The Emergence of Long-Lived, High-Value Data Collections

Alexander S. Szalay[1] *and Barry C. Barish*[2]

1. Dept. of Physics and Astronomy,
   Dept of Computer Science,
   The Johns Hopkins University,
   Baltimore, MD 21218
   Email: szalay@jhu.edu
   Phone: 410-516-7217

2. Department of Physics,
   California Institute of Technology
   Pasadena, CA 91125
   Email: barish@ligo.caltech.edu
   Phone: 626-395-3893

Consideration area:

data archiving, long-term data preservation,

The Emergence of Long-Lived, High-Value Data Collections

Alexander S. Szalay[1] and Barry C. Barish[2]

3. *The Johns Hopkins University, Baltimore*
4. *California Institute of Technology*

Many large-scale astronomy projects created unique high-value data sets at a cost of hundreds of millions of dollars. While the projects are active, they are mandated to create highly sophisticated, smart archives, used by a large community. However, there are no significant efforts about maintaining these data beyond the point when their instruments reach the sunset. This will soon lead to a serious data loss unless there is a focused effort to come up with long-term policies and sustainability solutions. Here we outline the challenges involved and make a few recommendations toward a sustainable data infrastructure. While many of the problems are more general than astronomy, it is likely that the first real challenges in this space will impact current astronomical surveys and require immediate attention.

Enormous National Investment in Data

Large data sets are fundamentally changing science. We have many major research efforts, often spanning decades, collecting unheard amounts of data, typically as part of a large collaboration. Much of the analysis of these large, open data sets are carried out by a much broader community, who access the open data archive, typically through smart data services. There is an emerging trend in creating Science Portals, intelligent, collaborative data analytics environments, where much of the data analysis could take place. This whole effort has started a few decades ago, first in High Energy Physics, followed by many others. Today we have such projects on almost every scale of the physical world (LIGO, LSST, ALMA, SDSS, EOSDIS, EarthCube, OOI, NEON, Materials Genome Initiative, Human Genome Project, Fermilab, CERN Large Hadron Collider,…). The resulting data are so unique that they will have a useful lifetime extending to several decades, possibly as much as 50 years.

The overall cost of these projects is often in the hundreds of millions if not billions of dollars. Their data collections represent an enormous national investment. We need to explore how one could benefit more from economies of scale over the long term, and build a common data ecosystem, shared by all the projects participating in the data preservation effort.

Sharing much of the infrastructure for the 'active' projects has been made difficult by the realization that a lot of them require very domain specific tools and expertise through much of the vertical software stack. Oceanographer's tools have little in common with genomics, particle physics, or astronomy. As a result, much of the development of the data processing and analysis stack has been done in isolation, often "reinventing the wheel", even for components that could be otherwise shared.

The Challenge of Long-Term Preservation

However, it gets worse. As the first generation of these large scale projects are getting close to the end of their lives as far as data collection is concerned, as the original instruments are slowly becoming obsolete, a new challenge is emerging: *what happens to the data after the* *instruments are shut down?* This is a much harder problem, than it may first appear. In order to tackle the problem, we need to establish a common "currency" on which one can make easier comparisons and try to formulate a rational decision making process. Let us try to establish the three different aspects of the business model for these surveys: the price, value and cost of the data.

The Price of Data

The data collection created by a Big Science project represents a major public investment of a few hundred million dollars. This includes the capital investment in the experimental facility, the data infrastructure, the cost of operating the instrument, reducing the data, and building and operating an open and accessible data archive. This is the price of the data, this is how much the federal government (in some cases augmented by private foundations and individuals) has paid to create this singularly unique resource. Generally, this process is well understood, and all of the aspects of the projects are well under control while the experiment is running.

The Value of Data

We can also ask how we could estimate the value of the data. It is clearly reflected in how much science it generates. While it is difficult to put a monetary value on the results of scientific research in an algorithmic fashion, we can use another approximate metric. Each scientific paper published in a refereed journal represents a research effort that costs approximately $100K (an estimate, but certainly more than $10K, and less than $1M). This is the amount of research funds spent on paying for students, postdocs, research tools, computer time, to be able to write a credible scientific publication. The number of papers based upon the analysis of a given data set are then measuring how much the members of the research community are willing to spend from their own research funds to work on this particular data set, they vote on the value of the data with their research dollars.

The Cost of Data

This is the third component of the problem. This is measuring how much it costs annually to curate, preserve and keep serving the data to the community in an open and accessible way, after the original instrument has been turned off, and there is no new data added to the archive any more. This is more than archiving, as the data use is through intelligent software interfaces, often based on a large database, combined with a collaborative data analysis platform. This requires a lot more than just copying data on disks. Operating systems change, database systems change, web browsers change, computing hardware changes, and the user's expectation is also increasing with time. A few years ago they were happy to download a few flat files and analyze them at their workstations at their home institution, today they are expecting access to iPython notebooks, and GPUs, but soon they will want to reprocess petabytes of data on hundreds of computer nodes interactively. Much of the cost is not so much in saving the bytes, but rather keeping the services alive, and up-to-date. As the cost of storage and even computing cycles keep decreasing every year, the dominant part of the costs are mostly in people.

Comparison of Price, Value and Cost

From our 20 years of experience with the Sloan Digital Sky Survey, the price of the data to date has been about $200M. The project's data has generated to date about 9,000 refereed publications, i.e. attracting about $900M of research over this period. After operating the archive for 20 years, we estimate the cost of maintaining the necessary technological advances into the future is approximately $500K/year. Let us express this annual cost in terms of the price: $500K/$200M = 0.25%/year. We can see that a 5% addition to the project's budget would secure the archive for 20 more years. Or, if the continued operation of the archive results in just 5 refereed papers in a year, it is still a reasonable investment to keep the archive alive.

The same numbers for the Large Synoptic Survey Telescope, the current national flagship project for astronomy, are similar. The price is expected to be around $1.2B by the end of the project, and the cost to be about $6M/year, i.e. $6M/1.2B = 0.5%, still in the same ballpark.

These costs are quite trivial compared to the price of the data, yet we have no coherent plans or long term funding mechanisms in place to address this problem. A potential loss of one of these data sets would create an enormous damage to science, and endanger the national willingness to continue more future experiments, if we cannot demonstrate that past investments are adequately protected, preserved and cared for.

The Challenges Ahead – A Sustainable Data Infrastructure

Up to now much of the long-term data preservation has been though journals and publishers. In a way "we threw the data over a fence" and hoped that there is somebody on the other side who will catch it. There were journals and publishers who did this, for a profit. The journals were accessible through the research libraries. Today, subscription fees are skyrocketing, and there is an open war between the remaining few mega-publishers and the science community. It is clear that the whole system is in flux, and the digital revolution is disrupting yet another aspect of science. The role research libraries have played in the past needs to be reinvented for the digital age.

With the increasing amounts and complexities of data, scientists will still need intermediaries who perform the function of preserving and presenting knowledge to scientists, but it is increasingly likely that this needs to become part of the open science enterprise rather than handed off to commercial publishers.

We have talked about the cost of the long-term data preservation, but not about how services would be maintained. Generally, we (as a community) have invested a fair amount of funds into Cyberinfrastructure, developing a variety of data services and software tools. The underlying (often implicit) expectation is that these services will at some point become self-sustaining. Let us analyze the different aspects of this problem.

On one hand, the goal is to provide open and free data services for everyone, in line with open science, and the increasing democratization of science. On another hand, open data is only useful if it is accessible in a practical fashion, implying smart services operating at no cost to the user. We could (and almost definitely will) place open data into a commercial cloud, but if every user had to pay with their credit card for both the data access, computations and data download, there would be a lot of unhappy users. Furthermore, in this model somebody still has to pay for the ongoing improvements and evolution of the hardware/software environment and the user interfaces. Some projects may be able to solicit crowdfunding, but even the arxiv.org has not been able to survive on voluntary donations only.

In many ways the three facets of long-term data preservation: (i) open/free data, (ii) usable and accessible data and (iii) self-sustaining funding model are similar to the three legs of project management: features, cost and time. You can pick any two and the third is then given. If we want free and accessible data, somebody will have to pay for the services. If we want free and sustainable, then we can drop the data on disks, and leave it there, and hope that nothing bad will happen. Finally, if we want accessible and sustainable, we have to revert to the traditional publishing model where the access is restricted to paying customers.

Who Do We Trust Over the Long Term?

We should look through the required organizational criteria for providing long-term data preservation. We need to find one or more organization

i. with a long track record with a predictable, stable future
ii. is trusted by the science community
iii. that understands knowledge preservation
iv. is technically capable
v. can run under a sustainable model
vi. has no single points of failure

Let us elaborate on these criteria. For the sake of this discussion let us assume that a predictable future means 50 years. Over the last two years some of the federal agencies have not been successful in maintaining a safe haven for their data. Without the science communities' help there would have been a major disruption turning into a disaster for legacy data. It is not clear that even National Laboratories or Federal Research Facilities established around a large instrument will have a stable mission and existence over 50-year timescales.

Probably the longest existing stable organization on Earth is the Catholic Church, followed by Universities and their libraries. University Libraries have been in the service of scientists for many hundreds of years. Probably none of us has any doubt that our major research universities will still be around in 50 years, and they will have a library. These libraries may be quite different from today's organization or holdings, but their mission of preserving scientific knowledge for future generations remains the same.

While today the digital skills in a typical library may not be on par with some of the more advanced technological centers, they are already disrupted by the digital revolution, and have started aggressively implementing rapid changes required by today's technologies and the changing landscape of scientific publishing confronting the needs of the open science enterprise[1]. An association of research libraries combined with clusters of domain scientists and digital archivists may provide an almost inevitable solution to the long-term challenges. Having several libraries participating as part of a large federation can avoid single points of

[1] http://sites.nationalacademies.org/pga/brdi/open_science_enterprise/

failures. We should definitely explore models in which these libraries could be part of a distributed network participating in long-term data preservation.

Long Term Career Paths

Most of today's data archives are overseen by a high-level management group consisting of the PI's of the instrument, typically internationally recognized domain scientists. While they are tenured either at a national laboratory or at a research university, the people who run the data facilities are on soft money. Their careers depend on continued grant support for the experiment. The senior data scientists (or better "data architects") are uniquely indispensable, they make the "trains run on time". Their expertise in the domain science, combined with deep understanding of the underlying data systems, how they have been built, and what functions should they provide to the community makes the projects successful. They form the bridge between the senior management, the user community, and the lower level programmers who perform the daily maintenance tasks. Yet, they have no stable career paths.

This contradiction becomes even more pronounced when the experimental facilities reach their sunset. It is relatively clear, that at that point something has to change. While it is acceptable that people at a University Department or at a National Laboratory are actively spending their time on managing the data from an ongoing experiment, as data sets move from being "live" to "legacy", the "long-term care" of these data may be done more efficiently in another organizational framework, using more economies of scale. The senior data architects, who have spent several decades of their lives, most of their scientific career, living with the same data collection, are the ones who can best oversee the transition of the data towards a long-term hosting. So, wherever the data goes, they have to go with it, even if the low-level programmers, system managers do not. Yet, we have no existing career paths for these indispensable experts, the equivalents of 'instrument builders' for our data.

Sustainable Funding – a Data Trust?

Libraries are used to raise funds, endowments. If they acquire a rare specimen, like an old codex, it needs resources to preserve and protect. One possible avenue is to seek such endowments to build a Data Trust, where the endowment's interest income would cover the cost of data curation. For the future, one can imagine public private partnerships, where federal agencies would provide a one-time contribution towards a rare data collection, or for example a project running for 30 years would start making annual contributions into a fund that would keep accumulating with compound interest, like a retirement account. We made a simple calculation. If the SDSS project started to put about $120K/year into a 5% interest bearing account in 1992, by today there would be enough funds to support the curation of the data forever off the interest. 5% of the total project costs (less than a typical contingency) would secure the existence of the data for the foreseeable future after the sunset of the instrument. Indeed, the German Federal Government is considering a similar approach, a fixed annual contribution toward the Nationale Forschungsdaten-Infrastruktur (NFDI)[2].

[2] https://www.bmbf.de/de/empfehlungen-zum-management-von-forschungsdaten-3036.html

Discussing these ideas with various people, it has been stated that the US Federal Government has no mechanisms to create such trust funds. However, we found at least two examples for such solutions. The National Endowment for the Arts (NEA) is an independent federal agency[3]. Through partnerships with state arts agencies, local leaders, other federal agencies, and the philanthropic sector, the NEA supports arts, learning, affirms and celebrates America's rich and diverse cultural heritage, and extends its work to promote equal access to the arts in every community across America. Also, The Highway Trust Fund[4] is a transportation fund in the United States which receives money from a federal fuel tax. Such an Endowment can also provide a wonderful opportunity for partnerships where private foundations sympathetic to support large data collection efforts would contribute either with money or with computing and storage resources to this national effort.

Recommendation

The high-valued data sets obtained through enormous, focused investments like large astronomical surveys are special, they will be providing unique scientific value for generations. The Decadal Survey should recognize long-term preservation of these high-valued data sets as a critical outstanding problem for Astronomy, and that in their report they urge the agencies to study the challenge in detail and seek a sustainable solution.

[3] https://www.arts.gov/about-nea

[4] https://www.fhwa.dot.gov/highwaytrustfund/