Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. 376, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds.

A New Approach to Tagging Data in the Astronomical Literature

Anastasia Alexov, John C. Good

Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA, USA

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 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 unobtainable from the Apache web server logs. 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[1]. IRSA plans to create tags for queries to the rest of its services in late Spring 2007. The archive provides a simple web interface[2] 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 sets.

1. Overview

The NASA Archives and Data Centers are participating in a project to provide closer linkage between their data sets and the published literature. IRSA began serving data tags in summer 2006. Authors may include data tags returned by IRSA and other data providers in their papers through the AASTeX 5.2 dataset macro. Tags appearing in the electronic edition of the Astrophysical Journal are linked to a name resolver at ADS and enable readers to locate and access the data themselves. IRSA users are able to recover the data by using the tag as input to a simple web form. The American Astronomical Society cites the

[1]http://irsa.ipac.caltech.edu/applications/Atlas/

[2]http://irsa.ipac.caltech.edu/applications/DataTag/

following benefits of data tagging: immediate access to the data used in the paper; support for seamless transition between the electronic journals and the data providers; higher visibility of papers to researchers.

Data tags have generally been used to identify individual science products, such as spectra and images. IRSA has taken a different approach: the data tags encode the query that generates the data rather than the data sets themselves. This choice reflects that IRSA services generally return filtered data sets (such as subsets of source catalogs) and dynamically created products (such as image cutouts). IRSA supports a web interface for astronomers to check the validity of the data tags and recover data from the queries represented in the tags.

IRSA currently creates tags for over 20 different data sets: ten *Spitzer* Legacy & First Look Survey (FLS) datasets, five *Infrared Astronomical Satellite* (*IRAS*) datasets, 2MASS 6X Lockman Hole, Large Galaxy Atlas (LGA), *Mid-course Space Experiment* (*MSX*), MAST Spectral/Image Scrapbook, *COSMOS*, The Sloan Digital Sky Survey - Data Release 3 (SDSS_DR3), Michelson Science Center (MSC) Palomar Testbed Interferometer (PTI) & Keck Interferometer (KI), *Infrared Space Observatory* (*ISO*) Short Wavelength Spectrometer (SWS) and *Infrared Telescope in Space* (*IRTS*).

2. Advantages: Tagging Queries, not Data

There are a number of benefits to astronomers in tagging queries instead of data:

- Offers a simple representation of dynamically created and filtered datasets (*e.g.* a spatial query that returns data from many distinct data sets).
- Allows astronomers to refer to complex data sets with a single tag.
- Reduces the risk of stale data links in the event that data are removed from an archive or are superceded.
- Ensures that users always receive the best available data.

The tags offer a final benefit to IRSA: by keeping track of all the queries in a log file, the archive has a complete record of all queries made to it. Such usage statistics are unavailable through web server logs.

3. Operational Consequences

IRSA's data tagging approach has some operational consequences. Every query is logged in a file, which must be archived and backed-up 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 data items; therefore, archive users must identify the subset of this ensemble referenced in the journal article (unless a helpful author provides this information). A 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. The user would get the best, recent set of data, but not neccessarily the exact dataset used in the publication.

4. Query-Logging at IRSA

When data tagging is complete, every query to IRSA will be logged in order to create a Data Tag for that query. Each month, a new log is created; it keeps track of each query by service. It is a continuously-updated log (per month) and indicates the date of each each query, the name of the IRSA service, and the name of the individual log file which stores the query parameters. A sample of the start and end of one search query is shown in Table 1. IRSA uses this information to build the Data Tag string for publication. Using the example in Table 1, the Data Tag for ADS publication would be: ADS.IRSA#Atlas.2006/0630/091305_331.

Table 1. Sample Lines in IRSA's Monthly Query Log

DateTime	Application	LogFileName	Status
2006-06-30 09:13:05 PDT	Atlas	2006/0630/091305_331	START
2006-06-30 09:13:17 PDT	Atlas	2006/0630/091305_331	END

A second log file (one per query) lists the http query search parameters used, and any other additional information for use of IRSA metrics. Table 2 shows one search parameter per row.

Table 2. Example of IRSA Search Parameter Log File

Query Parameter	Value
mission	COSMOS
collection_desc	Cosmic Evolution Survey with HST (COSMOS)
regSize	0.05
radius	0.025
radunits	deg
searchregion	on
locstr	150.043358 2.400321 eq J2000
irsa_execname	Atlas/nph-atlas

Using the parameter log file, IRSA can use the data tag to reconstruct the original query and re-run it on request. For the example in Table 2, the HTTP query is:

```
nph-atlas?mission=COSMOS\&
collection\_desc=Cosmic+Evolution+Survey+with+HST+\%28COSMOS\%29\&
regSize=0.05\&radius=0.025\&radunits=deg\&searchregion=on\&
locstr=150.043358+2.400321+eq+J2000
```

5. Data Tags in Publications

Using the ADS Astronomical Literature search engine, a journal article can contain links to the data (Schwarz 2005). When available, Data Tag links are presented on the ADS article left menu bar (Accomazzi & Eichhorn 2007), under the *Related Links* header, *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. For the tags served by IRSA, each link will perform a data query; the search results will list the data used in that article.

6. IRSA's Data Retrieval Interface

IRSA's Data Retrieval Service[3] uses a publication Data Tag to return data from the query represented by the tag. Entering an IRSA Data Tag in the search form will rerun the IRSA service using the same search parameters used to create the Data Tag. This service allows users to check their Data Tags in preparation for publication and offers journal readers a web interface to access data used by authors.

7. Data Tagging Plans

IRSA will add Data Tag capabilities to all user services by late Spring 2007. The *COSMOS* Special Issue of the Astrophysical Journal (Spring 2007) is scheduled to use IRSA's Data Tags. IRSA plans to develop tools for gathering usage metrics from the log files.

Acknowledgments. We would like to thank D. Kirkpatrick, B. Berriman and A. Laity for their assistance on the conference poster. Data Tagging was initiated for the *COSMOS* project for use in their ApJ Special Issue journal. This work was partly funded by the *COSMOS* project, as part of the development of its archive by IRSA. The Infrared Science Archive (IRSA) is funded by NASA under contract to the Jet Propulsion Laboratory.

References

Accomazzi, A., & Eichhorn, G. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 467

Schwarz, G. J. 2005, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP), 375

[3]http://irsa.ipac.caltech.edu/applications/DataTag/