Digital Archiving of Astronomical Data to Support Publication and Long-Term Preservation

http://ldp.library.jhu.edu/

Johns Hopkins University

Sayeed Choudhury, Tim DiLauro, Alex Szalay, Robert Hanisch, Ethan Vishniac

One of the most fundamental aspects of scientific scholarly communication is the ability to cite and examine data in a persistent manner. Without this ability, the very essence of the scientific method, with its requirement of validating results, becomes compromised. Large-scale astronomy projects such as the Sloan Digital Sky Survey[1] have gathered data at unprecedented rates, raising new challenges and opportunities.

This explosion in data-driven science has led to fundamental changes in practice and modes of inquiry, prompting NSF to advance the evaluation and development of Cyberinfrastructure to support large-scale, digital science projects. Both the Library of Congress' National Digital Information and Infrastructure and Preservation Program[2] and the NSF Blue-Ribbon Panel on Cyberinfrastructure report[3] stress the essential aspect of digital archiving of datasets to ensure long-term access.

Our project[4] – a collaboration of astronomers, a scholarly society, its publishing production partner, and research libraries – has the following goals:

- Provide long-term, persistent storage and access for cited datasets
- Explore a new model for data publishing (with libraries as digital annexes for journals)
- Simplify data publishing for individual astronomers
- Develop services to place processed data online
- Supply a catalyst/template for other disciplines and organizations
- Accelerate sociological change
- Recognize university libraries' key role in digital archiving and data curation
- Increase integrity of scientific publication

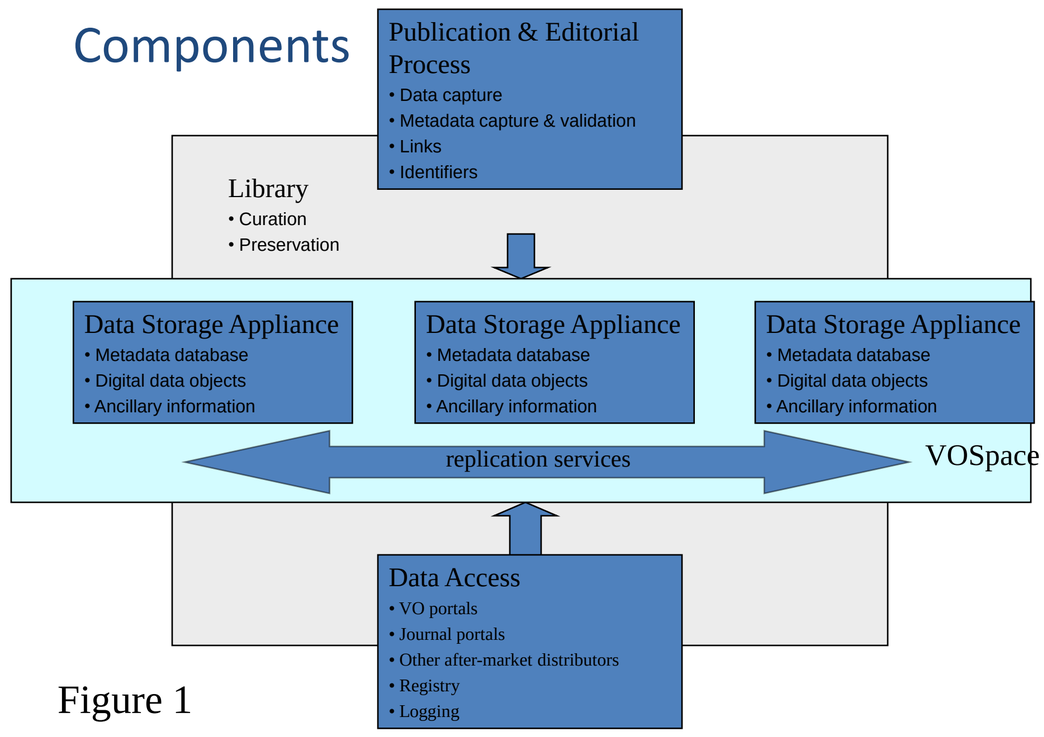
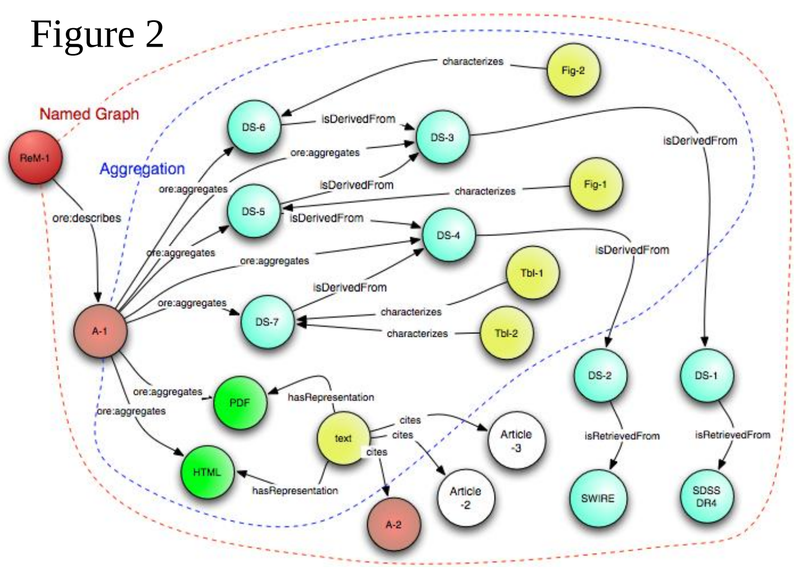


Figure 1

The components of the pilot system (Figure 1) reflect the variety of expertise that must be brought to bear:

- Domain experts understand their data and can develop appropriate standards and services for their community.
- Publishers continue to interact with authors and manage the editorial process. They can enact policies requiring data deposit and verify that standards are adhered to.
- Libraries are developing expertise with repositories and can ensure that data is available for the long term. They will be responsible for linking data to appropriate services.

Figure 2



One of the most interesting challenges of the project thus far has been the management of the article submission workflow. The challenges are several:

- Gather more metadata and datasets from authors without significantly increasing their workload.
- Simplify deposit process for authors and publishers.
- Enable article/dataset links without significant impact on publisher systems.

To accomplish these goals, we have chosen Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[5] and Simple Web-service Offering Repository Deposit (SWORD)[6] as enabling technologies.

Figure 2 shows a simplified model of an article using ORE concepts. The Resource Map ReM-1 in the upper left describes the article aggregation A-1 and captures relationships among resources both within and outside of that aggregation.

Figure 3 illustrates the workflow relationships among an author, the publisher, and one or more data archives and is described below.

- Author produces article

  The article is an aggregation of the document, the datasets underlying Figure 1 and Table 1. An ORE Resource Map (ReM), ReM-1, is created to describe this aggregation. We are working with Microsoft (Pablo Fernicola & Alex Wade) to integrate ReMs and tools for their manipulation into the Microsoft Office platform.

- The article aggregation is transmitted to the publisher

  Microsoft is integrating SWORD technology into the Office platform and there are standalone SWORD clients that could be adapted to use a ReM to configure the deposit. Several open source repository technologies, including DSpace, FEDORA, and E-Prints, are adopting SWORD for content deposit. That means there will be plenty of tool development that we needn't fund/manage directly.

- Publisher receives article

  The publisher receives the article, creates appropriate place holders in its own system, and generates a new ReM, ReM-2, which reflects this. This new ReM is made available to the data archive.

- Archive obtains datasets

  The archive uses ReM-2 to obtain the datasets associated with Figure 1 and Table 1. It stores these and any associated metadata, including links back to the article. The archive then creates a third ReM, ReM-3, reflecting the location of the new dataset objects, and makes that available to the publisher.

- Publisher links to datasets from article

  With the information in ReM-3, the publisher has a complete picture of the physical and conceptual elements of the article and, thus, can link them together.

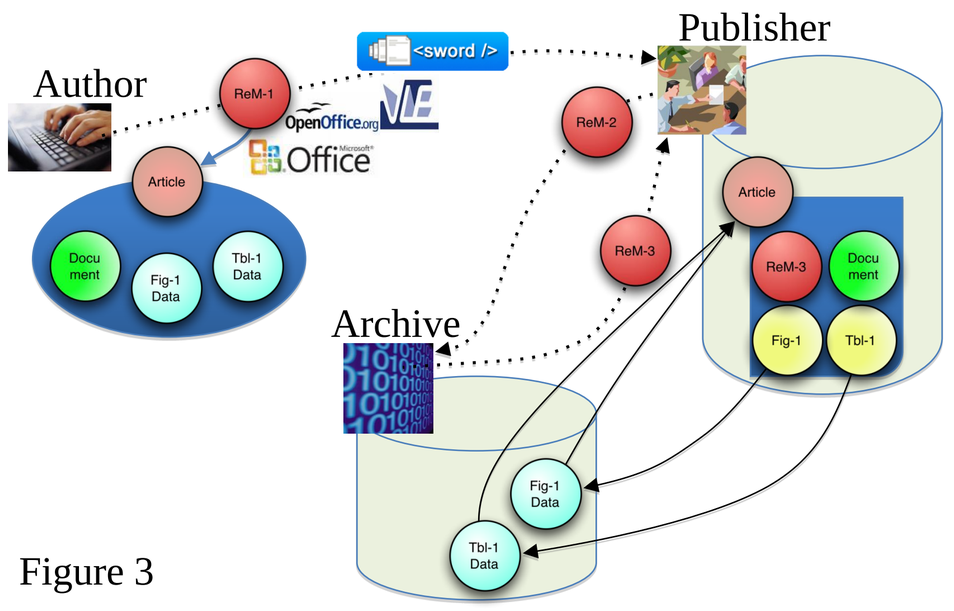


Figure 3

References

[1] Sloan Digital Sky Survey, http://www.sdss.org

[2] NDIIPP, http://www.digitalpreservation.gov

[3] National Science Foundation Cyberinfrastructure Report, http://www.nsf.gov/od/oci/reports/atkins.pdf

[4] Digital Data Preservation for Scholarly Publications in Astronomy, http://www.ijdc.net/index.php/ijdc/article/view/41

[5] OAI ORE, http://www.openarchives.org/ore/

[6] SWORD, http://www.ariadne.ac.uk/issue54/allinson-et-al/

JOHNS HOPKINS UNIVERSITY

External Research

Microsoft® Research

2009