Data Repositories at the CfA

Arnold Rots, Pepi Fabbiano, Raffaele D'Abrusco, Tom Dame, Eric Keto, Warren Brown, Sean Moran, Ed DeLuca, Mark Weber, Trae Winter, Kathy Reeves, Alberto Accomazzi, Larry Rothman, Youli Gordon, Roman Kochanov, Josh Grindlay, Sylvain Korzennik, Daina Bouquin, Matt Ashby, Randall Smith, Ian Evans

Context and Scope

- This group of 21 staff members aimed to put together a comprehensive white paper on the state of data repositories at the CfA and the unmet needs
- All of us are involved with one or more data repositories, which means that we know what we are talking about
- It also indicates that a significant fraction (at least 10%) of the CfA scientific staff is involved with repositories, meaning that it is a substantial part of the Center's mission
- As such, data preservation needs to be an essential component in the Strategic Plan and this paper intends to provide a blueprint of how it may be integrated

Scientific Data at the CfA

- CfA as the world's largest astronomical institution produces and stores a large volume of world-class scientific data
- Scientific challenge: how to maximize the science extracted from these data?
- Technological challenge: the amount and complexity of the data requires a professional approach to support the scientific challenge
- The preservation of CfA-held data in permanent, accessible, and searchable archives can solve this challenge, and can do so for a modest investment of resources

Why Data Archives Are Important

- For CfA to stay a world-leading institution, it needs to keep upping its game in the "data experience" provided to researchers, as well as to funding institutions, and to posterity
- This requires a quality of support that is no longer reasonable nor desirable to shift onto researchers, but requires more dedicated professional support
- There are four main themes in identifying the importance of data archives for 21st century astronomy and astrophysics

Importance of Data Archives (1)

- 1. Enabling great science
  - An important way to make the most of the CfA environment is to make its datasets as widely available and easily accessible as possible, and to ease the development and integration of new projects
  - But to do that better requires dedicated support
- 2. Enabling data reuse to multiply the value of the data
  - At present most data at CfA are not easily discoverable and cross-searchable outside the nearest area of expertise, even though data preservation and access are now mandated by our national funding agencies, NASA, NSF, as well as SI
  - Making data more readily available enhances the reputation of the institutions that do so and statistics on the use of such data archives is increasingly being used as a measure of the success of funded research

Importance of Data Archives (2)

- 3. Preserving unique datasets for posterity
  - Long-term preservation for posterity does fit under the Smithsonian's mandate
  - DASCH, for example, shows how century-old data sets can have great value in astronomy
  - It would seem that our role in preserving and serving such datasets indefinitely for the benefit of the U.S. and all humanity is practically thrust upon us from two directions, in CfA's role:
    - as a world leader in its involvement with a large number of astrophysical observatories and projects of all types and sizes
    - and as part of the Smithsonian

Importance of Data Archives (3)

- 4. Access to analysis and data mining tools
  - If data are available through established interfaces (IVOA, SVO), existing tools conforming to these interfaces can be employed in scientific analysis, thus increasing the productivity of CfA scientists
  - We advocate that the CfA, as a major astronomical data center, take a proactive stance regarding the exploitations of these data
  - **Data science** expertise at CfA should be actively encouraged, requiring a spectrum of cross-disciplinary skills and methodologies, some of which are already present at CfA:
    - Time-resolved astronomy (DASCH)
    - Spatially-resolved dataset (solar activity, solar features, etc)
    - Techniques for multi-wavelength classification (SEDs)

Areas of Support

- There is currently no CfA-wide policy, model, or support for the **development**, **maintenance**, and **preservation** of our data archives
- Instead they are scattered and are often hard to find, or query, largely due to their widely varying levels of financial and technical support
- To remedy this situation and to allow all CfA projects to comply with SI and other data policy mandates, we identify four main areas where support is urgently needed and should be included in the Strategic Plan

Areas of Support (1)

- 1. Discoverability and Access
  - Like most prominent astronomical institutions, CfA needs an **archive portal** that provides the community access to all our data repositories
  - It should provide a comprehensive overview of all data available from CfA and make the data products searchable and discoverable
  - As a simple first step the website should contain a prominent page with links to the existing access pages
  - The next step would be the design of a common portal based on standard access protocols, such as those provided by the IVOA

Areas of Support (2)

- Project Data Management Plans (PDMP)
  - As PDMPs are mandated for many if not all of our research projects, support for developing such plans is crucial
  - This can be overwhelming to those who have never prepared one before
  - CfA needs to provide support for PDMP development in the form of guidance, templates, and consultations
  - This is not a huge job, but such support is essential

Areas of Support (3)

- Project Support
  - There is considerable variety among the data repositories at CfA, in their scope, objectives, target audiences, size, resources, funding, and the expertise of their staff:
    - **Large projects** generally have the resources and expertise to manage the development and maintenance of their archives; still, there is the issue of who will take care of a project's archive after it has ended
    - **Small projects** with limited funding may need more support, but it will often be consultative in nature.
    - **User-contributed high-level scientific data products** (like those included in publications) generally are fairly small in data volume, but allowing them to be discovered is challenging

Areas of Support (4): categories

- System design
- Hardware configuration
- Hardware acquisition
- Enterprise level storage
- Disaster recovery plan
- Cloud computing expertise
- Virtualization
- Networking and firewall issues
- Software tools needed
- Databases
- Interfaces
- Policy and institutional requirements
- General consultation
- Disposition after project termination

Areas of Support (5)

- Project Support (cont'd)
  - Establishing **links between related datasets and between datasets and the literature** (the ADS being one of our repositories) has gained prominence
  - This will continue to gain in importance in a future where data mining and handling of Big Data are becoming part of standard research tools
  - And they provide meaningful performance metrics
  - The principles and tools for this type of data repository tools will need to be detailed

Areas of Support (6)

- Data Science at the CfA
  - While data science is not astrophysics, data science is and will be more and more needed to fully exploit the CfA data
  - We should invest in a coordinated effort for developing and making available the more advanced tools needed for the scientific exploitation of the data
  - We need support for expertise in this area and for coordinating these efforts, to maximize the scientific return of astronomical research at the CfA and everywhere CfA data are used
  - This includes the development of sophisticated data mining and statistical tools for the analysis of increasingly larger surveys, in which CfA scientists are and will be involved

Realizing the Support

- The good news is that much of the needed expertise and resources is already present at CfA:
  - The CF clearly has the technical expertise and could shift its focus to that type of support; this may also include access to SI resources in this area
  - The library's expertise in organizing and disseminating information and resources is an obvious match
  - The existing community of data repositories, especially the better supported ones including the ADS, are well positioned to share their experience and expertise

The Bottom Line

- Some funding will be required
- But we do not need to start from scratch and some resources are already available
- For one thing, *improved communication and sharing within the CfA data repository community will help tremendously*
- **But it has to be clear that responsible care for our data is a core part of the CfA mission, an essential part of the scientific endeavor, and recognized as such in the Strategic Plan**