A New Approach to Tagging Data in the Astronomical Literature

Anastasia Alexov and John Good

NASA/IPAC Infrared Archive (IRSA)
California Institute of Technology (Caltech), Pasadena, CA 91125

http://irsa.ipac.caltech.edu/applications/DataTag

ABSTRACT

Data Tags are strings used in journals to indicate the origin of the archival data and to enable the reader to recover the data. The NASA/IPAC Infrared Science Archive (IRSA) has recently introduced a new approach to the production of data tags and recovery of data from them. Many of the data access services at the IRSA return filtered data sets (such as subsets of source catalogs) and dynamically created products (such as image cutouts); these dynamically created products are not saved permanently at the archive. Rather than tag the data sets from which the query result sets are drawn, the archive tags the query that generates the results. A single tag can, then, encode a complex dynamic data set and simplifies the embedding of tags in manuscripts and journals. By logging user queries and all the parameters for those query as 'Data Tags', IRSA can re-create the query and rerun the IRSA service using the same search parameters used when the Data Tag was created. At the same time, the logs give a simple count of the actual numbers of queries made to the archive, a powerful metric of archive usage.

Currently, IRSA creates tags for queries to more than 20 data sets, including The Infrared Astronomical Satellite (IRAS), Cosmic Evolution Survey (COSMOS) and Spitzer Space Telescope Legacy Data Sets. These tags are returned by the spatial query engine, Atlas (http://irsa.ipac.caltech.edu/applications/Atlas/). IRSA plans to create tags for queries to the rest of its services in Winter 2007.

The archive provides a simple web interface (http://irsa.ipac.caltech.edu/applications/DataTag/) which recovers a data set that corresponds to the input data tag. Archived data sets may evolve in time due to improved calibrations or augmentations to the data set. IRSA's query based approach guarantees that users always receive the best available data.

Overview

- The NASA Archives and Data Centers are participating in a project to tag their data sets in the astronomical literature. IRSA began serving data tags in summer 2006
- IRSA stores a query as a "tag"; this tag can be published in a journal and used to re-run the same query to generate data results
- IRSA currently creates tags for over 20 different data sets: 10 Spitzer Legacy & FLS datasets, 5 IRAS datasets, 2MASS 6X Lockman Hole, LGA, MSX, Scrapbook, COSMOS, SDSS_DR3, MSC PTI & KI, ISO SWS and IRTS
- IRSA provides a web interface for users to check data tags and recover data corresponding to a published data tag

Advantages: Tagging Queries, not Data

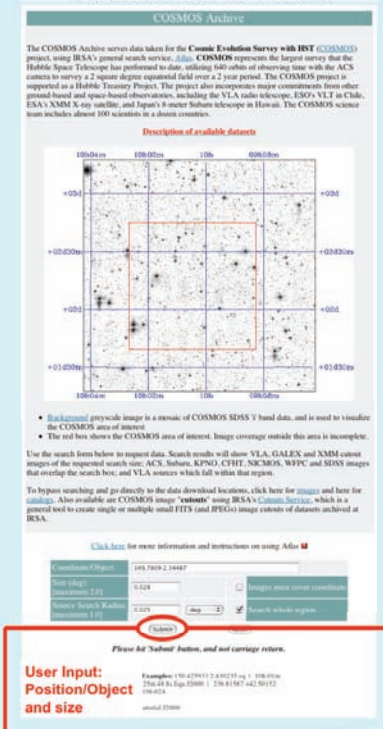
- Limits risk of stale data links.
- Tagging queries enables access to **dynamically created and filtered datasets**. Examples of such data sets are catalog subsets, image cutouts and complex data sets made up from images, spectra and catalogs.
- Publishing many individual data tags for files which overlap a region of interest is made simple via a query tag, since only **one tag can encompass many files**.
- Users will always receive the **best available** data.
- **Log files** provide archive **usage statistics** that are not available through the web server logs.

Operational Consequences

- Every IRSA search **query is logged** in a file, which **must be archived permanently**.
- If search programs at IRSA change (by name and/or major functionality) the **backwards compatibility** of the Data Tag infrastructure must be maintained.
- Query tags can include **large numbers of files**; therefore, user must **identify the pertinent files** referenced in the journal article.
- A Data Tag may generate a **different result than the original query** if the data set has changed (re-calibrated, fixed defect) since the Data Tag was created.

IRSA

Spatial Searches (COSMOS example)



User Input: Position/Object and size

Search Results:

Approach to Query Logging at IRSA

Data Tag String created:
ADS/IRSA.Atlas#2006/1002/141147_24776

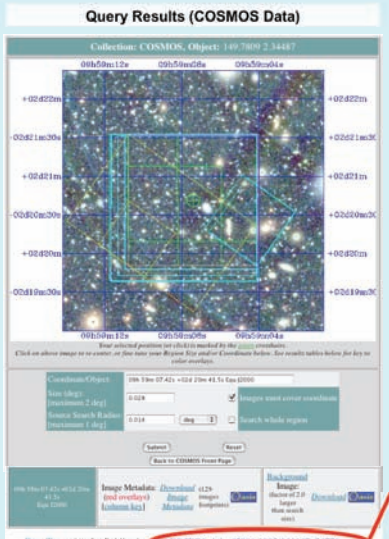
- Each month, a new log is created; it keeps track of each query run by services
- Once the "Submit" button is pressed on an IRSA data search form, the search query is logged:
  1. one continuously-updated log (per month) indicates the the date of each query
  2. a second log file (one per query), lists the http query search parameters used, and any other addition information for use of IRSA metrics
- The monthly logfile lists the date each query took place, the name of the IRSA search service which was run, and the name of the individual log file which stores the query parameters. This information is equivalent to the Data Tag.

Query parameters encoded in (2) so that HTTP query can be re-created:

(2)

```
http://irsa.ipac.caltech.edu/cgi-bin/Atlas/nph-atlas?
  mission=COSMOS&
  hdr_location=%5CCOSMOSDataPath%5C&
  collection_desc=Cosmic+Evolution+Survey+with+HST+%28COSMOS%29&
  regSize=0.028&
  radius=0.025&
  radunits=deg&
  searchregion=on&
  locstr=149.7809+2.34487
```

Query Results (COSMOS Data)



Data Tags can be published and used to rerun the same query

ADS/IRSA.Atlas#2006/1002/141147_24776

(1)

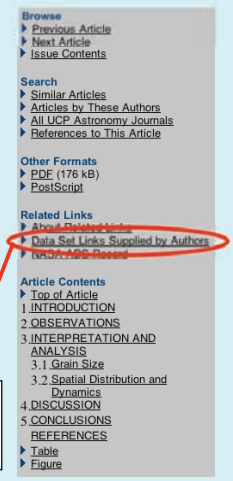
Excerpt from Log file (1) for October 2006

```
|DateTime                  |Application|          LogFileName|Status |
 2006-10-02 14:11:47 PDT    Atlas      2006/1002/141147_24776  START
 2006-10-02 14:11:59 PDT    Atlas      2006/1002/141147_24776  END
```

Publication

Sample ADS Article Data Links

Using the ADS Astronomical Literature search engine, a journal article can contain links to the data. When available, Data Tag links are presented on the ADS article left menu bar, under the "**Related Links**" heading, "**Data Set Links Supplied by Authors**" subheading. The link takes you to a list of Data Tags for that article, each of which can be used to retrieve data. In IRSAs case, each link will perform a data query; the search results will list the data used in that article.



Example list of data links from a journal article:

ADS/IRSA.Atlas#2006/1002/141147_24776

ADS/IRSA.Atlas#2006/1005154634_705

Embed Tags In LATex Macro

IRSA's simple Data Retrieval Interface

Data Tagging Plans

- IRSA will add Data Tagging capabilities to their user services by winter 2007
- ApJ COSMOS Special Issue (Spring 2007) is scheduled to use IRSA's Data Tags
- Develop tools for gathering usage metrics from the log files