PIDs are version specific
  - Growing realization that users often prefer current default
    - Improved calibration, more complete datasets, etc.
  - Version information can also be appended to the “root” PID
- Compound datasets or data objects
  - Made up of smaller components
    - Multiple files, multiple data objects, etc.

Drilling Down into Compound Data Objects

- From the user's perspective
  - For instance, a user interested in masses of galaxies
  - "Get me papers on galaxy mass estimates"
    - Respond with a list of pointers to papers in ADS
  - "Get me galaxy mass estimates"
    - Don't provide the list of papers, provide pointers to the electronic versions of the tables in those papers
  - "Get me mass estimates for M81"
    - Just provide the relevant number(s)

How to Drill Down?

- The obvious way is to implement this by introducing a hierarchical structure in the tokens tacked on to the end of the PIDs
  - However, watch out for interference with versioning
- A table (figure, …) can easily be identified as a component of a paper, but a single cell in the table???
- Need vastly more metadata about the content of compound data objects
  - How do we know what users may want to retrieve?
- Would it be sufficient to store the quantities contained and the objects covered, then parameterize?

Forward

- At this time we are trying to solve the issue of how to cite simple datasets and data objects
- Next we will need to put effort into discoverability
- But I strongly believe the demand for more sophisticated data discovery, this drilling down into compound data objects, is just around the corner
- Several people at RDA seemed to be thinking in the same direction – that is a hopeful sign
- In whatever we are working on now, though, we need to keep the future perspectives in mind