VIRTUAL ASTRONOMICAL OBSERVATORY

# Indexing the Virtual Observatory

Tom Donaldson, STScI

# Motivation

## Answer user questions as quickly as possible

- What data is available in a region?
  - Slow, but workable for small regions
  - Degrades towards impossible as region get larger
- What regions have data that satisfies some condition?
  - Not possible with pre-TAP protocols.
  - Possible with TAP or ObsTAP for a single data service.
  - Potentially very slow and challenging for cross-service criteria (e.g., What regions have both HST and Chandra observations?)

# Registry Features

- Service metadata
  - E.g., who has IR data?
  - Provisions for coverage footprints
- Helpful because metadata is centralized
  - Queries can apply constraints across data providers.
- Problematic because:
  - Metadata quality is inconsistent
    - Complexity and lack of guidance are factors, but not easy to address.
  - Many collections too heterogeneous to have specific metadata

# Spatial Indexing Can Help

- Centralized storage of MOCs is good example
  - Allows Yes/No coverage answers quickly, without needing to retrieve the data.
- Other coverage maps (HTM, etc.) possible.
- Density maps also helpful.
  - HEALPix or other data structures could support these.
- Registry could:
  - Store or ingest coverage maps in a variety of formats
  - Provide a centralized coverage service:
    - Who has data at…?
      - CDS and HEASARC provide such a service for resources they serve.
    - What regions have image data in all of IR, Visible and X-Ray?
    - Provide combined density maps.

# How Do Indexes Get Created?

- Registrant gives us index
  - e.g., they compute and contribute MOC
  - But how many will actually do this?
- We give/help with software that computes index
  - e.g., MOC, but could be other things like HTM
  - More registrants will participate, but probably not enough
- We harvest the data
  - Indexes will be much more consistently populated, but
  - Original VO protocols only support cones which do not make for efficient mining.
  - Even with a series of non-overlapping 1 square degree region queries, whole sky is > 40,000 queries.
  - How best to keep our indexes up-to-date?

# Suggestions – Near Term

- Store coverage maps (at least MOCs) in the registry
  - Decide how best to populate these maps.
- Index column metadata in the registry
  - Automatically harvest this using existing protocols.
- Develop simple http GET/POST queries that utilize those indexes and existing registry metadata:
  - Who has data at <Some Region>?
  - What regions have data that <Satisfy Some Constraints>? (e.g., have > 2 catalogs with redshift info)
- Ensure that protocols are sufficient for harvesting data
  - Box (i.e., non-overlapping) region support
  - In TAP, how do we know what tables to harvest?
  - Support counts-only queries.
    - Especially helpful in lieu of coverage maps.

# Suggestions – Longer-Term

- For queries with large responses, offer options for summary responses (counts, histograms, ???)
- Create indexes of more metadata, e.g., ObsTAP columns via harvesting.
  - Can compute histograms and range information on the fly
- Explore indexing of some data column values (beyond positions)
  - Histograms/ranges of flux values.
  - Time ranges
  - Map/reduce storage to enable more mining techniques
- Serve up summary products
  - Density maps, histograms, ???