

Report of the April 2002 Meeting of the Science Archive Working Group

The inaugural meeting of the Science Archive Working Group (SAWG) met at NASA Headquarters on April 29-30, 2002 and was attended by the members Julian Borrill, Roger Brissenden, Joel Bregman (Chair), Damien Christian, Eric Feigelson, Menas Kafatos, William Oegerle (Co-Chair), Sally Oey, Tom McGlynn, and Richard White, along with Paul Hertz and Joe Brederkamp from NASA HQ (absent were Jonathan Borden and Carol Lonsdale). There were a variety of presentations by some, but not all of the archive centers and related programs that report or receive support under the SEUS and Origins themes. We report on the discussions surrounding the presentations where there were substantive issues under consideration.

National Virtual Observatory

There were three presentations on the National Virtual Observatory (NVO), by Alex Szalay, Bob Hanisch, and George Djorgovski, which covered the general technical capabilities of such a system as well as its scientific benefits; the recently funded NSF/ITR activities were presented in detail. Also, four of the existing NASA data centers (HEASARC, Nick White; MAST, Marc Postman; IRSA, George Helou; ADC, Jim Green) reported upon their organizations and outlined their activities that relate to the NVO. The NVO concept is becoming defined and the SAWG was impressed with both the thoughtfulness of the program and with the degree of cooperation between the NVO lead personnel, the various data centers and the people in both universities and in industry. The NVO group is aware of the international efforts and hope to coordinate activities when feasible. The NVO is the natural evolution of the activities that have been occurring at the archival data centers and it is a feasible project that should have benefit well beyond the borders of the SEUS and OS themes in NASA. It will greatly enhance the usefulness and value of astronomical archival data, which we regard as a national treasure.

The NVO is built around the data archives, whose vitality is essential to this endeavor. The NASA astrophysics data centers have a strong record and have developed a coordinating council (the ADEC) to help define cooperative ventures and to avoid duplication of effort. It, along with individual data centers, contributes to NVO activities. In contrast, there are no optical and radio ground based archives at this time, a regrettable shortcoming that we hope will be rectified within the next few years. Archives from ground-based data will be an essential component of the NVO.

While we support the NVO concept, there remains a great deal to be defined. In particular, a formal proposal and work plan to NASA needs to be developed before NASA and the SAWG can comment upon the detailed issues of schedule and budget. In general terms, aside from suggested new research funds, the incremental cost for NASA to develop its part of the NVO may be relatively modest and should be significantly less than would be required from the NSF, due to the cost of developing ground-based archives. Regarding the research component, should the NVO greatly enrich archival data analysis, research funding should become commensurate with the change in scientific activity.

The NSF ITR program is helping to develop pieces that will contribute to the NVO and we view this as a testbed that will lead to a more complete system in the future. The NVO team

has shown great care in the management of the project, which will be essential to its success. We look forward to the initial demonstration projects that will be presented in January 2003 at the AAS meeting.

National Herschel Science Center

George Helou presented the IPAC plans for supporting US-based investigators of the Herschel mission. The placement of the National Herschel Science Center at IPAC makes optimal use of existing expertise, places the science support center nearby to US science instrument builders, and will ensure that software systems and archives are based on existing successful systems. The SAWG endorses the decision to locate the National Herschel Science Center at IPAC.

The labor levels for developing the science center were provided and the SAWG had some concern. The labor level was expected to be 38 FTEs during the prime phase of Herschel and this may be compared to a level of 13 FTEs for support of XMM (and previously, ROSAT) at the HEASARC during their prime phases. There was insufficient information available to properly review the required level so we recommend that a clear definition of the requirements and software deliverables for supporting the US community be developed and used to justify the staffing levels proposed.

We also note the reliance on the European component of the program and suggest that once the requirements have been baselined, a set of dependencies be identified and an appropriate agreement with the Europeans be developed.

CMB Data Center

Gary Hinshaw presented a proposal to create a Cosmic Microwave Background (CMB) data center by consolidating the Active Archive for the MAP mission with the COBE and SWAS archives within the Laboratory for Astronomy and Solar Physics (Code 680; as there were few details, this was more of a “concept” than a “proposal”). The proposal creates a thematic data center with expert support for the community and consolidates resources to maximize the science expertise and output product quality. We note that the GSFC Space Science Directorate (Code 600) Visiting Committee recommended in April 2002 that the MAP Active Archive be consolidated in this way. Also, the pre-launch agreement is to house the MAP data in the ADC Active Archive. We note that NSSDC Astrophysics encompasses three components: the Permanent Archive; the Active Archive, and the ADC. The proposal for a CMB archive applies to the Active Archive and does not affect the contents of the Permanent Archive. It is possible that resources from the ADC may be redirected to partly fund a CMB archive and this must be balanced against the value of the ADC activities that would be reduced or eliminated.

The resources required to operate the CMB data center were estimated at ~4 FTE, which is approximately double the existing level of personnel. At the present level, the data from MAP are ingested and available to the community, but with little or no support in the areas of dealing with questions from users or providing expert commentary. The SAWG concurs that these data sets should receive proper curation and that there will be significant value added to both the science and community. During the period in which MAP data is the main archive data set, the nature of this archive might be quite different from other data archives. In other imaging archives, there are typically many objects per field, so a variety of users download the same data

for many different purposes. From this model, we have an understanding of the necessary staffing levels. The MAP data set is entirely different, so the usual staffing model may not apply. The MAP data is likely to be used by a few expert users, such as the teams involved in other CMB experiments or those dealing with foreground and background issues in this waveband. As the staffing levels were not particularly well justified, this issue may need to be revisited on a yearly basis.

The issue of location of a CMB archive was not discussed, as it was assumed to reside at GSFC and it is sensible to have the expertise and the data at the same location. However, the IRSA will be the home for Planck, the next large CMB mission after MAP (and other far-IR missions will reside at the IRSA). If a CMB archive at GSFC has no additional mission after MAP for a substantial number of years, it may not be a viable long-term archive. Prior to the formation or enhancement of a CMB archive, NASA should consider how the long-term needs of the community would best be served.

The concept was put forward that the CMB data center might incorporate data sets from sub-orbital missions. This is likely to be a useful addition but it places new demands on those missions, which would require additional funding. The addition of funding to permit archiving of sub-orbital mission data might reduce the total number of such missions, so the CMB community should be queried as to whether this is a sensible trade-off.

ADC Activities

One of the most difficult issues that the SAWG faced was in assessing the vitality and importance of the activities being carried out by the ADC. They conduct such activities as: accumulation of tables; ability to search for table (ADC Data Viewer); plotting of different columns of data tables (CatsEye), a graphical interface to astronomical databases (IMPRESS), development of XML tools, a method to query catalog holdings (AMASE), some EPO activities; and they are a mirror for Vizier service, which come from the CDS in Strasbourg. This was the lowest rated of the archive sites in the last Senior Review, although the absolute rating still indicates that the Senior Review believed it to be of value. The level required to continue all these activities is about \$500k/yr but a decision was made within Code 600 (the Space Science Operations Data Office) to reduce the funding level to the ADC by about half. It has been difficult to understand the tasks of ADC staff members prior to and after staff reductions, so we cannot comment effectively on the proper number of FTEs to conduct various tasks. The following concentrates on the functions of the ADC rather than on staffing levels.

The primary activity of the ADC centers on their tables and since the CDS also ingests tables, the issue of duplication of effort is central to the discussion. This issue was not adequately developed in the presentation by Jim Green, although he responded to further inquiries by email, which were very helpful. According to his response, on an annual basis, the ADC ingests about 1/3 of all astronomical catalogs. It has been difficult to confirm this although we requested and were sent a list of the tables ingested in 2001. During 2001, the ADC ingested 33 tables (or table sets from papers), while the CDS ingested about 50 per month, nearly 20 times more (only 1 table in the month of April 2002 is acknowledged to come from NASA/GSFC). Even for a reduced staff size, this seems like a low rate of table ingestion and inconsistent with the presentation and email of Jim Green.

The ADC has developed an SGML-to-XML ingest pipeline and they provide quality control to this process, sometimes finding errors in table entries. Their Data Viewer is very

capable and in some ways superior to VisieR, its competitor. Usage of this site has been less than at the other archive sites, although usage has increased in the past few months and by improving the connections to these tables (e.g., from the IRSA), this may become a heavily used site. However, some features offered by the ADC are not broadly used by the community, and given the alternatives, are not likely to become widely used by the community. One example is their graphical interface to databases (IMPReSS), since there are other viewers that are quite capable and are more widely used (e.g., SkyView). Most of these tools probably do not have the correct architecture to be incorporated into the NVO activities, although this was not discussed either. There is a general issue of too many graphical viewers, each of which have been developed at the various archive sites, and this is something that the ADEC should deal with.

Regarding their XML development work, the ADC were pioneers in this area. With relationship to the NVO project, there is a contract from the NSF-funded program to ADC/Raytheon for ½ FTE in 2002, 1 FTE in 2003, and ½ FTE in 2004, according to Bob Hanisch. This is primarily for support of Ed Shaya, who is valuable to the NVO program. However, the NVO project is not dependent upon the ADC for XML development. If funding is available, support for XML development is certainly useful, but this should not be traded against funding that would go forward to ingesting tabular data.

It is probably important to have one US site that is involved in the ingestion and query of tabular data. This would naturally connect to the efforts by publishers of journals in North America (such as ApJ), who have been involved in connecting many parts of articles with databases. Currently, the ApJ provides ASCII versions of all tables since the advent of electronic on-line articles (a really inexpensive way of augmenting the archive is to make a simple ASCII file of name of tables and the volume that they come from, since they are easy to download). Also, if we rely entirely on foreign institutions, we will have little input or control of the process and the needs of the NASA community may not be adequately served, with little recourse. Having more than one institution involved in tabular data activities is likely to lead to competition, which will help to improve efficiency and quality of offerings (competition between some of the other data centers has been quite important). Also, links to these tables from the other data centers need to be improved. One might consider whether such table activities should continue within the ADC or moved elsewhere.

Other Items of Note

We are relieved to see that a significant amount of data has finally arrived in the XMM archives, although the situation is still not ideal. The interface to the archive does not work on a variety of machines and there is a limit as to the number of data sets that a site can download per week. Also, it would be helpful if there were a mirror site in the US rather than being required to download from the site in Spain, where the bandwidth is much more limited.

The reports by the other presenters were in line with expectations and the committee did not have particular concerns.