CADC White Paper for LRP Committee

15 February 2010

Submitted by:
David Schade – Group Leader

CADC

The purpose of the Canadian Astronomy Data Centre (CADC) is to develop and operate data management services for the benefit of research users. Its activities are driven by the needs of the Canadian research community.

The idea behind the CADC is that a close marriage of scientific and technical expertise produces advances in applications of technology that make the doing of science better and more convenient. The group began with astronomers who were highly-skilled in computer science and has evolved into a group consisting of astronomers and computer scientists. The group is structured as 3 sections: Operations, Software Development, and Science. The head is the Group Leader and each section has a manager. These groups work together on defined projects, for example the Gemini Science Archive. The CADC has 20 staff members.

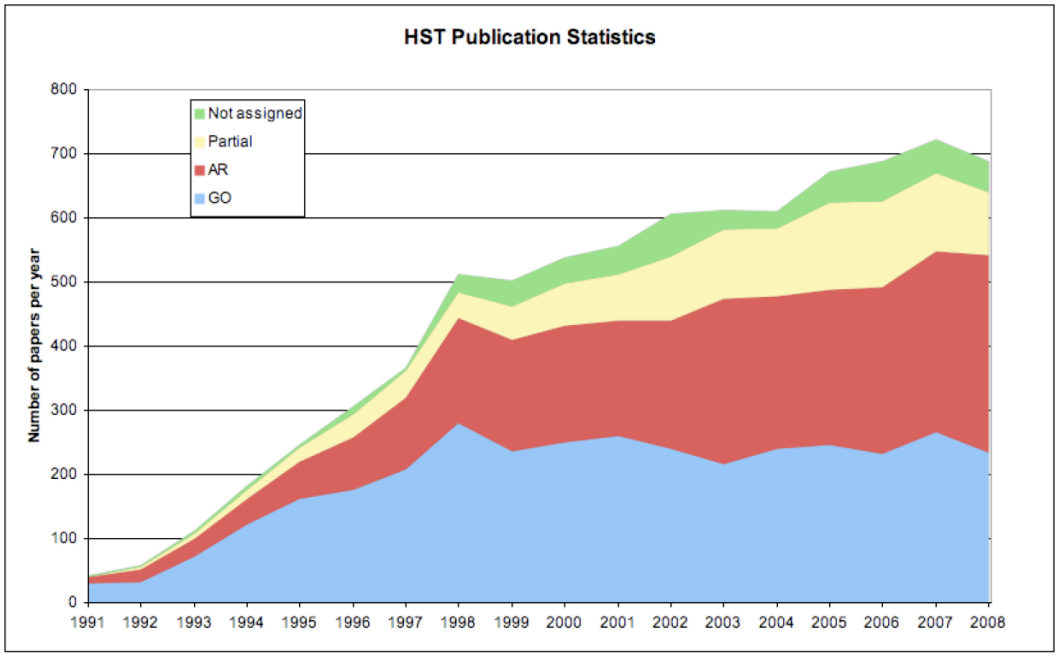


Figure 1: Statistics from the Hubble Space Telescope mission. For many years the number of papers that are partially or totally based on archival data has surpassed (by more than 2-to-1 in 2008) the number based on proprietary data. The CADC HST numbers are included in these statistics and we have made an important contribution by developing advanced data products for WFPC2 and ACS. (From STScI)

The CADC was formed in 1986 in response to a community need for access to HST data. In 1988 CFHT data began to be archived at CADC and, after launch, HST data began to flow through CADC as one of three data centres.

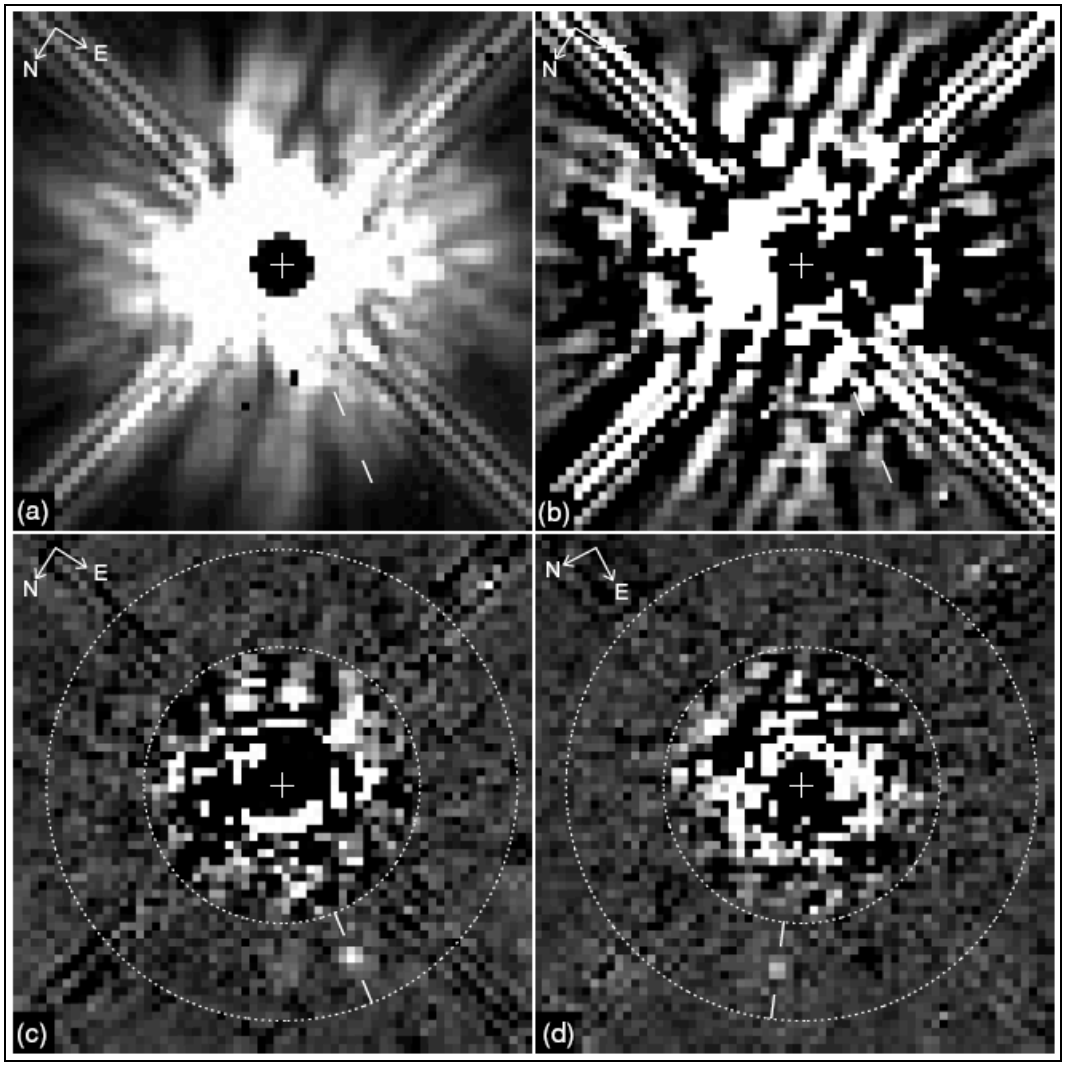


Figure 2: An example of the value of data centres like CADC is the recovery of planet HR 8799b, discovered in 2007, in archival HST/NICMOS data from 1998 (LaFreniere et al 2009, ApJ 694, L148).

CADC is unique. The group has evolved in such a way that it is perhaps the only integrated multi-disciplinary group working on end-to-end astronomy data management. All data management decisions are science decisions. The process that starts with the production of data at the telescope extends as a continuum to the scientific analysis stage on the astronomer's desktop. Decisions about data handling made at the observatory have direct impacts on all phases of data management through to the research paper.

Performance against LRP 1999 goals

Recommendations of the LRP 1999

> "• ***The LRPP strongly recommends*** that NRC's outstanding CADC develop its ability to manage archives of data from upcoming space and ground-based observatories. Funding should be provided to develop innovative data mining techniques that maximize the scientific usefulness of multi-wavelength observations in astronomy."

The LRP recommendation has two distinct components. The first is that CADC continue to develop its data management techniques for observational data and to extend these techniques to deal with the rapidly-growing scale of datasets. The second component is that CADC develop its vision for extracting knowledge from large, multi-wavelength datasets. This vision (referred to as "Data Mining") is equivalent to the vision of the international "Virtual Observatory" movement whose goals are to provide integrated access to global datasets and to develop and deploy advanced services to explore those datasets. Broadly speaking, the future direction of CADC depends on the balance between these two components.

The LRP panel recommended $300k/year for CADC and this money was spent on 4 hires: 2 additional science staff positions, a database administrator, and a software developer/designer.

The LRP Mid-Term Review

The LRP Mid-Term Review (MTR) in 2004 concluded that the "CADC has largely met the goals set out in the LRP. To meet the increasing challenges created by the rapid growth of the data volumes worldwide, it is now timely to consider how best to plan for increased effectiveness of the CADC and to ensure that Canadian scientists will have the tools to retrieve and analyze their data from LRP and other large facilities in an effective manner."

However, the MTR report was concerned that large computing and storage capacity "may be outstripping NRC's capacity to provide the necessary resources" due to the "explosive growth in astronomical data worldwide." The panel felt that "a fundamental reassessment needs to be conducted at an international level to determine how to manage the worldwide data explosion associated with international observatory facilities." The panel recommended "that NRC-HIA conduct a review of Canada's role in global data management and the CADC's contributions to this role" and that "LRP support for CADC should be continued to help maintain the strengths of the existing programs".

Further, the panel recommended that CASCA conduct a review of the data retrieval and analysis requirements of all LRP facilities and use this as input to the NRC review of CADC's role.

The reviews of astronomy data management recommended by the MTR panel did not take place. There has not been a fundamental reassessment at an international level to determine how to deal with data management issues for new projects.

Canada needs a clear data management policy. The development process for major astronomy facilities typically does not include end-to-end data management plans until very late in that process. This practice has been persistent even as the scale and sophistication of data management have grown to the point where data management is nearly on the same effort and cost level as instrument development.

Performance against LRP 1999 goals

The first of the LRP recommendations is clear and evaluating performance against that goal is simple. The LRP panel was concerned about the rapidly-increasing scale of data being produced by new facilities, in particular MegaPrime. CADC has dealt successfully with this challenge. An MOU was signed between CADC, CFHT and Terapix to manage the data produced by the Legacy Surveys and CADC successfully handled all of the data from MegaPrime. CADC has also taken on a major role in

the JCMT Legacy Surveys (defined by MOU) including being a data processing hub as part of JCMT operations in addition to storage and distribution.

The second goal was to enhance "data mining" capabilities. The CADC vision of the future is was ambitious and far-reaching and required substantial cooperation with international partners. This vision is fundamentally equivalent to what has become known as the "Virtual Observatory". Much work has been done both locally and within the International Virtual Observatory Alliance (IVOA) in pursuit of this vision.

The CADC has a number of successes it can claim with respect to data mining. Massive processing has been done on HST data collections (WFPC2 and ACS to produce advanced data products. (A true "data mining" result was produced as a by-product of our processing of the full WFPC2 catalogue. We discovered a significant drift in the pointing of HST during long visits that was previously unknown.) Large-scale processing and catalogue generation have been done by Terapix and by MegaPipe. CADC has developed several prototype services for multi-wavelength data query.

A major change in the way that CADC collections are accessed is produced by the ability of users to write scripts that remotely access CADC collections and download data. In 2009 this "scripted" or "programmatic" access accounted for 80% of our delivered data volume to several hundred users who exploited this capability.

The IVOA has developed capabilities for Registry and for Simple Image Access and for a number of other services. Some of these protocols are widely deployed (hundreds of SIA services including those at CADC) but if the metric of success for the Virtual Observatory (VO) is the delivery of services to research users then the Virtual Observatory has had very modest success up to the present point. The number of users is fairly small. The fundamental problem of seamless user access to multi-wavelength datasets is completely dependent on international collaboration. That collaboration is taking place but has been slow to produce results.

In 1999 we predicted strong growth in the use of databases to deliver catalogue content to science users and as platforms for database-level data mining. This growth has been slow largely because catalogue production and distribution have been slower than our optimistic estimates. The SDSS databases have been a major exception and demonstrate the types of capabilities that we expect to become more widespread. It is very difficult to produce fully-automated pipeline processing that produces science-ready catalogues.

In summary, we have made progress in developing VO and data mining capabilities. We are on the verge of exploiting a great many years of work on developing VO standards in the deployment of the next generation of distributed data access and on deploying a general Table Access Protocol for distributed database access. The CADC, working within the IVOA framework, has played a leading role in both of these developments.

CADC in 2010

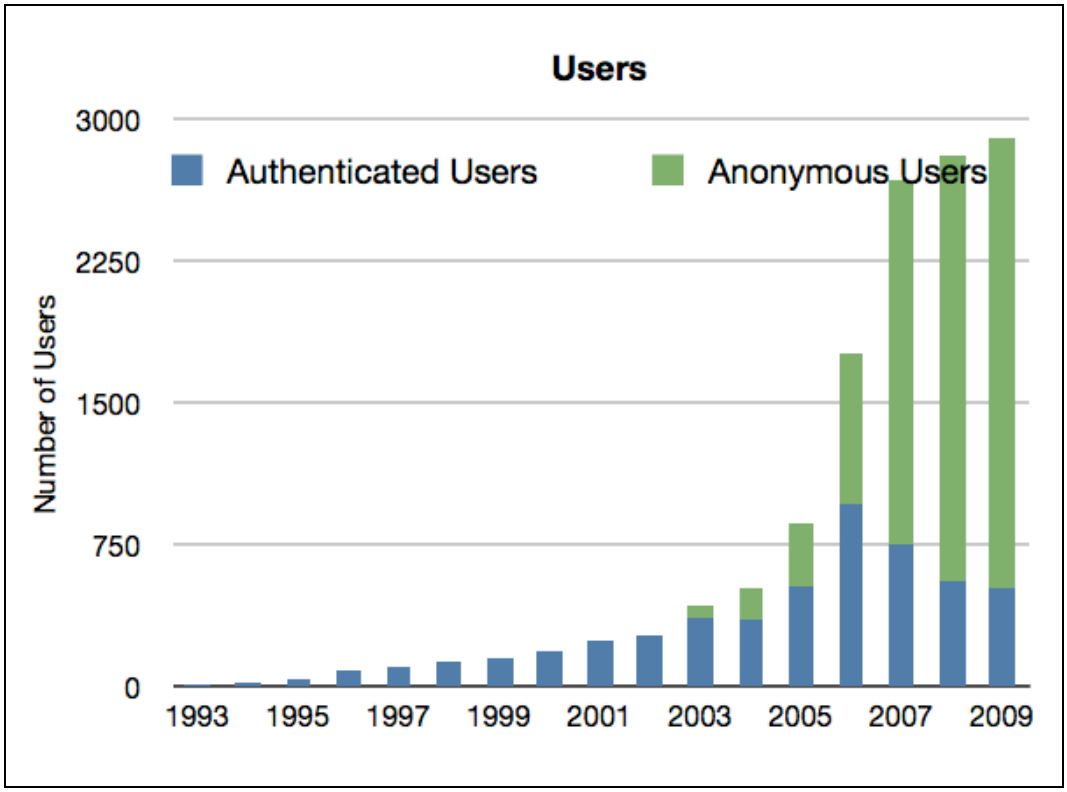
CADC Status

CADC has evolved dramatically in the past decade. In 2009 we delivered proprietary data from Gemini, JCMT, and CFHT as well as serving public archival data. In 1999 CADC served 151 users with data from CFHT and HST. In 2009 CADC served data to 2900 distinct users of 7 different archives. In 1999 we served only archival (non-proprietary) data. There has been extraordinary growth in the number of users that download data from our collections. This is an indicator of success. But the most direct indicator of success is the number of refereed publications produced by our services. We cannot reliably get that number because our users do not regularly acknowledge the use

of our services. From a past survey of users we estimated that 25% of those that use our services for publication actually acknowledged us in those publications[1].

There are two aspects of CADC that are very different from the situation in 1999. First, CADC is now an integral part of observatory operations for Gemini, CFHT, and JCMT. Second, it is an integral part of science activities with CFHT and JCMT Legacy Surveys and with major surveys through the CANFAR project. The two ends of the data management streams are the observatory and the science user. The activities of CADC form a continuum between these two endpoints. CADC is integrated into observatory observations and survey data management. Now we want to move toward tighter integration into the work of science teams and we have made steps in that direction, most recently with the CANFAR project.

A major strength of CADC is that it produces and operates persistent, stable services that are maintained over long time scales. The challenges for the future are twofold. First, to ensure that the stability of CADC services is maintained (in an uncertain funding environment) and, second, to ensure that CADC evolves along with the community. The most effective way to achieve a sharply-focussed understanding of the needs of the community is to be integrated with science teams, for example by having staff members working as part of those the teams. This approach will drive us to develop the right services to support science. And those CADC services will then be integrated into the operational practice of science teams.



[1] Information from publication sources is incomplete. The NASA ADS full text search does not yield current results. Searches using Google Scholar and astro-ph full-text searches yield 1137 and 615 (423 since 2004 from astro-ph) papers respectively (which contain the string "Canadian Astronomy Data Centre" or alternate spelling) which is an indicator of our impact but not a reliable one. It includes many non-refereed sources. Google scholar shows that CADC was referenced by 121 publications in 1994-1999, 311 publications in 1999-2000 and 622 publications in 2004-2009. There are serious reliability issues but the trends indicate growth in the science impact of CADC along with the growth in number of users. We need to improve our ability to assess impact.

Figure 3: The growth in the number of distinct users that downloaded CADC data. Authenticated access is necessary for proprietary data (some Gemini, CFHT, and JCMT data). Most access is now anonymous.

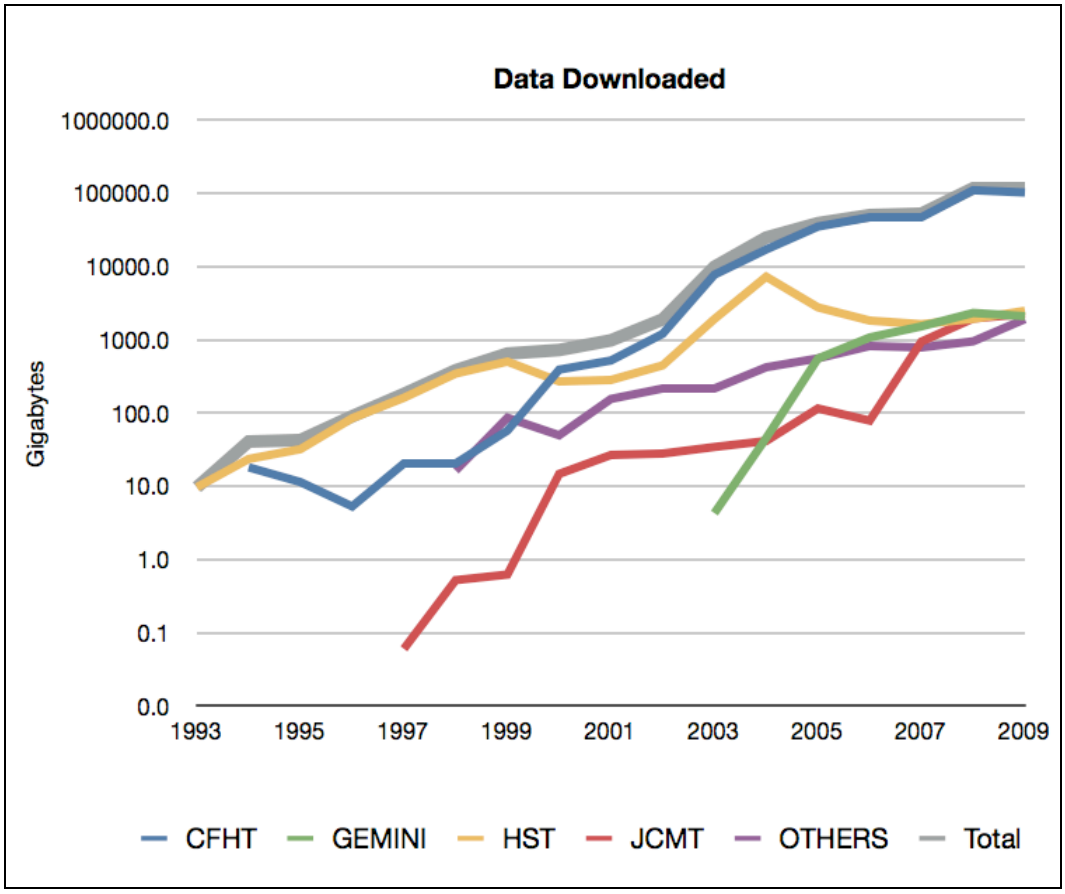


Figure 4: Total data volumes downloaded and volumes by data collection.

CADC Successes 1999-2009

Successes over the past decade have included:

- Developing and operating the Gemini Science Archive
- Hosting the Canadian Galactic Plane Survey
- Developing Advanced Data Products for HST WFPC2 and ACS
- Collaboration in developing the Hubble Legacy Archive
- Operating the MegaPrime pipeline
- Safely storing and delivering CFHT MegaCam data
- Managing the CFHT Legacy Survey data
- Creating programmatic access to allow users to write scripts to access CADC collections
- Vigorous contributions to International Virtual Observatory Alliance standards, development, and governance
- Development of the JCMT Science Archive which includes an operational data processing role for CADC

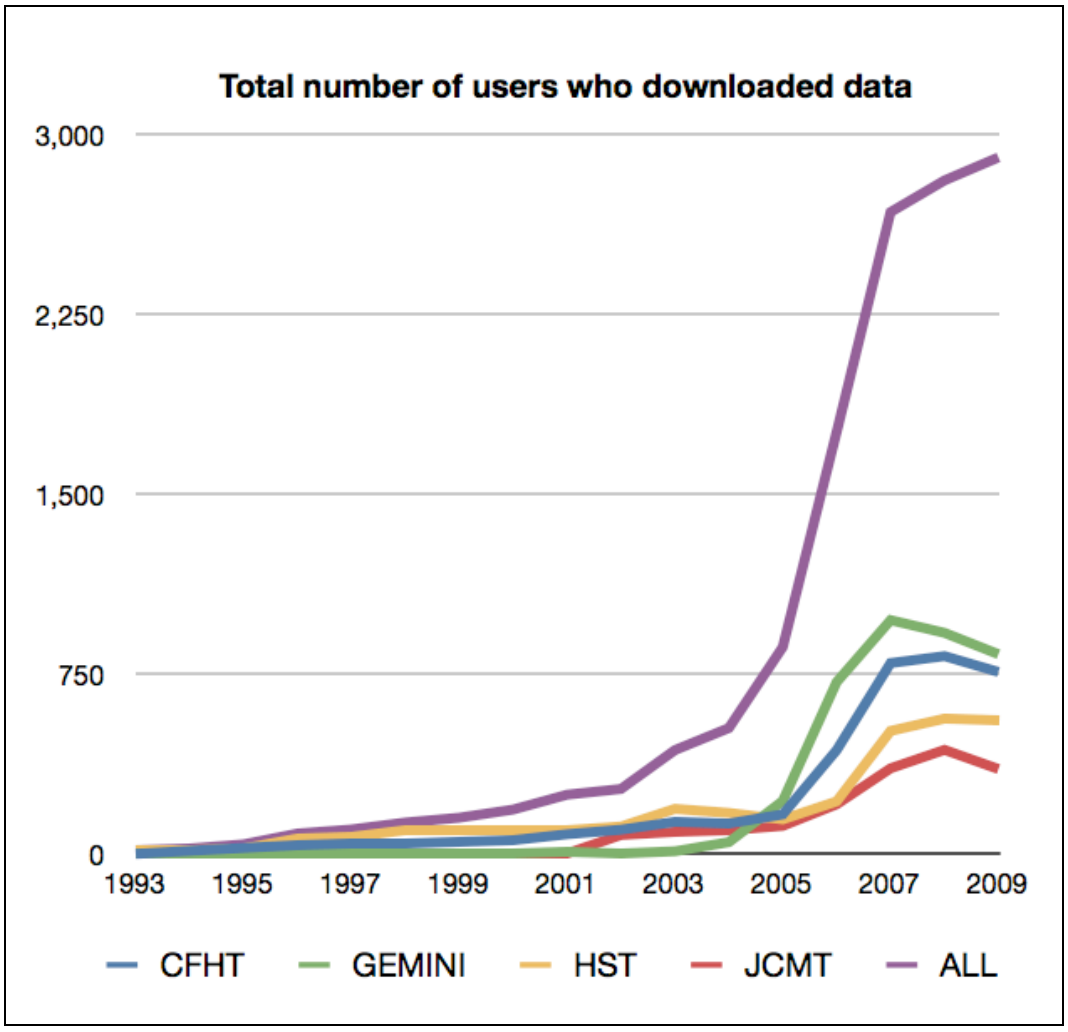


Figure 5: Number of users that download data by archive and in total.

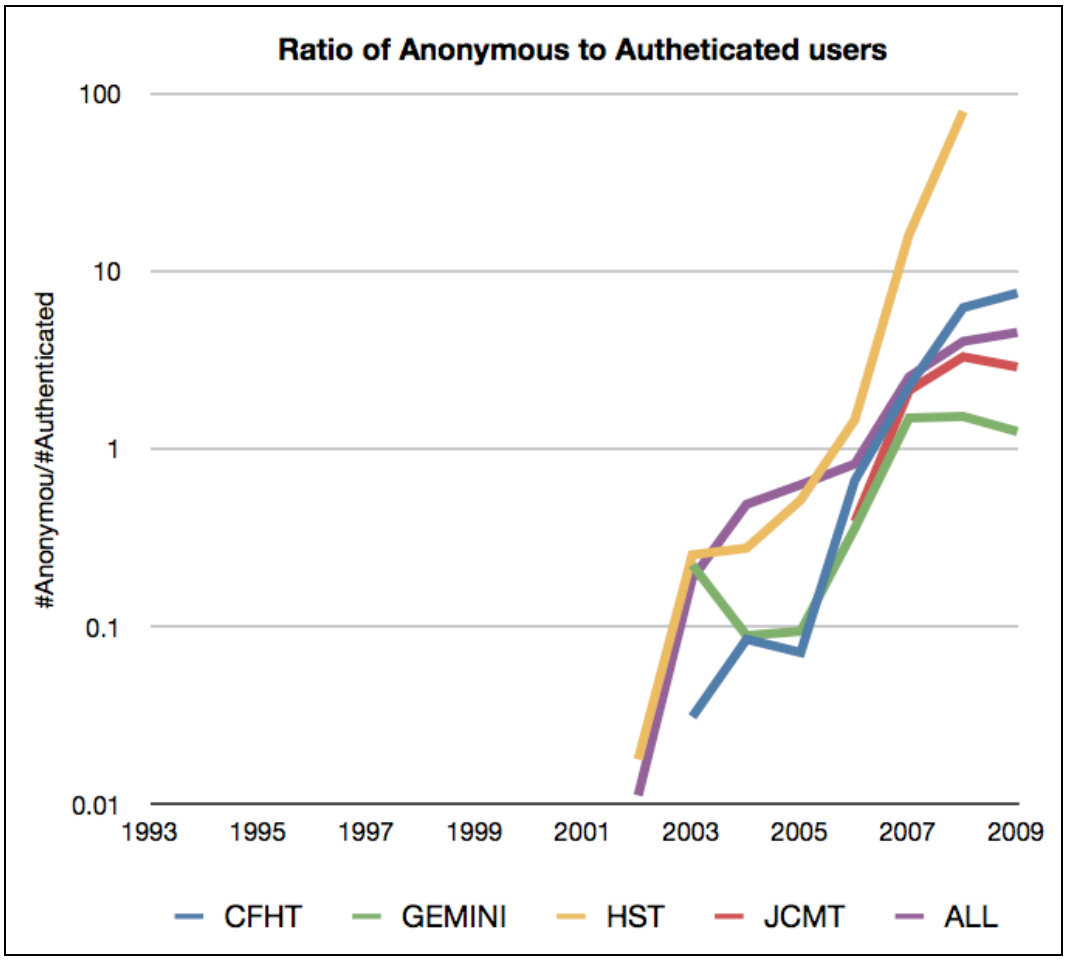


Figure 6: The ratio of anonymous to authenticated users. Anonymous access is always to public data (all HST data that CADC serves is public). Authenticated access is often (not always) to proprietary data. The figure indicates that archival research on non-proprietary data is important for most archives (ratios in 2009 of 7.5 for CFHT, 2.9 for JCMT). Even Gemini data is access more often (ratio of 1.25) anonymously than when it is proprietary.

Strengthening the CADC model

The CADC has succeeded because of a tight partnership between astronomers and technical staff (software developers and operations). The purpose of the science staff is to maintain the focus of the group on community science goals. They act as proxies for the general community and they interact with the general community and bring knowledge of the community's needs back to the group. They apply their technical expertise to linking science goals to solutions. They also work on various projects, defining requirements and applying their science expertise to data engineering work. The partnership of science and technical staff reporting to the same Group Leader is critical to the future of CADC.

CADC has taken on more serious operational roles in its collaboration with CFHT, Gemini, and JCMT. The benefit is improved data quality and science usability at the source through interactions between observatory and CADC staff. It is part of the integration of CADC into the astronomy landscape. It means that it is feasible for scientists to have access to data from CADC a few minutes after the observation is complete.

The other end of the spectrum is the science user. We have been responsive to community needs judging by the growth in our usage. But we need to continue to improve particularly because the

community is so diverse that a small science staff cannot be an effective proxy for the whole community. We cannot have an expert in every science area and wavelength regime on staff.

A promising solution to this problem is suggested by the structure of the CANFAR project which is a collaboration of the University of Victoria, the University of British Columbia and HIA/CADC. The science activities are in the universities as well as at HIA and CADC acts as the technical hub in support of the science teams. Technical development depends partly on a UVIC-based infrastructure development team in High Energy Physics. This project will produce an interface for astronomers into Infrastructure as a Service (IaaS) on Compute Canada hardware but it will also provide close support to science teams.

The CANFAR science teams rely on CADC for processing, storage, databases, and collaborative tools. They are close collaborators in the development of these services and CADC benefits from the fact that these users express their needs directly and constantly. This small number of postdoc users represent the needs of many astronomers on the team and we expect the systems that are developed with their input to have general usefulness to the entire community. It is important that the science teams in such a project are widely distributed in Canada and in science area.

We need to keep this type of vigorous interaction happening between users and CADC. We need a mechanism, including funding from NRC and other sources, to do this. As part of this model, we propose at least two postdocs within CADC who can do data-intensive science and who can be drivers of development and debuggers on a day-to-day basis. The postdocs should be part of university-based research teams but reside at CADC and could be part of a jointly funded (NRC/NSERC?) program.

Cost and Value

The total budget for CADC is around $2m per year. About 25% of this is capital. In the past 10 years the majority of our capital requirements have come from sources external to HIA (for example the NRC Major Initiatives Committee, $3.0m 2002-2010).). While CADC has enjoyed success at securing capital in this way, there is no guarantee for the future; accordingly, we are now exploring vigorously the extent to with the Compute Canada infrastructure can be incorporated into the evolving needs of data-intensive astronomy. Over 60% of our staffing costs are covered by revenues (for example Gemini, and the fee-for-service CANFAR contract with the University of Victoria). These sources of funding would not otherwise go to astronomy. CADC has been successful in attracting these revenues to support its activities but also produces vulnerability. At this point in time, about 7 of our 20 staff are term positions that will end in 2011.

CADC provides value in a number of ways. We offload data management costs from observing facilities and projects. These activities would need to be funded by other means if CADC did not do them. We manage data for Gemini, CFHT, JCMT, CGPS, MOST, BLAST. It is difficult to make an accurate estimate but the cost of these services to the observatories or facility teams would certainly be more than $300k in staffing and a larger amount in capital and operations.

The CADC develops and operates services to support research that are excellent compared to similar facilities globally. Further, we support Canada's role in international data sharing reciprocity. We are part of the astronomy culture that maintains a very low level of restrictions on data. We share Canadian collections (CFHT one of the most important) and Canadian researchers have access to valuable datasets from the U.S. and Europe. This is good for science.

It is a great benefit to have multiple data collections in one data centre. The centralization of these activities in a specialized science-focused group produces savings. More importantly, it drives an understanding of the need for integration of multi-wavelength data and it drives the work toward achieving that integration. Forefront astronomical research in 2010 is a multi-wavelength activity.

Long-term curation of data collections is another area where CADC makes an important contribution. We have preserved and protected the valuable data produced by expensive observing facilities.

Issues

1. CADC activities must be driven by the needs of the Canadian research community and must operate as an integral part of that community. The most effective way to achieve this is to establish collaborative projects, similar to CANFAR (whose funding will end in Spring 2011), where CADC interacts on a day-to-day basis with scientists and science teams.
   - **Issue:** A model needs to be developed which ensures ongoing vigorous interaction between CADC and its user community. The model should have CADC staff and university-based scientists functioning together as a team in a fully integrated manner on a day-to-day basis. The model should include science members distributed among Canadian universities using CADC as a stable and persistent technical hub. The benefits derived from such a model will benefit all Canadian researchers.
   - **Issue:** The role filled by postdoctoral researchers at CADC is critically important in the early phases of prototyping and debugging; it brings in fresh viewpoints and skillsets, and provides immediate feedback to developers. We propose that at least two postdoctoral researchers in data-intensive astronomy should be located at CADC. They would ideally act as part of university-based science teams in data-intensive areas.
2. The absence of a Data Management policy for Canadian and international facilities with Canadian contributions is a growing problem. Canada typically participates in major facility projects without clear agreements with its international partners about data access and management. At this point in ALMA, JWST, ASTROSAT and other projects there is not yet a defined end-to-end data management structure. The absence of planning also creates instability for CADC (which links to item 5 below) at a time when the amount of work needed in astronomy data management is increasing rapidly.
   - **Issue:** Canada needs to develop a policy for addressing Data Management issues in the ground-based and space-based projects in which it participates. Data management should be discussed early in the project development process. The experience and expertise of CADC should be considered as one of the resources available that may, in some cases, be beneficial in this planning process. CADC should be considered as a potential partner in data management for new projects.
3. Compute Canada and the regional grids need to be part of a stable model for delivering computing infrastructure to science users including observational astronomers. The current model neither serves this community nor is stable. As more of CADC's critical infrastructure is moved onto Compute Canada facilities we are increasingly tied to the destiny of that organization which is very new and whose future funding is uncertain. (This item links to uncertain capital support for CADC which depends on the successful use of Compute Canada resources.)
   - **Issue:** The Canadian astronomy community needs a stable model for the long-term provision of computing infrastructure. The Compute Canada model is attractive but it needs to be part of a well-defined, long-term commitment from government.
4. The long-term preservation and curation of astronomy data needs to be addressed. CADC is a central repository for data from observing facilities with Canada as a partner. We will soon be faced with difficult decisions, for example continuing to host the JCMT archive after the telescope is closed, which requires ongoing spending to preserve the legacy value of data.
   - **Issue:** We recommend that funding should be provided specifically for the purpose of long-term preservation of the value of astronomy data. CASCA can help define the urgency of this problem and NRC can lead the way in this area which is of global concern.

5. CADC is vulnerable because of its reliance on revenues. Canada has invested for over 20 years in creating a national resource that is of great value in supporting the work of the research community. Ongoing A-base commitments cover less than 50% of the group and the group has no ongoing capital budget. About 1/3 of the group are on term positions that will end in 2011.
   - **Issue:** A funding model for CADC needs to be developed which reduces the risk to the future existence of the group due to reliance on external revenues.
6. CADC met the goals of the LRP recommendations in 1999 with the help of the recommended 4 positions. These science and technical staff continue to be critical to the mission of CADC.
   - **Issue:** The four new staff positions created in response to the LRP recommendations in 1999 need to be continued.

Vision for the future

CADC will be fully integrated into the scientific mission of the Canadian research community and will make strong contributions to end-to-end data management for astronomy projects with Canadian involvement. Global integration of data collections and services as part of an international Virtual Observatory will be achieved in the coming decade.