ARTICLE EXCERPTED FROM:

INFORMATION STANDARDS QUARTERLY
SUMMER 2014 | VOL 26 | ISSUE 2 | ISSN 1041-0031

ISQ

TOPIC

OPEN ACCESS INFRASTRUCTURE

NISO
How the information world CONNECTS

Clifford Lynch

The Need for Research Data Inventories and the Vision for SHARE

CLIFFORD LYNCH

Disclaimer & Disclosure: I am a member of the SHARE steering group. SHARE's design is still actively evolving and undergoing prototyping and validation, and what I describe here are a mixture of my own ideas about SHARE and the broader enterprise of research data management, as well as fundamental functions that have already been adopted explicitly into plans for the SHARE system.

There is a major movement calling for public access[1] to the results of funded research, both in the US and globally. These results include both publications (most notably journal articles) and underlying observational or experimental data. In the US, the funders include federal agencies (where the White House Office of Science and Technology Policy is coordinating a government-wide effort to open up federally funded research), state governments, and private foundations.

In parallel with these developments has been a growing focus on the importance of research data management across all fields of scholarship. That is to say—essentially the idea that appropriate stewardship of data used in or arising from research is essential to preserving, communicating, and replicating scholarship and that, in fact, great opportunities exist to improve the pace and effectiveness of scholarly inquiry broadly if relevant data can be discovered, reused, recombined, and re-purposed in creative ways. Funders and disciplinary scholarly communities have also taken measures to advance these ideas.

With the broad adoption of these ideas, it has become clear that the research and higher education community needs to better understand and manage the research outputs that it produces. SHARE (SHared Access Research Ecosystem) is a joint project of the Association of Research Libraries (ARL) and the two key higher education presidential associations, the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU); ARL, with generous grant funding from the Alfred P. Sloan foundation and the US Institute for Museum and Library Services (IMLS), is leading the implementation effort. My own organization, the Coalition for Networked Information (CNI), with its deep expertise in both research data management and emerging developments in scholarly practice and scholarly communication, is also helping through its participation on the SHARE project steering group.

What I want to do here is to briefly summarize the potential role of SHARE in the overall scheme of managing research data, with some emphasis on the importance of standards (both existing and to be developed) for making this vision a reality. Note that there are parallel efforts within the SHARE development to address research publications, but I won't discuss those further here.

Most fundamentally, SHARE functions as an *inventory* of research data that is produced by scholars within the higher

CONTINUED »

SHARE

Great opportunities exist to improve the pace and effectiveness of scholarly inquiry broadly if relevant data can be discovered, reused, recombined, and re-purposed in creative ways.

education community. The system would ultimately include data elements such as what the data is; who created it and their affiliations; what organization and what program or grant funded its creation or capture (if any); where it is currently stored; who is funding the management of the data and how long that funding is guaranteed; and some notes on any access or use restrictions (e.g., embargoes, human subject constraints) that may apply to the data. Populating all of these data elements will require integration of a substantial number of different data sources, and in the early days of SHARE will be sparse; this will improve over time, as both the data gathering system and the sources it gathers from evolve.

SHARE is *not* itself a repository for data, but simply a place to record deposits and associated metadata. It is agnostic to the use of any specific repository and indeed seeks to span as many repositories as possible. These will include disciplinary, institutional, and funder-provided repository services.

Note that while this sounds simple, it is rife with scoping challenges that will need to be sorted out. Only a modest part of research data is "files" or "datasets" coming from individual investigators; often investigators *contribute* to very complex shared or pooled community scientific information systems (e.g., Genbank, the Protein Data Bank, the Astrophysics Data System, etc.) and how to reflect these contributions is unclear—as is how to reflect the ongoing stewardship of such data, which depends on the assurance of sustained support for these complex community data systems more broadly. There is also observational or cultural data that is collected and stewarded by a great assortment of entities (including research libraries on behalf of one or more scholarly communities), or that may even support a multiplicity of scholarly, commercial, and broader public uses: synoptic sky surveys, Web crawls, weather, geospatial and remote sensing data, and the Twitter archive. Projects and collaborations span institutional and national boundaries: scholarship is a global undertaking. Contributors or co-contributors of data include not just academics but government, research, and even commercial groups (consider the pooling of information now occurring between major drug companies and academic researchers, for example). Exactly what should be represented in the inventory?

In the SHARE architectural model, this inventory is stored in a component called the registry. The registry is "fed" by a series of services that make up the notification component, which gathers data from many sources and can also redistribute that data to other interested "subscribers" besides the registry. As data is fed from the notification system into the registry, efforts are made to normalize and consolidate data, which will be an ongoing challenge. It is very likely that there will be functions within the registry, as well, that try to continue to improve the quality of data normalization and consolidation.

Data picked up by the notification system can come either from external events occurring in environments that have been modified to post these events to SHARE, or from software that harvests metadata from the catalogs associated with existing repositories, for example. Events of interest might include the award of grants; the submission of progress reports to funders or achievement reports to host institutions; deposits of data to various repositories or scholarly information systems; the acceptance of a data management plan (hopefully with some of that data being in structured form that can allow the identification of intent to create and deposit data as part of a funded project—imagine building this into widely deployed tools like DMPTool); citation of deposited data in the literature; and reappraisal events and transfers of stewardship responsibilities. Clearly, the system relies upon a mass of standards (existing, under development, and/or as yet undefined) for harvesting, for structuring data, and for "vocabularies" for purposes like the identification

of organizations and funding sources. Simply enumerating relevant current standards and standards efforts would take an article longer than this one.

Personally, I am convinced that in the emerging world of international research data management, we are going to see more movement of data from one repository to another, and transfer of stewardship responsibility or funding sources to underwrite ongoing management—much more often than we are accustomed to as we have managed the traditional base of research publications. It is already common to make research data available for limited time periods through pre-funding built into grant budgets, setting up the need for periodic re-appraisal, and transfer of stewardship, though it is unclear who will conduct this or how it will be done. But the type of inventory envisioned as a core part of SHARE will be essential to managing these processes on a multi-disciplinary and multi-institutional large-scale basis.

Complementing the notification and registry components of the system are the discovery services; many of these services will simply incorporate data extracted from the registry into other discovery services within the research data management ecosystem. Because SHARE is so fundamentally and broadly multi-disciplinary in its coverage, I suspect that most researchers working in one or two specific disciplines will gravitate towards discovery tools (perhaps, for example, associated with specific disciplinary repositories or clusters of such repositories) that are optimized to understand the knowledge organization practices, ontologies, and vocabularies of specific disciplines. There will need to be at least some basic query interfaces to the registry itself, of course, to allow the most precise searching feasible on some structured data elements, such as funding sources.

A system like SHARE will be useful for many purposes. First and foremost, it will give researchers new tools to manage and reuse vital research data. It will help funders to understand the impact and outcome of their funding programs. It will help those responsible for the stewardship of scholarship to manage processes like reappraisal and transfer of stewardship. It will also provide visibility in the scale of current investment and future obligations related to the management of research results and outcomes, and help to clarify the rate of growth of these obligations.

| PI | doi: 10.3789/isqv26no2.2014.05

CLIFFORD LYNCH (clifford@cni.org) is Executive Director at the Coalition for Networked Information (CNI), an adjunct professor at Berkeley's School of Information, co-chair of the National Academies Board on Research Data and Information (BRDI), and a member of the Steering Committee of the SHared Access Research Ecosystem (SHARE).

My thanks to Elliott Shore and Eric Celeste for very helpful comments on an earlier draft of this; Diane Goldenberg-Hart helped immensely with the final version.

[1] At least in the United States, federal funders in particular have used the term "public access" rather than the related "open access" to describe their goals. The distinction and ambiguities here are important but beyond the scope of this short article. Note, as discussed later, that the SHARE system is agnostic to access limitations.

[2] Data registries similar to SHARE are under consideration in several other nations at present, and one urgent open question to be explored is how these systems should best interconnect or interoperate.

RELEVANT LINKS

Association of American Universities (AAU)
https://www.aau.edu/

Association of Public and Land-grant Universities (APLU)
http://www.aplu.org/

Association of Research Libraries (ARL)
http://www.arl.org/

Astrophysics Data System
http://adswww.harvard.edu/

Coalition for Networked Information (CNI)
http://www.cni.org

DMPTool
https://dmp.cdlib.org/

Genbank
http://www.ncbi.nlm.nih.gov/genbank

Protein Data Bank
http://www.rcsb.org/pdb/home/home.do
SHARE (SHared Access Research Ecosystem)
http://arl.org/share

White House Office of Science and Technology Policy (OSTP) Memorandum: Increasing Access to the Results of Federally Funded Scientific Research
http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf