Workshop Report: Preservation at Scale

Digital preservation practitioners from Portico and from the National Library of the Netherlands (KB) organized a workshop on "Preservation at Scale" as part of iPres 2013. This workshop aimed to articulate and, if possible, to address the practical problems institutions encounter as they collect, curate, preserve, and make content accessible at Internet scale.

Preservation at scale has entailed continual development of new infrastructure. In addition to preservation of digital documents and publications, data archives are collecting a vast amount of content which must be ingested, stored and preserved. Whether we have to deal with nuclear physics materials, social science datasets, audio and video content, or e-books and e-journals, the amount of data to be preserved is growing at a tremendous pace.

The presenters at this workshop each spoke from the experience of organizations in the digital preservation space that are wrestling with the issues introduced by large scale preservation. Each of these organizations has experienced annual increases in throughput of content, which they have had to meet, not just with technical adaptations (increases in hardware and software processing power), but often also with organizational re-definition, along with new organizational structures, processes, training, and staff development.

Over thirty people from seventeen countries attended this workshop. The speakers were:

1. Marcel Ras, Program Manager International e-Depot, and Caroline van Wijk, Project Manager, Koninklijke Bibliotheek: "Balancing between technique, organization, roles and long-term strategies"
2. Tobias Steinke, Project Manager Deutsche Nationalbibliothek: "Ingest levels for handling preservation of different object types in a large collection"
3. Maureen Pennock, Head of Digital Preservation, The British Library: "Organising preservation at scale: The British Library's Digital Preservation Strategy (2013 – 2016)"
4. David Rosenthal, Chief Scientist, LOCKSS Program, Stanford University Libraries: "Economics and operational issues in distributed preservation at scale"
5. Ross King, Chairman of the Board of the Open Planets Foundation: "The SCAPE Project"
6. Andrea Goethals, Manager of Digital Preservation and Repository Services, Harvard Library: "Challenges and lessons learned in migrating an entire repository"
7. Vinay Cheruku, Director of Content Management Systems, Portico & JSTOR: "Scaling Up – issues and solutions"
8. Tim Dilauro, IT Architect, Digital Research & Curation Center, The Johns Hopkins University: "Lessons learned at the Data Conservancy and Johns Hopkins University Library"
9. Lars Bjørnshauge, Managing Director at DOAJ: "Challenges in preserving OA content and the long tail of very small publishers"
10. Zhenxin Wu, National Science Library, Chinese Academy of Sciences, "Digital Preservation Center of NSLC"
11. Peter Burnhill, Director of Edina: "Using The Keepers Registry To Assist Collaboration Between Keepers"

Boats and three-legged stools were recurring leitmotifs in the presentations and ensuing discussions. The preservation community will require bigger–and more stable–boats to ride the incoming tsunami of content. Acquiring and preserving more and more content will necessitate curation, access, and reuse both at larger and at a finer levels of granularity. It will also require significant collaboration (our three-legged stools) within and across organizations.

There were a number of broad categories addressed by the workshop speakers and participants:

1. Technological adaptations
2. Institutional adaptations
3. Quality assurance at scale and across scale
4. The scale of the long tail
5. Economies and diseconomies of scale

Technological Adaptations

Many of the organizations represented at this workshop have gone through one or more cycles of technological expansion, adaptation, and platform migration to manage the current scale of incoming content, to take advantage of new advances in both hardware and software, or to respond to changes in institutional policy with respect to commercial vendors or suppliers.

These include both optimizations and large-scale platform migrations at the Koninklijke Bibliotheek, Harvard University Library, the Data Conservancy at Johns Hopkins University, and Portico, as well as the development by the SCAPE project of frameworks, tools and test beds for implementing computing-intensive digital preservation processes such as the large-scale ingestion, characterization, and migration of large (multi-terabyte) and complex data sets.

A common challenge was reaching the limits of previous-generation architectures (whether those limits are those of capacity or of the capability to handle new digital object types), with the consequent need to make large-scale migrations both of content and of metadata. These system changes in turn resulted in organizational changes in institutions. Other shared challenges included the optimizations required for data transfer and data storage at scale–data that are variously delivered in large and small packages of both few and many files.

A distinction that emerged in discussion is that between optimization to increase throughput, and scalability (if you have to double throughput tomorrow, you could do it by doubling your same hardware running the same software). There was much discussion of the interplay of such linearly scalable technologies, along with other software engineering best practices such as modular design, data-driven configuration of workflows, and clean application programming interfaces, against evolving requirements for an institution's preservation repository.

Institutional Adaptations

For many of the institutions represented at this workshop, the increasing scale of digital collections has resulted in fundamental changes to those institutions themselves, including

changes to an institution's own definition of its mission and core activities. For these institutions, a difference in degree has meant a difference in kind.

For example, the Koninklijke Bibliotheek, the British Library, and Harvard University Library have all made digital preservation a library level mandate. This shift -- from relegating the preservation of digital content to an organizational sub-unit, to ensuring that digital preservation is an organization-wide endeavor -- is challenging, as it requires changing the mindsets of many in each organization. It has meant making choices and reallocation of resources from other activities, recognizing that the organization cannot do everything. It has necessitated strategic planning and budgeting for long-term sustainability of digital assets, including digital preservation tools and frameworks – a fundamental shift from one-time, project-based funding. It has meant comprehensive review of organizational structures and procedures, and has entailed equally comprehensive training and development of new skill sets for new functions.

Scaling up has resulted as well in a growing recognition of the necessity for working across and among, not just within, institutions. It will be increasingly necessary to engage others -- national libraries, funding agencies, governmental units, federations like LIBER and IFLA, publishers -- in taking up their responsibility within the life-cycle of digital objects.

Quality Assurance at Scale and Across Scales

A challenge to scaling up the acquisition and ingest of content is the necessity for quality assurance of that content. Often institutions are far downstream from the creators of content. This can mean data transfer quality issues, particularly for high-volume transfers, such as complexity in synchronizing fixity checks and detecting damage in transit. It can mean the lack of provenance information, including such technical information as which version of what software produced a dataset. It can mean semantic deficiencies, such as an inability to interpret data sets or cells in an Excel spreadsheet. It can mean the inability to correlate content across repositories, as when contributors to the Keepers registry use differently formatted ISSN information.

Institutional quality assurance policies can effectively bottleneck even systems built for large-scale throughput, both because some automated QA tools do not scale as effectively as the workflows in which they are embedded, and because quality defects often require non-automated (human) intervention for resolution. Thus there was much discussion of how institutions define just what is "good enough," and how those decisions are reflected in the architecture of their systems. Some organizations have decided to compromise on ingest requirements as they have scaled up, while other organizations have remained quite strict about the cleanliness of content entering their archives. As the amount of unpreserved digital content continues to grow, this question of "what is sufficient" will persist as a challenge, as will the challenge of moving QA capabilities further upstream, closer to the actual producers of data.

The Scale of the Long Tail
As more and more content is both digitized and born digital, institutions are finding they must scale for increases in both resource access requests and expectations for completeness of collections.

The number of journals in the world that are not preserved was a recurrent theme highlighted by Keepers, DOAJ, and LOCKSS. The exact number of journals that are not being preserved is unknown, but some observations include:

- Of the 100,000 serials with an ISSN, 79% is not known to be preserved anywhere. It is not known how many of the serials that do not have ISSNs are being preserved.
- Of the full-text OpenURL requests fed through Edina, 85% refers to content that is being preserved by fewer than 3 "Keepers."
- In 2012, Cornell and Columbia University Libraries (2CUL) estimated that about 85% of e-serial content in their libraries is unpreserved.

This digital "dark matter" is dwarfed in scope by existing and anticipated scientific and other research data, including that generated by sensor networks and by rich multimedia content. This scale of uncollected data, along with QA issues described above, stimulated discussions of various choices across the collecting versus preserving spectrum.

Economies and Diseconomies of Scale
Perhaps the most important question raised at this workshop was the question as to whether we as a community are really at scale yet? Can we yet leverage true economies of scale? David Rosenthal noted that as we centralize more and more preserved content in fewer hands, we will be able to better leverage economies of scale, but we will also be increasing risk of a single point of failure.

Next Steps
The consensus of the group seemed to be that, as a whole, the digital preservation community is not yet truly at scale. However, the organizations in the room have moved beyond a project mentality and into a service oriented mentality, and are actively seeking ways to avoid wasteful duplication of effort, and to engage in active cooperation and collaboration, including providing fora in which organizations can thoughtfully and frankly discuss shared challenges, capabilities, successes, and failures.

Some possible further steps for this group mentioned at the workshop included:

- Continued discussions of a possible registry of APIs.
- Identifying ways to move forward with the unpreserved e-journals.
- Projecting the BL three-legged stool of strategy, authority, and communication from inside one organization to across many institutions

Participants who responded to a poll after the workshop in general found the workshop very useful, and the presentations and discussion very informative. Topics that responders considered would have benefited from more extensive discussion included hearing from those

with experience handling billions of files, and experience with commercial solutions for large-scale preservation. Respondents also suggested the need for more information about archives specializing in web or audio/visual archiving, as well as companies grappling with preservation of their own material. Other suggestions included a dual-track discussion, with one track focused on technical issues, and the other focused on organizational issues (including succession planning, rights management, and best practices) of archiving at scale

Note: The notes from the discussion of each presentation are available at: https://drive.google.com/folderview?id=0B1X7I2IVBtwzcGVhWUF0TmJIUms&usp=sharing.