Repository replication. Specification for possible further work presented to the Aquifer partnership for consideration at its meeting on October 25, 2004

JP Wilkin et al
11/12/2004

At the August 5th meeting of the Aquifer initiative, there was discussion of the importance of digital asset management, persistence, shared repository development, and data replication. Greenstein reported that "Persistent digital asset management is likely to be key to Aquifer's success. How can Aquifer partners leverage their extensive respective experiences with digital object repositories in order to put into place a network of robust and appropriately replicated digital filestores?" John Wilkin was charged with convening a group to scope an initiative in this area and to make recommendations to a group convening at the next meeting for possible future work. The group's focus would be in two areas: (1) data replication, and (2) repository co-development possibilities.

Wilkin initiated a number of brief exchanges on e-mail to enumerate a list of issues, benefits, and opportunities in each of these two areas. The list was expanded by suggestions received in e-mail and in individual phone discussions. Subsequently, each issue or benefit was annotated with a statement commenting either on possible conclusions we might reach (with regard to issues) or the possible validity of the benefit. Additionally, each issue or benefit was annotated with a suggested strategy for further exploration. Although time did not allow for an extensive discussion or amendment of this list, we did feel that it should be put forward at the October 25th meeting for consideration as a framework that may serve as the basis for subsequent work.

Scope a piece of work that will help us explore prospects for data replication.

1. What issues or problems do we need to address?
    a. Defining (and understanding, mutually) data formats
    Possible conclusions
    It is almost certain that content replication would need to be enabled by unambiguously and commonly understood content. To this end, the effort would need to leverage the work of the DLF Global Digital Format Registry in an effort to create an architecture within which the format of content could be defined and declared to a repository manager.
    Strategies for exploration

    b. Expression of rights and responsibilities
    Possible conclusions
    This is a difficult but tractable problem. Rights expression languages such as that within MPEG-21 provide a framework to be able to provide the necessary articulations of rights and responsibilities. Of particular need would be a framework that brings together not just copyright information, but information on responsibility for publication as well as basic authN/authZ information. Some typical examples that the initiative would need to take into account are:

- Preservation surrogates: we own them collectively, and we handle issues of display and derivative rights collectively. Each institution can build specialized interfaces to subsets if it chooses, but ultimately the item itself should be delivered to the user from the repository.
- New publications without access restrictions: While, on the surface of it, these may seem similar to the Preservation surrogates (above), there is a strong sense of ownership and identity with these publications. Re-use should be facilitated by the architecture, though perhaps with constraints, but the actual "presentation" of the corpus should be off-limits to anyone but a responsible agent.
- New publications with access restrictions: A complement to the above, with the added need to perform authN/authZ.
- Commercially-produced and co-owned converted materials: Having these materials be in the repository would be important, and access is an important facet. It needs to be limited to subscriber institutions. Otherwise, this one probably looks like the "Preservation surrogates."

Strategies for exploration

c. Management of rights and responsibilities.

Possible conclusions

This is a tractable issue, and is necessarily an architecture piece driven by the rights and responsibilities (i.e., 1b, above) that are expressed by the cooperating partners. Without 1a, this work would be largely speculative and could not move forward

Strategies for exploration

d. Economics

Possible conclusions

We can only speculate on this issue, but we do feel that although the cost of the whole would be greater than anything invested in by an individual institution, the cost would not be greater that the sum invested by a small collective (three or more?) institutions. The benefits (e.g., broad geographic replication) also exceed what is currently being undertaken by an individual institution, thus making the economics even more compelling.

Strategies for exploration

e. Governance

Possible conclusions

The operational and legal responsibilities of a repository that participates in replication are significant, and the processes of defining and enforcing responsible curation are probably the greatest challenges of replication. No possible conclusions are presented here: our ability to find and agree to an effective model may be the Achilles heel of this concept, but we should nevertheless strive to find that model.

Strategies for exploration
A group of interested parties …?

Unaddressed issues to be included?

1. How do you "use" it? (e.g., DLXS)
2. What form does it take? (e.g., database, container, API)
3. What models can we borrow from? (e.g., GRID)

2. What compelling advantages are there to replication?

a. More cost-effective

Validity
Our sense is that there is a good enough possibility that doing something collectively can be more cost-effective, assuming that the collective effort serves the needs of individual institutions, that we should explore this by developing cost models where local investments shift to a shared effort. The Sakai initiative may provide some strong parallels.
Strategies for exploration

b. Greater data viability

Validity
Among all of the questions before us, this one seems most likely to be true: replication of data will lead to greater data viability, addressing some critical preservation questions before us. No further work on this question needs to be done, but the principle needs to be re-articulated at every turn as we explore the advantages and disadvantages of replication.

Strategies for exploration

c. Better access (skirting interoperability?)—helps to facilitate developing the sorts of tools and services we imagine *when and if* the rights and responsibilities permit it.

Validity
Whether or not data replication will improve *access* is probably an open question. It may help by giving tools more immediate access to data, and it may also lead to a cacophonous proliferation of representations of the same

data. Nevertheless, we feel that exploration through several "use cases" or studies would be helpful.

Strategies for exploration

Evaluate options for repository co-development effort.

1. Are there candidate systems upon which we can build?

   Possible conclusions
   There are a number of viable candidate systems within which a repository co-development effort could take place. Obvious candidates include DSpace (and the emerging 2.0 architecture), Fedora, and the recently released CDL preservation repository, each a multi-institutional effort with different advantages and strengths. It may also be the case that the Sakai framework, essentially a portal architecture that has been directed first at a general-purpose Collaboration and Learning Environment (CLE), could be used for as a framework for developing a preservation repository. Active discussions are taking place in the Sakai Educational Partners Program (SEPP) Library Discussion Group around the relationship between Sakai and repositories such as Fedora and DSpace.

   Strategies for exploration

2. If not, does this argue against co-development?

   Not applicable.
3. If co-development is to be pursued, what sorts of models for collaboration are available to us, and what are pros and cons?

   Possible conclusions
   A number of models have virtues for an effort such as this, and should be seriously considered if a coordinated initiative is to go forward. These include:
   a. "Tight" federation of developing institutions with central governance.
   b. "Loose" federation of developing institutions with consultation.
   c. Single lead institution with shared development space.
   d. Independently established non-profit.

   Strategies for exploration