What constitutes success in a digital repository?

Kenneth Thibodeau
National Archives and Records Administration
8601 Adelphi Road
College Park, MD 20740-6001
Ken.Thibodeau@nara.gov

Workshop on "Digital Curation & Trusted Repositories: Seeking Success"

ABSTRACT

International collaborations have produced a standard describing the functions of a digital repository and the characteristics of one that can be trusted. These results provide an abstract frame of reference for evaluating such repositories, but meaningful evaluation requires that they be supplemented by empirical data on the purpose of each repository and the institutional, cultural and resource context in which it operates. Informed evaluation will consider how a repository balances the competing objectives of preservation and dissemination, whether it is defined primarily in terms of a community of producers or a community or users, and the extent to which it operates in isolation or in collaboration with other institutions**.**

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**] – *Value of information*,and H.3.7 [**Digital Libraries**]

General Terms

Management, Documentation, Performance, Reliability, Standardization.

Keywords

Repository, digital preservation, Open Archival Information System

1. INTRODUCTION

What constitutes success in a digital repository can only be addressed in context, specifically the context of the purpose the repository serves and of the environment in which it operates. No repository can be said to be truly successful in a meaningful sense unless it fulfills its purpose. Thus, criteria for success must be derived from its statement of purpose. Similarly, the metrics for estimating success against the criteria must be formulated in light of the culture, constraints and opportunities existing in the environment.

Conformance with the Open Archival Information System (OAIS) reference model and satisfaction of the attributes and responsibilities laid down in the RLG/OCLC report on trusted digital repositories may be preconditions for success, but they are not sufficient. One might assert that a basic purpose that all digital repositories must achieve can be derived from the OAIS standard's definition of an archival information system as "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community."[1] In this view, the implicit purpose of a digital repository is twofold: to preserve some information and to deliver products and services, derived from the preserved information, which satisfy the needs and/or desires of its designated user community. Following this train of thought, a statement of purpose which is specific to a given repository can be articulated by describing the nature of the information objects it preserves, the requirements for their preservation, along with the characteristics, needs, and desires of its designated user community. This approach is consistent with the RLG/OCLC report's description of a trusted digital repository as "one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future."[4] One might say, then that a digital repository is successful if it functions as an OAIS in a reliable and trustworthy manner.

However, taking this view into real contexts, there can be substantial variations in what information a repository should preserve, what is entailed by preservation, as well as in what

it means to serve the needs of a designated community; that there can be significant tensions between the goal of preserving some set of information objects and that of serving the needs of a given community; and that there can be major differences in the importance attributed to preservation versus service, perhaps to the point where one eclipses the other.

Libraries are typically organizations whose primary purpose is to deliver relevant information products – typically publications – to their users. There is a huge difference between the library service model and that of a repository whose primary purpose is to protect the interests of its owner. Companies in the pharmaceutical industry, for example, are making substantial investments to preserve electronic laboratory notebooks, but the purpose of this preservation is not – as it might be in a library – to enable future research in the history of science. Rather it is to enable the company to protect its intellectual property in potential patent litigation. The delivery service for such a repository is narrowly defined: to provide documentary evidence the company's lawyers can use to demonstrate that the company had made a certain discovery by a date certain.

Contrast this with purposes served by preserving digital information in other industries, aerospace, for example. In the aerospace and other manufacturing industries there is a long-term need for reuse of data for maintaining and adapting aircraft and other mechanical systems which are often kept in operation for many decades. The product data will be reused to manufacture replacement parts or to modify the aircraft to accommodate new components, such as new navigation instruments, new cargo handling equipment, or new passenger seats. In order to be reused, the data will have to be fed into engineering and manufacturing systems in the future. No one knows what such systems will be like 20 or 30 years from now; therefore, no one knows what data formats they will require. Satisfying the needs of this designated community requires a different model of preservation. This need cannot be satisfied – as may be the case with library books and pharmaceutical notebooks – by preserving specific documents in their original form. Keeping product data for its intended reuse requires the ability to transform it to some unknown future format, and the success of the transformation cannot be measured in terms of fidelity to the information object that was originally entrusted to the repository, but in whether it enables production of a replacement part which exactly duplicates the original one, or which enables successful adaptation of the aircraft[5].

A very different context is that of a government archives in a democracy. For institutions such as the National Archives and Records Administration, the legally designated community of users is anyone who has an interest in records of the U.S. Government or the information they contain, for whatever reason. This clientele is very diverse ranging from the government itself, through other governments at national and lower levels, through academic scholars, law firms, other businesses, TV and movie producers, to individuals doing family history, and in the case of electronic record even to individuals looking for records of their own life. NARA supported the development of the OAIS standard from the beginning, and required all companies who bid on the development of the Electronic Records Archives to conform to this model in their designs; however, the diversity of its user communities inhibits tailoring the system to the characteristics, such as knowledge level, of the designated user community, as stipulated by the OAIS standard. Government archives, thus, need to define their service model according to two criteria that are essentially agnostic of the characteristics of the 'community' of users; namely, the requirement for preserving and being able to provide authentic records and, within the constraints of this requirement, providing dissemination services to persons with a right of access.

The need to take into account the specific purposes and environments of individual repositories across a wide range of possibilities does not mean we cannot construct a general framework addressing success.

2. FRAMEWORK FOR EVALUATION

A framework for organizing information needed to evaluate the success of digital repositories can be articulated along three axes: orientation, coverage, and collaboration.

1.1 Orientation

One can define a spectrum of purposes ranging from prospective to retrospective. In a retrospective repository the emphasis is on preservation of assets while a repository with a prospective orientation will optimize the ability to satisfy demands of a user community. If the primary purpose is retrospective, to ensure that assets existent at any point in time are preserved intact for future times, then access to those aspects is a bi-product. Conversely, if the primary purpose is to support the needs and demands of a user community for information, long-term preservation of information assets may not be even necessary and at most it

will be instrumental to the primary purpose. Preservation, in the strict sense, may not be necessary where repurposing or adapting information to capitalize on opportunities offered by new technologies for delivery is important. Obviously, even in repositories which highly prioritize service to users, often there will be an instrumental need for preservation because: there can be no delivery of assets that no longer exist. In many cases, it will also be necessary to preserve information about the context in which information was originally created and used to enable appropriate interpretation of repurposed or derived products.

Criteria for success can be articulated relative to orientation. A repository with a retrospective orientation should be evaluated on how well it preserves the essential characteristics of its information assets. How well a retrospective repository servers potential users is likely to be of lesser importance. The interests of users might be better served by information brokers or value added services which act as intermediatories between the repository and the designated community of users. In contrast, evaluation of an institution with a prospective orientation should emphasize how well it satisfies the information needs and demands of its user community, with preservation being assessed primarily only insofar as it is necessary or instrumental to dissemination.

2.2 Coverage

A second dimension for evaluating success of a digital repository is how well it covers the universe of assets it should or might hold. Coverage should include both acquisition of assets and execution of functions against those assets. Specific criteria for successful acquisition depend on orientation. The universe of information objects that should be acquired by a retrospectively oriented institution may be defined with respect to the output of a designated producer community. Relevant considerations for evaluation include: what products of the producers are targeted for acquisition and what percentage of the targeted products does the repository acquire? In contrast, the target universe of assets for a prospectively oriented repository might not be so well defined. It could vary as a function of the interests of its designated user community. One should ask: how widely and how thoroughly a prospective repository searches for materials responsive to its users' interests and how effective it is in acquiring those assets; and whether its relationships to the sources of those assets are attuned to user need, and whether its processes for ingesting, storing, and managing the materials are optimal for their intended use.

2.3 Collaboration

In estimating the success of a repository, it is necessary to consider the environment in which it operates; in particular, whether it can achieve its purpose operating in isolation or whether it collaborates with other organizations in order to achieve success. This 'isolated' v. 'collaborative' spectrum defines a third category of criteria for evaluating success of a digital repository.

A repository may be said to operate in isolation if internally it fulfills all the functions described in the OAIS model from the receipt of Submission Information Packages (SIPs) to the export of Dissemination Information Packages (DIPs). A repository may be said to be collaborative if it relies on any other institution(s) to fulfill any of the functions assigned to an OAIS. However, one needs to distinguish arrangements where the repository contracts with an outside service bureau for one or more services from those which require true collaboration. A repository is responsible for proper management of any contracts. Therefore, contracts for services would fall towards the 'isolated' end of the collaborative spectrum. A more collaborative arrangement would exist where the separate institutions independently execute missions or pursue goals which they recognize as complementary and decide to work together to leverage each other's strengths.

Over the entire spectrum from isolation to collaboration, one needs to evaluate whether a repository recognizes and exploits possibilities for collaboration. Some repositories may be constrained legally or by their institutional setting from entering into collaborative relationships. In the face of such constraints, one could not fault a repository for not exploiting collaborative possibilities. Absent external constraints on its options, it is legitimate to assess whether a repository which operates in isolation could improve its performance through collaboration. Similarly, one should consider whether a repository which does engage in some collaboration might have other opportunities for improvement through additional collaborations.

For actual collaborations, a relevant criterion for success is whether a collaborative relationship actually improves the repository's performance over what it could achieve acting in isolation. Evaluation according to this criterion should take into account the nature of the collaboration. An example of a collaborative relationship is that between the Florida Center for Library Automation (FCLA) and its client libraries in pubically funded colleges and universities in Florida. FLCA provides the repository for digital versions of library-owned

collections. It relieves individual schools' need to preserve the digital materials, but does not provide direct services to library users in any of these schools[2]. A success model based on the OAIS standard would ask how well a repository fulfilled all OAIS functions. However, this would be inappropriate for institutions participating in the FLCA program. One should not, for example, fault the FCLA for not providing any direct services to library users. In effect, the FCLA and its client libraries have split OAIS functions, with FCLA providing digital preservation and the libraries providing service to end users.

This arrangement is asymmetric, but there are other permutations that could be described as collaborative, including peer-to-peer relationships, collaboration on voluntary standards for repository functions, sharing of best practices and lessons learned, and perhaps agreements on coverage.

Te LOCKSS (Lots Of Copies Keep Stuff Safe) collaboration, exemplifies a peer-to-peer relationship. Each library in this collaboration performs the full range of functions needed in a digital repository, or at least any variations in functions are at the discretion of each institution, with the sole exception of preservation of digital materials. For preservation, each library is responsible for maintaining the authenticity, integrity and availability of its digital collection so that, if needed, other partners can obtains copies.3]

The criterion of whether collaboration improves performance can be applied to both the FLCA and LOCKSS cases, but the application needs to be tailored to the context. In Florida, the question for each participating library is whether the centralization of repository and preservation services in FLCA reduces its costs and/or provides it will access to greater or better resources for these functions than it could achieve in isolation. In the LOCKSS alliance, use of the approach probably entails additional expenses over operating in isolation, because each member needs to acquire and operate a LOCKSS appliance, which enables coordination and sharing of copies on an as-needed basis. The question for a library participating in LOCKSS is whether the possibility of recovering lost or damaged items from other partners is worth the additional investment.

3. CONCLUSION

Abstract models of what a digital repository should be, what functions it should fulfill, and whether it merits trust need to be supplemented by empirical consideration. We need to situate digital repositories along several axes:

1. Orientation: does a repository emphasize preservation of assets or the satisfaction of the demands of a user community?

2. Coverage: does a repository aim primarily to preserve all or at least the noteworthy products of a given producer or set of producers or to build a collection best suited to the needs of its designated user community, regardless of source?

3. Collaboration: does a repository operate in isolation or collaborate with other organizations in order to achieve success?

Answers to these questions describe the space in which a repository operates, and allow us to contextualize criteria for evaluating how well a repository achieves its objectives, given its resources and constraints.

REFERENCES

[1] Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS)* (July 2001) www.ccsds.org/documents/pdf/CCSDS-650.0-R-5.

[2] Florida Center for Library Automation. www.fcla.edu/FCLAinfo/aboutinfo.html.

[3] Maniatis, P., Roussopoulos, M., Giuli, TJ , Rosenthal, D., Baker, M., and Muliadi, Y. "LOCKSS: A Peer-to-Peer Digital Preservation System", *ACM Transactions on Computer Systems*, Vol. 23, No. 1, February 2005, 2–50. www.eecs.harvard.edu/~mema/publications/TOCS2005.pdf

[4] Research Libraries Group. *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report.* RLG. Mountain View, CA. May 2002. www.rlg.org/en/pdfs/repositories.pdf

[5] Project Group "LOTAR." *White Paper for Long Term Archiving and Retrieval of Product Data within the Aerospace Industry (LOTAR). Technical Aspects of an approach for application.* Version 1.0. 2002.. www.prostep.org/file/17291.WP_LOTAR.