

Advanced Architecture for the Infrared Science Archive

G. B. Berriman, N. Chiu, J. Good, T. Handley, A. Johnson, M. Kong,
S. Monkewitz, S. W. Norton, A. Zhang

*Infrared Processing and Analysis Center, California Institute of
Technology, Pasadena, CA 91125*

Abstract. This paper describes the data mining, catalog-cross comparison, and visualization services available at the Infrared Science Archive (IRSA), NASA's archive node for infrared astronomy data. IRSA is a living archive, which maintains contemporary datasets and continuously develops services to exploit these datasets. Over the past three years, IRSA has devoted most of its resources to support the requirements of the massive 2MASS survey datasets. Given the high volumes of 2MASS data, the services and infrastructure supporting them provide insight in understanding how a future NVO may operate.

1. IRSA's Charter

The Infrared Processing and Analysis Center (IPAC) at Caltech was charged with archiving data sets produced by the Infrared Astronomical Satellite (IRAS). The success of this mission and the demand for its data products made IPAC a leading center for archival research and data distribution, and led directly to the development of the Infrared Science Archive (IRSA) as the archive node for NASA's infrared astronomy missions. IRSA now provides public access to the catalogs and images from the 2MASS and MSX missions, as well as from the IRAS mission. IRSA's requirements are derived on one hand from the specialized needs of projects, and on the other hand from the needs of users analyzing the data. IRSA also holds ancillary catalogs required to allow exploitation of the infrared datasets, including USNOA 2, NRAO VLA Sky Survey (NVSS) and Faint Images of the Radio Sky at Twenty-centimeters (FIRST). A full list of holdings as of August 2000 can be found at the IRSA web site.¹

2. Access To IRSA Services

While all IRSA services can be invoked via a program interface or via remote HTTP or Java client interfaces, users generally invoke them in server mode through a web client. Processing is performed server-side in the IRSA environment. Results are made available to the user as tables or images that can be

¹<http://irsa.ipac.caltech.edu>

downloaded. Broadly speaking, the following services can be applied to the data held by IRSA:

1. Completely general catalog queries,
2. Image queries and visualization, with customization of visualization tools applicable to individual missions,
3. Cross-comparison between catalogs, and
4. Statistical representation of large datasets.

IRSA receives on average over 220 requests for data each day, and over 99% of these requests are successfully processed. The architecture of the IRSA services is described by Good (2001).

3. IRSA As A Living Archive

A key feature of IRSA is that it is a *living* archive. That is, by providing robust, contemporary archives and by continuously developing services that have the power to exploit them, IRSA permits the development of new scientific products and opens up new avenues of research. The support provided for IRSA's largest customer, the 2MASS project, has demonstrated the power of such an archive. 2MASS is uniformly surveying the entire sky in three near-infrared bands to detect point sources brighter than about 1 mJy in each band, achieving an 80,000-fold improvement in sensitivity over the first full-sky survey of Neugebauer & Leighton. This deep survey has generated datasets that are by far the largest obtained in any astronomical survey, with roughly 12 TB of images, and an internal catalog of sources now containing over 1 billion entries in an Informix database.

Efficient mining of these huge 2MASS datasets places extraordinarily large loads on IRSA services. Research into special techniques has led to the development of optimized algorithms and software for rapid searches of large databases. As an example of the new science that can be performed with the help of these services, astronomers culled from the 2MASS catalog the very red candidate objects from which the first brown dwarfs were identified (Kirkpatrick et al. 1999, and references therein).

Future surveys will produce data volumes even larger than those generated by 2MASS, and will certainly produce ever more spectacular scientific advances. The services developed by IRSA and the infrastructure to support them therefore provide a window into how a future NVO is likely to function. The remainder of this paper is therefore given over to discussions of special features of IRSA services that permit efficient data mining, and current research at IRSA that will provide the next generation of data mining, visualization and analysis tools.

4. Special Features of IRSA Services For Data Mining Large Volume Datasets

4.1. Catalog Queries and Indexing of Database Tables

Driving the 2MASS data mining requirements is the need to provide efficient and completely general querying methods. Querying has been made efficient in two ways. First, queries are run in parallel fashion across as many processors and

I/O channels as possible. IRSA chose a Sun Microsystems E6500 server and an Informix database because they can be highly optimized for parallel querying. Second, IRSA spatially indexes catalogs using nested hierarchies of increasingly smaller bins. Search mechanisms traverse those branches of the tree to isolate database entries that meet the constraints imposed by the query. IRSA employs three spatial indexes:

1. The Hierarchical Triangular Mesh (HTM), developed at Johns Hopkins to support the Sloan Digital Sky survey, divides the sky into a nested series of equilateral triangles.
2. In magnitude/color space, three dimensional box partitioning with logarithmic steps away from the diagonal.
3. R-tree indexing of image metadata where images cover a large area; R-trees take into account the spatial extent of the elements contained within them.

4.2. Catalog Cross-Comparison and Distributed Queries

Here lies perhaps the greatest challenge to the NVO, and here also lies perhaps the greatest potential for ground-breaking new science. IRSA has developed tools that allow for efficient distributed queries and which handle the complex DBMS functionality involved in cross-comparison of catalogs. The heart of the methodology is the use of three-way joins between tables, and sets of candidate source associations (known as relationship objects). Ma et al. (2000) describe the power of this method in more detail, and future research in this area.

4.3. Image Metadata

IRSA separates images from their associated metadata. The metadata reside in a database catalog, with one record per image. Indexed searches can be made on any parameter, as well as spatially indexed searches based on position. IRSA can therefore efficiently locate images in catalogs it does not hold itself.

4.4. Statistical Representation of Large Datasets

Generally speaking, statistical representations of catalogs or query results are a powerful way of initially studying a large volume of data, allowing a user to quickly refine a search to locate objects of interest, such as those with extreme colors. IRSA provides a service to derive a histogram of a pre-binned representation of the IRSA database.

5. Current Research at IRSA and Development of Next Generation Services

One of IRSA's next generation services is described in detail elsewhere in this volume. Many astronomers have expressed the desire to perform on-the-fly coordinate transformations with the minimum of function calls. Zhang et al. (2001) describe how an Informix datablade solves this problem. A datablade is simply a function or library embedded into the Informix Database Management System. Users simply embed the input and target coordinates into their SQL queries; no further function calls are necessary.

In the next two years, IRSA anticipates that it will ingest the catalogs, images, and spectra from the Infrared Telescope in Space (IRTS) and, in collaboration with the NASA Goddard Space Flight Center, the data sets from the Cosmic Background Explorer (COBE). IRSA will provide integrated access to the Infrared Space Observatory (ISO) data archives, held in Vilspa, Spain. Further in the future, IRSA will archive catalogs from the SIRTf and SOFIA missions. The previous section described research into catalog-cross comparisons. The full value of these data will be realized when they can be combined with distributed multi-wavelength data. Much of IRSA's research and development is therefore aimed at providing a portable, Java-based architecture that will support efficient access to and interaction with multiple, distributed data sets. This architecture, called the On-Line Archive Science Information System (OASIS), is described in detail by Good et al. (2001), and is modeled after the layering methodology employed by Geographical Information Systems (GIS). As part of the OASIS development effort, IRSA is cooperating with STScI, CDS, and ADC to establish an XML output format for catalog search results, and is pursuing similar standards for the transfer of image metadata.

Underpinning the OASIS front-end will be a persistent archive request and management mechanism. IRSA expects to deploy this infrastructure in Spring 2001. It is designed to wrap around existing services, while supporting existing HTML form and CGI technologies. The request manager is an Enterprise Java bean system that uses the WebLogic Applications Server to accept multiple requests from users, control information, maintain state information, and communicate results to users.

Acknowledgments. We thank the NASA Mission Operations and Data Analysis and Science Applications of Information Technology programs and the Digital Sky project for financial support. We thank Drs. G. Helou, W. P. Lee, C. J. Lonsdale, and J. Ma for their technical support.

References

- Good, J. C., et al. 2001, this volume, 52
Kirkpatrick, J. D., et al. 1999, ApJ, 522, L65
Ma, J., et al. 2001, in ASP Conf. Ser., Vol. 225 Virtual Observatories of the Future, ed. R. J. Brunner, S. G. Djorgovski, & A. Szalay (San Francisco: ASP), in press