Active and Social Curation: Keys to Data Service Sustainability

Jim Myers, Margaret Hedstrom, University of Michigan

Researchers need to be able to discover data in order to use it and hence, discussions of scalable data systems often focus on universal data catalogs as an early, and sometimes preeminent goal. However, such a characterization of requirements for data services can be problematic in that it ignores many other high value services that researchers need. Further, by placing too much emphasis on the requirements for discovery over those for scalable data publication and use, designers and developers risk significantly underestimating the costs and challenges involved in providing useful, scalable, and sustainable services. Most obviously, if discovery is an early emphasis, the system's value depends on the amount of data and metadata submitted as well as the system's usability relative to mature disciplinary repositories. This results in critical mass barriers as well as a catch-22 situation in which the costs of generating metadata must be borne long before any value is derived. Over-emphasizing discovery also perpetuates a 'broken market' in which costs are borne by data producers while value accrues to data users.

By recognizing the lifecycle of data and the processes it participates in, a number of opportunities emerge to minimize the barriers and delays to achieving beneficial use of data, to maximize the return on investment, and to enable market forces and self-interest to drive adoption of good data management practices, including metadata provision. At the most basic, taking a broader view of data services simply means recognizing the impact of the Amdahl's Law style issue that arises from making discovery faster and/or easier: the impediments to speeding up the end-to-end use of data and the obstacles to obtaining the most value from data, simply shift from the discovery phase to other parts of the data lifecycle. If the challenges of managing large complex data sets locally, assembling and publishing new data, capturing metadata, transforming and integrating data, etc. are not addressed, they will become the most pressing issues. At a deeper level, the idea of supporting the community-scale data lifecycle recognizes that data are used and managed by many different groups for many different purposes over time. Given that all of these groups must participate to achieve the ultimate goal of data services that support and drive science, it becomes important to look at the cost/benefit analysis of participation at the individual and group level. The common challenge today for data services to acquire sufficient metadata is, at its core, a problem of not providing enough value to metadata producers to motivate their full participation despite the general agreement that well documented data would be a community asset.

The SEAD project (http://sead-data.net/), supported through the NSF DataNet program, was founded to develop and deploy coordinated data services that address these lifecycle concerns as a way to achieve scalable and sustainable data service infrastructure. Over the course of two years, SEAD has worked to refine this general plan into actionable directions, a clear set of core

capabilities, and an initial set of robust deployed services focused initially on the sustainability (ecological and societal) research community. A primary motivator for SEAD was the recognition that modern computing technologies can remove many of the constraints that have influenced traditional curation practice, including strict requirements for the amount and quality of metadata. Data services built without such constraints, including the growing range of social media applications, are able incentivize data and metadata submissions in ways that traditional discovery catalogs do not. While SEAD's services are still being actively developed, it's architecture and conceptual approach to curation, as well as its specific capabilities and growing experience in supporting the long-tail of research, are highly relevant in building and sustaining a national service.

SEAD provides a federated storage and discovery mechanism, and can provide metadata to other data catalog services (e.g. DataOne, Thompson-Reuters), but the discovery portal is not the sole, or even primary interface through which researchers access SEAD service. SEAD provides secure project-level active content repositories where data and metadata can be incrementally added over the course of a research effort. Within a project's space, members can upload data in any format and add metadata using any vocabularies they wish (including tags). For known formats, interactive previews (maps, zoomable images, movie viewers, spreadsheets, temporal graphs) can be created and metadata within the files can be extracted for display on data pages and indexing for search. Local applications and third-party services can read/write data and metadata to/from the project space, making it possible to automatically capture provenance and instrument/model-derived metadata. When a given data collection is deemed ready for publication, it can be assessed for compliance with the policies of long-term archives, additional metadata can be added or inferred, and web services manage the transfer and packaging of a data copy for long term storage. Publication does not remove data from the project space and the team can continue to develop a new version of it. The collection is given a persistent identifier (DataCite DOI) and can also be associated automatically with online project and individual profiles and registered with third-party catalogs, all of which can drive traffic to a project-branded interface for data download and further faceted search within and across the project's collections. Data can be referenced across spaces and SEAD can track provenance and other relationships across it's federated holdings, which will enable researchers to discover who is using existing datasets and where derived data sets (subsets, alternate formats, modeled results, quality checks, papers, etc.) and social commentary may be found.

It is easiest to understand SEAD's feature set and direction in terms of the ideas of active and social curation. Active curation (AC) involves recording data and metadata as close to the source as practical and driving that acquisition through the deployment of capabilities that help data producers manage their research. Whereas uploading data at the end of the project is often seen as a burden, uploading it to a SEAD project space makes it easy to share within a distributed team and makes metadata immediately useful - visible when browsing shared data and usable for sorting/filtering/organizing data. AC replaces the model of requiring a big form for metadata submission to an incremental model where metadata can be added over time - a Flickr-style model where you can add a photo today and see it, and then come back and tag it,

add it to collections, etc. as you wish. This approach requires a flexible/'schema-less' repository and interfaces that can display data with as much, or as little metadata as currently exists. Overall, this approach changes the economics of metadata for producers - adding metadata is no longer an all or nothing affair and it is done out of self-interest rather than altruism.

Social Curation (SC) drives this economic analysis further, looking at ways that cross-group interactions can further motivate best practices. SC broadens the connection between private group spaces and public catalogs by enabling rapid data 'pre-print' capabilities, keeping project branding on data to drive peer recognition, and adding services that motivate the documentation of data connections to creators and the literature. In SEAD, specifying the creators of a data set and documenting the use of data in a paper results in automatic updates to online public profiles for the people involved and for the project. SC is a very rich area and encompasses many forms of social feedback - suggesting tags and metadata terms based on prevailing community use, providing reports that document the use and impact of data products, etc. SC also emphasizes the publication of comments, reviews, and derived data products by third parties and linking that information to primary data products, e.g. capturing the work to verify, validate, assess reference data assets that would not fit in todays reference data catalogs. Such capabilities can reduce the amount of duplicate work done across groups and provides additional visibility for such 'gray data' efforts. As with AC, SC influences underlying architecture - relying on persistent global identifiers to allow tracking of provenance and other data relationships through multiple systems, across groups, and over time.

Together these approaches, and the overall design principle of assuring that data services provide value across the community and throughout the lifecycle shift development priorities and dramatically shift how data services are perceived. They also change the dynamics of scaling - groups use the data services for their own purposes, ones that do not rely on the existence of a critical mass of data in national catalog, enabling the catalog to grow organically from day 1. Further, with a variety of incentives for the use of common data formats and rich metadata, the ability for these to be evolved by communities over time, and more emphasis on service endpoints and third-party extensibility, many costs, risks, and planning tasks can be reduced and distributed.

Fortunately, and somewhat by design, the core architecture proposed for NDS, providing a scalable store for 'arbitrary' data and metadata is the same regardless of whether a national catalog or active and social curation inspired services are emphasized. However, experience in SEAD has shown that it is easy to let preconceptions about specific use cases/interfaces influence repository design and service interfaces. It is therefore important for the NDS consortium to establish governance and design/development management processes that appropriately recognize the value of this broader lifecycle approach and assure that ongoing work across our data services community can be leveraged and incorporated.

Acknowledgements: The SEAD DataNet Team including Praveen Kumar (co-PI), University of Illinois Urbana-Champaign; Beth Plale (co-PI), Indiana University, and Kavitha Chandrasekar, Inna

Kouper, Robert McDonald, IU; Dharma Akmon, George Alter, University of Michigan; Mostafa Elag, UIUC; Rob Kooper, Jong Lee, Luigi Marini (NCSA), and other current and former team members.

NSF Award #OCI0940824