Data Centre Operations in the Virtual Observatory Age

David Schade

(1) Canadian Astronomy Data Centre, Herzberg Institute of Astrophysics, National Research Council Canada

5071 West Saainch Road, Victoria, B.C., Canada

EMail: David.Schade@nrc.ca

ABSTRACT

Data centres are entirely engaged in getting data out of the door to scientists who want to do research with it immediately. That is our business. Data curation over the long term is of lesser concern to our funding agencies. However, we are highly motivated toward adding value to data by creating more useful products and services because it moves data out the door to users more quickly. More usage this year means continued funding next year. Fortunately, when we add value to data to get it out the door today, we have added value permanently to those data. Users a decade from now will benefit just as much as users do today.

At the Canadian Astronomy Data Centre we have made strong commitments to a few distinct key areas. These include Advanced Data Products and Cross-Archive data query. We continue to work intimately with data providers at several major observatories (Gemini, CFHT, JCMT, HST). We have produced advanced data products for many years and usage patterns support our view that his is essential work. Cross-archive data query is fundamental to good data access across our own collection and across the world data collection. Our work with observatories has evolved rapidly in the past few years and profound lessons follow from this work. The vision that was popular some years ago was that VO could be implemented as a “thin-layer” on top of existing data collections. It was soon realized that substantial re-engineering of metadata needs to be done on old data collections before they can be integrated into the “new” VO styles of access. So the requirements of VO push back past the data access layer into the metadata of the collection itself. It is now evident that the requirements of VO push back past the collections housed at data centres all the way to the observatories where metadata originates. It could be argued that the real frontier of VO is at the telescope and that the most important work that we do to enable VO is our collaboration with observatories. This makes VO a still more daunting prospect. But nobody said it would be easy. Or maybe they did. If so, they were wrong.

Keywords: up to 8 keywords, comma separated, may be added here

INTRODUCTION

The primary role of the Canadian Astronomy Data Centre and most modern data centres is the organization of data and the distribution to scientists in order to support science in the short term. They have never had a serious role in long-term preservation and most have not been funded to preserve data. The digital age is young and it will take time to develop approaches and policies that make sense for managing science data.

Data centres need to justify the money spent on them. They do this by providing valuable services. They provide “added-value” services to support research and they do this in a variety of ways. A generation of astronomers has grown up expecting services like these and the high rate of usage of data centre services proves their worth. The success of data centres is built upon an intimate collaboration of science and technical staff.

Long-term preservation of data is not a core role of data centres and many are not interested and not well equipped to address that problem. New thinking is required on this subject. It is not sufficient to assume that data centres will grow into this role naturally. At CADC we have so far escaped confronting this challenge but we will be forced to deal with it in the near future.

The Virtual Observatory vision sees an international community of data collections and services operating together in a seamless fashion. This vision is one of the drivers toward integration of data services in astronomy. This vision seems to be exactly what is called for based on the international, multi-facility, multi-wavelength nature of science teams. There are considerable challenges involved with making the VO vision a reality.

ARCHIVING AND ADDING VALUE TO DATA

Ther term “data archive” is a burden. Archiving is a word that does not adequately describe what most astronomy data centres do. Archiving has an air to it of “old”, “dusty”, “rarely accessed”, and “accessed under carefully-controlled conditions”. It is close to the opposite of what data centres actually strive for. I prefer the term “data centre” to describe our facilities and “data management” to describe our activities.

The Canadian Astronomy Data Centre (CADC) was created in 1986 with the mission of getting Hubble Space Telescope data into the hands of Canadian astronomers. As an astronomer I published papers using archival Hubble Space Telescope data and Canada-France-Hawaii Telescope data. My motivation for going to CADC was to be closer to collections and processing capability so that I could do science better and more quickly. I did not choose to work at the CADC so that I could “archive” data for posterity.

Why and how do data centres add value to data?

Data centres need to provide a valuable service related to the data that they handle. There are many ways to add value to data and data collections. Simply getting all of the data from an observatory into one place and assuring astronomers that it is available from that place is a process that constitutes “adding value”.

Organizing data in ways that are meaningful to researchers is another basic way of adding value. Historically, this has meant extracting metadata from “FITS” headers and building query-able database tables from that metadata. (FITS is a standard for image transport that was developed over 25 years ago and includes metadata in the file header that describes the data it contains.) Header metadata has often been the only metadata that was available. The next step in adding value would be a careful selection of which metadata to use and an evaluation of the metadata quality, possible including correction of header metadata where it needs to be improved. Through this process it has become clear that metadata management is a central role of data centres. It is an increasingly expensive component of data centre activity.

Another important step in delivering a service of value to astronomers is the creation of good interfaces that support database queries and data access. The overall design of a data management system consisting of file systems for storage, databases to house catalogues, and servers to deliver data adds considerable value.

Data processing is an activity that is clearly a “value-added” activity. The purpose is to take the contents of a data collection (for example, raw images of the sky) and to apply processing algorithms to produce an “Advanced Data Product”, for example a stacked image which is an average or median image of higher quality than the individual images. This is a major step in the activities of a data centre. Not all data centres need to undertake this activity. It requires deep science domain knowledge and this elevates the role of science staff in data centre operations.

Metadata management requires science expertise. The design of interfaces also requires science expertise and clearly data processing requires very significant guidance from scientists. These considerations suggest that the organizational structure of data management groups need to incorporate scientists as well as a technical staff with a range of skills.

Successful data centres have an intimate, effective working relationship between science and technical staff.

Long-term care and preservation of data

My understanding of observatory policy before the digital age is that the observatory generally maintained ownership of data (usually photographic plates) but that the observer normally took their plates away with them. In my experience the plates often never returned to the observatory and the only "archives" were those represented by the personal collections of the observers at their home institutions. There are counter-examples, such as the Harvard College Observatory plate collection of over a half-million plates taken over a period of about 100 years. And there are important survey collections like the Palomar Observatory Sky Survey which has been reproduced and distributed throughout the world. But Principal Investigators normally had considerable control over the disposition of the data they produced.

The digital age started with a data policy and practice unchanged from the pre-digital age. Some observatories made a copy of digital data for safety reasons. But Principal Investigators (PI's) copied their data at the telescope and took them away with them. Thereafter, they maintained control over the distribution of those data.

Why should we put resources into preserving data into the future? The principal reason is that data contain science value that does not vanish either after a single research use or after a specific period of time has elapsed. In particular, the time sampling of historical data cannot be replaced by any other means. It must be preserved. There is an argument that "the sky is my archive" and it can be observed again, usually with greater sensitivity but this argument clearly fails to undermine the time-domain value argument for data obtained in the past.

The science value of data justifies it being preserved rather than discarded.

At this time, there is neither a data policy in Canada (or elsewhere that I am aware of) nor is there a means of funding long-term data preservation that prevents data from disappearing. But there are data centres in many countries that do astronomy. These data centres are doing work that greatly assists in preserving the value of astronomical data. But they do not have a mandate for long-term preservation.

We should not be looking to the data centres to carry out the mission of long-term data preservation. They were formed and staffed for a different mission. That mission is adding value and getting data out the door to produce more science in the short term. And our role is evolving more into the mainstream data management areas and even further away from an "archiving" role. The Canadian Astronomy Data Centre has never received significant funding to support preservation. We value data and we believe that it should be preserved but its important to recognize that preservation is neither an area of interest nor is it an area where we have real expertise.

Data centres do not have the structure to support long-term preservation of astronomical data. Nevertheless, the work that they do provides a foundation for preservation. The collection and engineering of metadata and the building of database catalogues to support querying are necessary, in addition to basic file storage, to preserve the value of data. If this work is not done, the value of data is lost irretrievably. Astronomy data centres have invented systems that need to be in place before long-term preservation of data (and more importantly, the value of data) can be implemented.

Could existing data centres take on the role of long-term data preservation? How would this activity be funded? They could take on the role if they supplemented their current staffing model (science, software development, operations) with staff whose role was to implement changes that would enable long-term preservation. The funding for this staff and their activities should not compete against other activities of the data centre. These could be staff from the digital archiving world. It's a model that deserves to be thought about.

At CADC (and elsewhere I suspect) the issue of long-term preservation of data has never been an important challenge. Older and lesser-used data collections have been preserved and migrated to new media at negligible cost. This is due directly to the rapid growth in data volumes produced by each successive generation of astronomical instruments. It cost us nothing to preserve, therefore we have successfully preserved. This situation is changing.

The James Clerk Maxwell Telescope produces very large data volumes and these are processed and housed at CADC. In a few years the telescope will close, the primary science from the new instruments will be largely complete, and the immediate motivation to continue funding the maintenance of the data collection will vanish. The new and stark reality of this situation is that we project that more than half of our data collection will be JCMT. Therefore, new storage costs will largely be spent on data from a closed facility whose mission is perceived to be complete. A similar situation will eventually occur for the Canada-France-Hawaii Telescope data collection. And it is unlikely that we will house new massive data collection on our site. So we will face the very real question of how much we value legacy data collections. Are we willing to spend real money to preserve them?

CADC EXPERIENCE

CADC has grown by a factor of 3 or 4 in the past 10 years. The number of astronomers downloading data from CADC has grown by a factor of about 25 (from 100 per year to over 2500 per year) in that time. We consider ourselves to be successful based on the roles that we play and the number of scientists that use our services. What do we do differently now that contributes to our success? What factors have contributed to our success?

There are two variables that are useful to describe data centre activities. The first is scale. How large are the data collections, databases, processing tasks. The second is scope. What is the scope of data centre activities? Are they limited to receiving data from facilities and storing and cataloguing those data and then allowing query and download. Do we work with our sources of data over the years to improve data and metadata quality? Do we produce Advanced Data Products? The scale of CADC operations has grown by a large factor in the past decade. At the same time, the scope of the work that we do has gradually expanded.

During the first 10 years of its existence CADC supported only “archival” access to data after the proprietary period had expired and the data were available to the entire research community. A critical point in the history of our data centre occurred when the report of the Long Range Plan for Astronomy in Canada recommended that we be given additional resources to develop the capacity and the new features necessary to handle the wide-field camera data from Canada-France-Hawaii telescope. This new camera would cause the scale of our operations to balloon by a factor of 100. The scope of what we do also changed. We became the exclusive channel for data distribution from the telescope to all Legacy Survey science users worldwide. There was no distribution directly from the telescope. We thus assumed the cost of distribution and relieved the observatory of that burden for survey users.

A second turning point in the history of CADC was when we began undertaking distribution of all data for the Gemini Observatories in Chile and Hawaii. The working partnership with Gemini has been close and productive and the inclusion of PI distribution in the mandate of the Gemini Science Archive at CADC was an important step. We are undertaking a similar collaborative arrangement with the staff of the James Clerk Maxwell Telescope to produce an advanced science archive at CADC serving all JCMT users.

In summary, both the scale and the scope of our operations have expanded significantly in the past decade. The CADC has moved from the “archival” fringes of the data flow between observatories and researchers and into the mainstream of astronomy data flow. Thus the value that we deliver to the science community has increased very significantly.

VIRTUAL OBSERVATORY

The Virtual Observatory (VO) is an important feature of the data management landscape in astronomy. What is the Virtual Observatory? It is a vision for the future of astronomy data management that guides the way that we do things. It is also a set of national VO projects that make up the International Virtual Observatory Alliance. It will someday soon be an identifiable entity that is used extensively by astronomers to do science.

The key concept that drives VO is integration. Seamless integration of distributed data collections and services of all types is the problem that VO needs to conquer. Inter-operability is a term often used in VO circles to describe what we need to do on the technical side to achieve what is perceived by the science user as integration.

An example that illustrates the VO vision is a case where a scientist has a science project in mind. She/he needs to explore and select data, assess its suitability and refine the selection. The selected set of data is then subjected to processing of some sort. The results are evaluated. Analysis of the processing output yields a scientific result. This whole process starting at data selection is likely to be iterated again in order to refine the result. In a VO world, many or all of these tasks would be executed on VO systems with no requirement for local data collections, storage, database, or processing resources. These resources and capabilities would be integrated by VO machinery.

The sociology of astronomy has changed during the past 20 years. Astronomy is very often done by large teams that span continents. Survey astronomy has grown in importance. At the same time the use of multi-wavelength datasets have become more important, even dominant, in many areas of astronomy. It is now much more common for scientists of use existing (archival) data collections to complement data produced from a specific observatory proposal. All of these changes are inter-connected and reinforce one another. All of these changes in the practice of astronomy are what drives the Virtual Observatory vision. The VO vision is a response to science practice.

The International Virtual Observatory Alliance has been working since 2002. It has accomplished a number of things and is producing a growing number of standards for inter-operability. It is fair to say, in my opinion, that the impact of this work on research scientists has been small, up to this point. The pace of development could be better. The management of distributed projects is notoriously difficult. In the cases of the IVOA no money changes hands between international proejcts, making the management problem more difficult.

The IVOA or other groups attempting collaboration between distributed data management groups should be careful to define the goals of their work in detail. High-level goals should be transformed into science and technical requirements and then a consensus should be developed to define the highest priorities. It is very easy to have ambitions that greatly exceed the resources available to make them a reality.

A problem that appears to be fundamental is that the VO has been seen by some as solely concerned with developing inter-operability between existing services. In fact, many existing astronomy data services do not have the capabilities that VO requires to deliver VO-like functionality. Inter-operability will not solve the problem. The services themselves need to be re-developed to a higher standard. There needs to be a wider appreciation of this fact and a plan needs to be developed to address this issue through substantial investments in data centres to build up the infrastructure and to do the data engineering that is required to support VO functionalities.

The Virtual Observatory vision cannot become a reality without strong investment in improving the infrastructure and capabilities of data centres and other data providers.

FUTURE OF ASTRONOMY DATA MANAGEMENT

The future of astronomy data management is integration of data collections and services. The VO is a part of this vision. We are not limited by technology. We have never been able to fully exploit existing technology. We are not limited by vision. We have a clear high-level vision for how we need to integrate astronomy resources to deliver value to the scientist. We are limited by our ability to design and implement real services that represent our vision. We are also limited by our ability to collaboratively define common goals internationally and to pursue those goals. But we are learning and we are making progress.

Beyond astronomy, the future of data management is in integration across the boundaries of scientific areas of study. One might first integrate data and services in astronomy and physics, then move to integrate these with biology and chemistry. There will exist Science Data Centres that handle multi-disciplinary datasets. Ultimately, sciences will integrate with social sciences, health, and other areas where data are collected. But I have to admit that, although this seems inevitable to me (the alternative is perpetually working in each field of data management in isolation), I have no idea how this will be achieved.

LESSONS LEARNED

Data management as we envision it now is a young field. The transition to digital data opened the door and changed everything. We are still reacting to that change. So the lessons that we have learned so far are fragmentary rather than comprehensive. They give us hints on how to move forward but that's all.

My most important lessons learned from my 13 years in data management would be as follows.

Science data management is in an early developmental phase and a tight partnership between domain experts (astronomers) and software developers and other technical staff is needed to define the way forward.

Data management work needs to be driven by science practice. For example, our vision that integration is a dominant theme for data management is driven by the fact that scientists integrate multi-facility, multi-wavelength data to do their research.

Data centres need to invest in working with data producers to help them understand the downstream impact of their decisions at the telescope on data and metadata handling. This will lead to improvements in the data produced at source.

Data centres will not be a driving force in implementation of long-term preservation of data. We need to think in terms of modified or completely new models to achieve preservation. But data centres are doing much of the work that will enable effective long-term preservation so the need to be part of the process.

The success of the Virtual Observatory movement depends on substantial work being done inside data centres and by other data producers to improve their collections and services before it will be possible to make these services "inter-operate" in a way that is valuable to astronomers.

Time scales for big changes are much longer (by a factor of four) than I thought.

It is easy to develop visions that greatly over-reach our ability to make them reality. It is sometimes productive to dream of the impossible. It is never very useful to try to implement the impossible.