CITATION

Creative Commons License: Attribution-NoDerivatives 4.0 International (CC BYND 4.0)
QUESTION

Since 1996, the digital preservation community has been emerging as evidenced by the increasing number of formalized standards, conferences, publishing options, and discussion venues.

What are the most significant community developments and why? What gaps remain in terms of community standards and practice? What roles should/could academic programs, professional associations, curatorial organizations, and individual researchers and practitioners play in those developments? What priorities and desired outcomes should there be for building the community’s literature?

RESPONSE

Ascertaining the most significant community developments is a task likely to cause a few religious wars amongst long-time digital preservationists. However, the following are the most significant events in this author’s humble opinion.

1. The realization by the overall Computer Science (CS) and Information & Library Science (ILS), among other domains and industries, that there is a digital preservation problem in the first place. This realization took place among individuals and organizations over the course of several decades, from the 1960s to the early 1990s. This is important because you cannot fix a problem if you don’t know you have one.

2. The 1996 Waters (et al.) report on the digital preservation problem. This report was a significant event because it outlined the problem(s) and what steps needed to be taken to ameliorate it.

3. The development of the OAIS Reference Model (RM) by the Consultative Committee on Space Data Systems (CCSDS) and the standardization of the model in 2002. The development of the OAIS RM is important because it the committee that created it consisted of, and was informed by, practitioners and users of data beyond the Space Data Community. It defined common preservation terms that would mean the same to all who used them (or, “should”). And, finally, the OAIS RM defined a common preservation repository standard against which digital repository managers could compare their own systems to determine, at least subjectively, their own preservation-worthiness.

4. The creation of the Digital Curation Centre (DCC) in the UK in the 2000s. Although the DCC is designed to serve UK Higher Education Institutions (HEIs), it has provided a central location for digital preservation practitioners to go to for information related to digital preservation. The
centre has also provided a platform from which further research and standardization in digital curation and preservation may continue.

5. The creation of the National Digital Infrastructure Preservation Program (NDIPP) in the United States in the 2000s. Much like the DCC in the UK, NDIPP had provided a central location in the USA from which digital preservation research and development, and the application of it, is promoted. As well, the program has provided an avenue through which private industry, government, research, and academia may come together to address the common problem of digital preservation. The Science and Technology committee of the Academy of Motion Picture Arts & Sciences (AMPAS) mentions NDIPP in their report, “The Digital Dilemma”, (2007) as being an important program with which private industry should be involved in order to coordinate resources to solve a problem (digital preservation) that all industries are facing.

6. The publication of the AMPAS report (2007) on the digital preservation problem within the movie industry. The movie industry’s products and libraries represent a large source of profit for the industry, as well as the cultural heritage of the respective countries that produce movies. The AMPAS report about the digital preservation problem is important because: (1) it meant that a major, high dollar industry was also seeking solutions to the digital preservation problem. This made it “not just a library issue”, and provided additional clout (financial and political) to the task of finding solutions. (2) The authors of the AMPAS report clearly stated that the digital preservation problem was not just a movie industry problem, it was everyone’s problem who used digital data, thus the solutions must be found by working together across private industry, government, academia, and other research institutions. (3) Reflecting the work done in ILS, the industry stated that digital preservation costs were far higher (1100% more) than non-digital preservation. This is the only non-research, non-academic report this author has read that shows the costs as determined by private industry. The authors of the report stated that standards will reduce costs, and the movie industry should resist implementing one-off solutions. This promotes the use of standards as an integral part of the digital preservation problem, even within a high-profit commercial industry.

7. The development of the concept of a “Trusted Digital Repository”, as well as the mechanisms to audit and certify that a repository is actually “trustworthy”. This includes the development of TRAC (“Trusted Repository Audit and Certification”), DRAMBORA (quantitative self-assessment of a repositories trustworthiness), other assessment criteria developed in Europe, and the development of TRAC into an ISO standard via the CCSDS called, “the Audit and Certification of Trusted Digital Repositories”. This development also includes the development of
standards with which to certify the certifiers. The significance of the development of an ISO standard for a “trusted digital repository” is that it gives practitioners and other repository managers a base set of policies from which they can build or assess their repository’s ability to survive over the indefinite long-term, especially when use in conjunction with the OAIS RM.

8. The development of outlets for publication, forums for discussion, web sites, with information on preservation, etc., has given practitioners and researchers avenues for their work that can be used for their own professional advancement. Providing incentives for researchers and practitioners to do preservation work is one way to ensure the necessary preservation is done. It also provides an iterative feedback loop, such that researchers and practitioners can adapt policies and standards as new information and research become available.

Some gaps do remain in terms of community standards and practice. The gaps are in some instances managerial, others are more technical.

1. The standards for preservation policies and repository design, such as “the Audit and Certification of a Trusted Digital Repository” and the OAIS RM are designed with large organizations in mind. What if you aren’t NASA or the Library of Congress? For example, what if you are the lone digital archivist for the Harley Davidson archive? Or a digital library on quilts? The standards outlined for preservation policies and repositories are stacked in favor of large organizations with large bureaucracies. The ILS community ought to develop a “lite” version aimed at small “mom and pop” repositories whose administrators curate important material but don’t need all of the overhead presented in the ISO standards for trusted digital repositories and the OAIS RM.

2. The same idea applies for data management training for researchers and other administrators of data archives in non-ILS domains. These researchers don’t want to spend their time on the full curation of the data, but neither are many of them likely to turn the data sets over to libraries and archives for stewardship in the near term. (The long-term is another issue.) Yet, in order to support science and the requirements of funding agencies such as the NSF and the NIH, the data must be preserved and shareable. The development of a “lite” curriculum for data management, aimed at non-ILS data managers, would be useful for non-ILS data managers, and would strengthen librarian and archivists’ roles as information managers by providing a consulting and outreach function to scientists and researchers.
3. The designation of certification as “trustworthy” does not seem to take into account “local” rules and regulations that a repository may have to take into account when designing the preservation system and the policies that must be applied to it. Some repositories may have to forgo international preservation standards in lieu of following national, state, county, or other regulations. Does that mean the repository is not “trustworthy”? Will the repository now be considered “2nd class” because it isn’t certified as “trustworthy”? Thibodeau has discussed looking at each repository on a case-by-base basis.

4. Preservation policy and repository standards have been designed from the top-down. Granted, the people designing the standards have worked with repositories themselves (usually) and so based the standards on their own experiences with the running of a repository. However, it is one thing to define standards, it is another to ensure their implementation. For example, this author’s masters paper work (2002) involved studying 100 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) data providers, to determine which Dublin Core elements were used or not used. At the time, practitioners were arguing for and against qualifying DC to be more detailed, per other metadata standards (Lagoze, c. 2001). Practitioners have certain standards for metadata quality. No one had actually examined how people were using DC. This author found that out of 15 DC elements, only 3 (title, author, & date) were being used the majority of the time. A separate but related study by Dushay & Hillmann around 2003 with the National Science Digital Library found the quality of the metadata content was abysmal. Follow up studies by Shreeves, et al. in the mid-2000s examined metadata quality in the OAI-PMH and also found it to be abysmal. These studies combined made the religious war over qualifying or not qualifying DC moot. Why qualify DC if only 3 elements out of 15 are being used the majority of the time, and the quality of the metadata content in those elements is abysmal? Perhaps the quality of the metadata content should be improved, then more elements should be used, and then practitioners can worry about qualifying DC? The same discrepancy may not exist for preservation policies, but one might think it would be interesting to find out what people are actually doing, as opposed to what they say or think they are doing with regards to compliance with standards. Then again, that is the purpose of auditing and certifying trusted digital repositories, so one may consider this argument circular!

5. Further examination as to what digital preservation is going to cost. If material must be curated from its birth, then it is also be true that decisions will have to be made early on as to whether or not material should be preserved at all. Even AMPAS noted that the movie industry
must change its mindset from “save everything”, which worked fine with film, and must now curate the digital move data.

6. A large gap in digital preservation is the transfer of data from one system to another, whether external or internal. The development of standard ingest tools would help reduce the costs of preservation. The Producer-Archive Interface Method Standard (PAIMAS) by the CCSDS (c. 2003-2004) has been one step in this direction, but it only outlines a method. A technically simple way to transfer data between repositories is a solution in need of an answer.

7. Metadata is one large gap still in need of a solution. The problem relates to both metadata quality (mentioned earlier) and tools with which to create metadata. Scientists and researchers who work with data have repeatedly stated, both in readings by others and in this author’s work on the DataNet project, that the “killer app” is a tool which helps them to appropriately and simply annotate their data. Like ingest, metadata is a challenging problem in search of an answer.

The roles that academic programs, professional associations, curatorial organizations, and individual researchers and practitioners play in these developments as a function are that individuals must identify the gaps. As a community, individuals must agree on those gaps, and then apply to their organizations for the time to work on the problems, or else, obtain grants to that they may work on the solutions. Eventually, the solutions may be taught as part of the ILS and CS curriculums.

In terms of the gaps identified above, researchers and practitioners should continue to work on the metadata and ingest problems, realizing that these are two huge gaps to preserving materials that cross all domains.

Some other areas organizations and individuals may play involve deciding what are the penalties if a repository is not an “OAIS RM” “TDR”? What are the rewards for being “trustworthy”? Should there be rewards or penalties? If so,
what? For example, there is a “charity navigator” that provides certain criteria against which a possible donator may determine whether or not they wish to give money. It does prevent people from giving money to organizations who, say, waste a lot of money on administration. But it is also true that some smaller organizations may not have the money to re-fit themselves to meet Charity Navigator’s criteria. Does this mean that they are no less worthy of donations, or that the money will do to waste? Not necessarily. It may mean that a charity may receives less money than they would otherwise, because Charity Navigator gives them a lower rating than an organization with more funding. This in turn gives more funding to the charity that already receives more funding, and less funding to a charity that already has less. If an organization does not have the time and resources to self-assess or receive certification as “trustworthy”, will they encounter any penalties, whether implicit or explicit?

This implies that standards should be a guide, and viewed as one part of a whole package.

One role individuals and organizations in the preservation field might play is that of consultant to small organizations that manage data and to individual researchers. One output of this could be an OAIS RM “lite” and an “audit and certification for trusted digital repositories” “lite”, aimed at repository managers who do not work for large bureaucratic organizations. This could be a document or standard or online training that a practitioner could do on his or her own time.

Currently, even the DRAMBORA self-assessment requires a large time
commitment from at least one, if not more, repository administrators. Part of this consultant role would involve education graduate students and researchers on data management. One output of this could be an online certification program that scientists and researchers could do on their own time, as they have time, on how to manage their data. This is slightly different from, but related to, personal information management. The data in this sense would be data gathered in the course of one’s work, not personal data, such as a digital photo album. This would include learning how to tag metadata, and thus begin to fix this problem where it starts – with the data creator.

Some possible outputs for the above problems are librarians and archivists continuing to provide consulting and outreach to scientists and researchers regarding preservation standards. The creation of “lite” versions of preservation policy standards and repository designs would be helpful for small repository administrators and those whose local standards might supersede international preservation standards. Technology is not an ILS strength, information management is ILS’s strength. ILS practitioners and researchers must continue to work with CS and other technical folks on developing ingest and metadata tools, especially with preservation in mind. These tools may also need to be designed with individual researchers and small repository administrators in mind. ILS practitioners and researchers must build upon their strength in information management, and not cede ground to CS, if they wish to remain relevant.
One outcome of a consulting role within other domains regarding preservation, and providing tools that aid in preservation, is to raise standards for data preservation within other communities. This will make the long-term preservation of that data and information easier for those who must eventually manage it. And, thus, make it more likely that the data will be preserved at all. This will also strengthen ILS as a field.