THE PORTICO STRAIGHT-TO-INGEST PROJECT

Sheila Morrissey
Portico
USA
sheila.morrissey@ithaka.org
https://orcid.org/0000-0003-2069-121

Amy Kirchhoff
Portico
USA
amy.kirchhoff@ithaka.org
https://orcid.org/0000-0002-3136-142

Abstract – This paper describes a soon-to-be-concluded six-month project at Portico. The goal of this project is to attack those cost drivers with automated, data-driven analysis of problematic content, revamped policies and manual procedures, and targeted development of automated management and remediation tools. The desired outcome, in addition to more quickly ingesting content into the archive, is the most efficient, economical application possible of technical staff to the automation of problematic content handling, as well as a streamlining, rather than expansion, of manual production processes for content continually growing in quantity and complexity.

Keywords – Sustainability, economies of scale, automation, managed ingest, minimal ingest

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation.

I. INTRODUCTION

The activities associated with selection, acquisition, and other pre-ingest processing (including quality assurance) of content are well-understood to be critical cost drivers for digital preservation (See, for example, [1], [2], [3]).

Sustainable preservation services must realize economies of scale and of the strategic application of automation, so that the growing size of an archive -- the volume of content processed, the ever-increasing amount of storage required -- doesn't mean comparably increasing costs. Additionally, these services must continually monitor what are new or, often, continually changing input streams of content, to ensure that variable content does not result in equally variable, unpredictable, and (in the worst case) runaway costs. As the Digital Preservation Coalition (DPC) noted in its review of digital preservation of non-print legal deposit materials at the British Library,

> "Increasing volume, complexity and unpredictability of content place considerable strain on digital preservation workflows in a variety of ways. Greater volumes (both in numbers of items and sizes of component files) place strains on the workflows that must process them, requiring more resilient software processes and greater workflow automation to enable issues to be resolved without backlogs arising.... Unpredictability requires greater flexibility to react to changes in content and its supply. Deposited data that doesn't conform to previously encountered norms must be detected and workflows adapted to process it. The accuracy and completeness of digital preservation activities will be impacted with adaptation to meet these challenges." [4]

Certainly this has been a challenge for Portico. Portico is a community-supported digital preservation service for electronic journals, books, and other content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Portico serves as a permanent archive for the content of, at present, 606 publishers (from 60 countries, and on behalf of over 2000 learned societies and associations), with 32,004 committed electronic journal titles, 1,379,448 committed e-book titles, and 220 committed digitized historical collections. The archive currently contains over 103 million archival units (journal articles, e-books, etc.), comprising over 1.67 billion preserved files. Portico is sustained by the support of over 1000 libraries in 23 countries.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Copyright held by the author(s). The text of this paper is published
under a CC BY-SA license (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.1145/nnnnnnn.nnnnnnn

II. Process and Procedures: Content Workflows

How does content make its way into the Portico archive?

First, Portico works out an agreement with a publisher. Besides ensuring the legal right to preserve content, the discussion in this pre-processing phase includes sharing sample content, so that Portico can analyze publisher metadata, as well as the conventions used in packaging together and delivering a collection of files that comprise, for example, the articles of a single issue of a journal.

Portico reverse-engineers those conventions into declarative XML "profiles," which direct the aggregation of individual files in a delivery into complex multi-file digital objects, such as journal articles. These profiles also enable Portico automatically to assign a functional descriptor to each file making up the complex digital object (what we term a "functional unit type"); to determine if an expected component is not in the package; and, once all files are grouped, to detect "left-over" files that are not attached to an archival unit, suggesting investigation is required to determine if the file a new component now being provided, or simply "noise" that can be ignored in subsequent deliveries (for example, "thumbs.db" files in Windows folders). We also develop an XSL transform to normalize publisher descriptive metadata into the Journal Article Tag Suite (JATS) or Book Interchange Tag Set (BITS). Both the original and normalized metadata are preserved in the archival unit.

Those discussions also elicit details necessary for automating publisher submission of content, typically via FTP. These details are again translated into declarative XML rules in Portico "Fetcher" and "Loader" profiles, which drive the automated fetching of content, and assembling of files into batches to be loaded into the content processing (ConPrep) automated workflow system.

When these steps comprising a "publisher setup" are complete, automated processing can begin. The ConPrep workflow, using rules in the profile, assembles the various files in a batch into archival units. Publisher metadata is validated against publisher-provided XML schema, and transformed to JATS or BITS. Technical metadata is assembled about each component file in an archival unit, and preservation metadata (including descriptive, technical, rights, and event metadata) is created and packaged with each archival unit. Accepted archival units are then ingested into the archive, where they are replicated, and where periodic fixity checking is performed.

Portico's content and processing model can accommodate updates to the content that publishers might provide. The updated version of the content is packaged with the original, and both versions are maintained, with accompanying metadata, as a single archival unit. They also accept, for example, technically "deficient" component files such as supplementary images that do not pass JHOVE validation.

What the processing model would not permit to be ingested into the archive are archival units that fail the profile rules described above: those missing mandatory components such as the PDF of an article if the publisher does not provide XML full text, or those with ill-formed or invalid XML descriptive metadata, or those missing XML metadata files entirely. Further, since the workflow is a batch processing system, all archival units in the same batch as a defective one are retained in the ConPrep system, until all problems for all archival units are resolved.

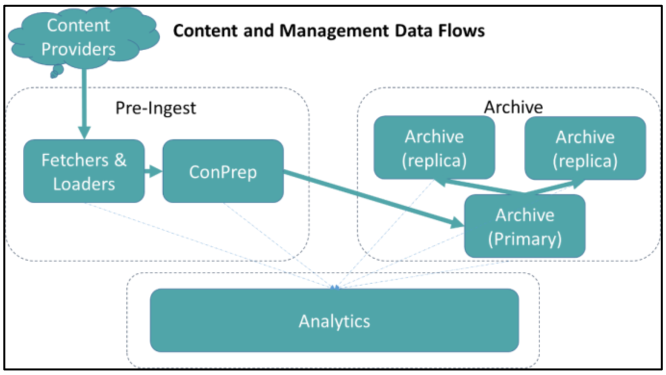


Figure 1

III. Processes and Procedures: Adapting to Growth and Change

By long-standing policy, Portico does not make editorial decisions: Portico does not "correct" the scholarly record. If publisher-provided metadata, for example, states an article belongs in issue 42, when it actually appeared in issue 24, Portico preserves that information as it was received. We preserve the content as it was published.

Nevertheless, there is a quality control function in the handling of publisher-supplied content. Portico's original ingest policies with respect to the cohesion and completeness of content, and the content workflows built in conformance to those policies, were predicated on the detection of certain classes of content defects (for example, missing article components such images, or invalid XML metadata), and on obtaining repaired content and metadata before ingest into the archive. These policies in turn were based on other assumptions:

- that publishers are both willing and able to provide corrections to content and metadata
- that all such defects are equally significant barriers toward making content available, accessible, and useful over the long term
- that there would be a minimal amount of content in problem state, since the content is actively in production at the publisher

The consequence of this policy of "perfect-only" content in the archive was that a significant body of content, across all publishers and content types, was stalled in the content ingest workflow holding queue – even if that content was published with those flaws. This content is expensive. It requires staff people to frequently touch it and manually manage the storage space in ConPrep. In addition, unlike content in the archive, content in the ConPrep queue is not replicated and is not subject to the archive regimen of regular checksum computations to detect "bit rot."

Because bibliographic metadata about archival content is an outcome from processing by ConPrep, the considerable amount of content in the holding queue also was not visible in Portico's holdings data, effectively making them "invisible" to us and to our participants.

And, perhaps most crucially from a cost management perspective, because the ConPrep system is, by design, a transactional system, focused on processing of specific content streams, Portico did not have capabilities for looking across content in those streams, or the processing information detailing defects detected, to make informed decisions on aggregating and prioritizing problem-resolution of content.

Additionally, Portico's very success in attracting more and more publishers to entrust content to our care meant that the staff burden of managing problem resolution, and publisher interactions, was continually increasing.

As we have described elsewhere [5], Portico's recently re-architected, horizontally scalable technical infrastructure is easily able to accommodate geometric growth in content over the past fifteen years. What has proved more problematic has been the relative increase in the number of small and medium publishers, as the total number of publishers and content streams has grown year by year. Currently, over 55 per cent of publishers providing content to Portico are classfied as small or medium.

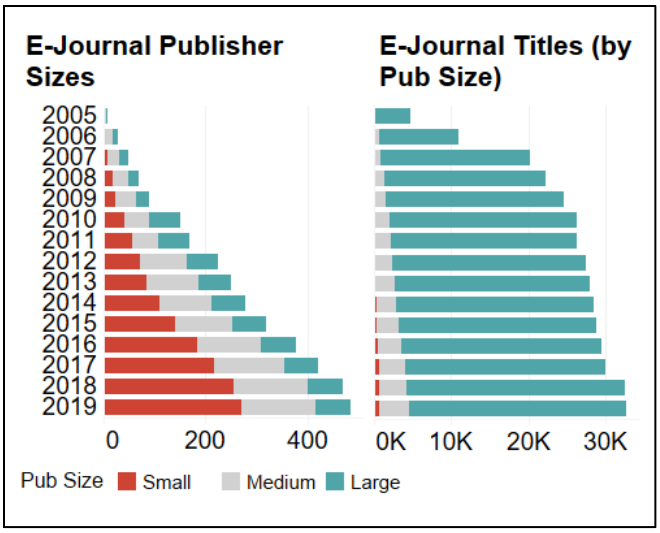


Figure 2

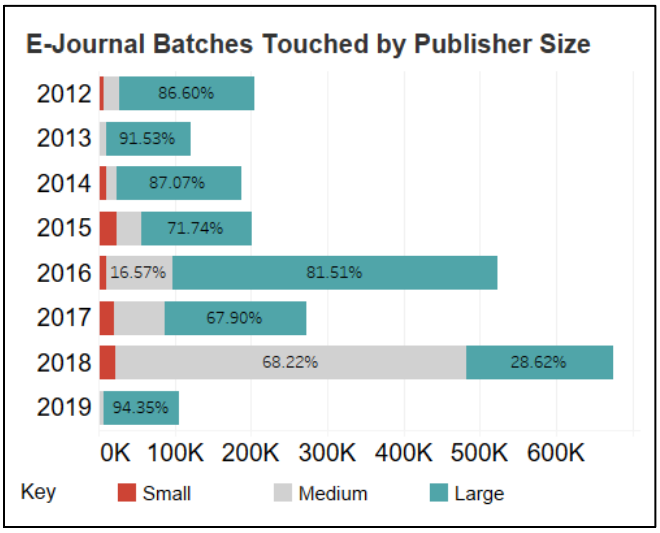


Figure 3

Given that many of these new content streams are in the "long tail" of scholarly publishing, comprising publishers with fewer technical resources, or simpler, sometimes manual production processes, our experience has been that the number of problems per provider has begun and will to continue increase as well.

Another way of saying this is that an increasingly small subset of the content coming into Portico necessitates an increasingly greater amount of staff time in manual interventions to deal with problems in the automated processing of that content. In 2018, a third of all batches, across all content types, required some sort of manual handling.

So, while we now had the capability to scale our automated systems horizontally in a reasonably effortless and economic fashion, the same was not true of our manual processes. A fundamental change to ingest policy was required to meet this challenge.

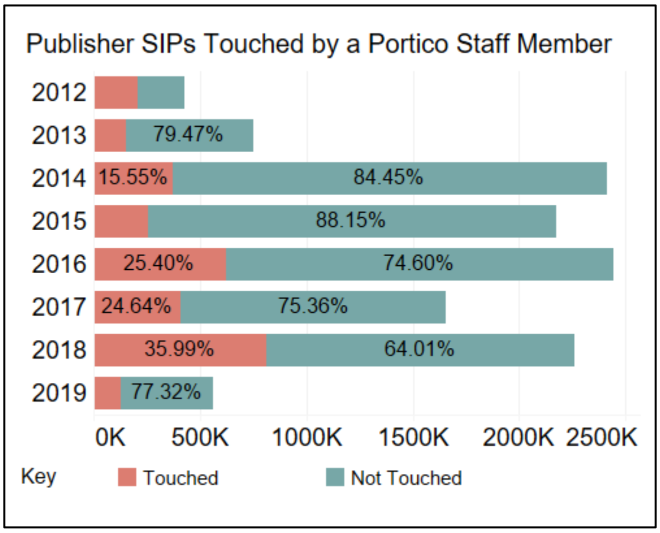


Figure 4

IV. The Straight-to-Ingest Project

A. Goals

The goal of the Straight-to-Ingest (S2I) project was simple: to make changes to our production processes (both automated and manual) that would enable us to move all content, regardless of its "cleanliness" or completeness, into the archive. This would enable us to eliminate the backlog of defective content in the ingest queue, to scale our manual production processes to meet the demands of increasing growth and variation in quality of the content entrusted to Portico, and to make informed decisions on how to prioritize the handling of the most seriously defective content.

This does not mean that we intend to ignore defects in content preserved in Portico. Rather, we have moved the point in our automated and manual workflows at which we deal with bad content, from its original location (correcting all defects in ConPrep before ingest) to new, post-ingest processing. There, we can employ the new analytic tools of the archive, as well as a comprehensive view of all content deposited, to inform our interventions to make repairs to defective content, if possible.

Specifically, S2I was designed to:

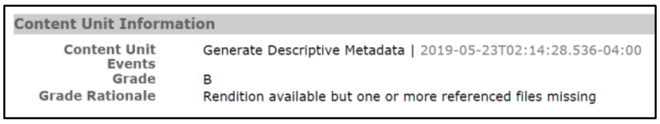
- scale our manual production processes to leverage the horizontal scalability of the new technical architecture to deal with increasing volumes, types, and sources of content, of widely varying quality, without necessitating increasing staff
- move all content into secure, managed, long-term preservation and out of the transactional content processing system where defective content is often stalled
- provide greater transparency to both publishers and libraries about the current state of content committed and submitted to Portico
- bring information about all content submitted to Portico into our holdings metadata, and provide a more complete and correct picture of content in our care
- leverage new analytics capabilities to provide us with a deep understanding of what problems exist in the content, and to enable us to make informed decisions about how to allocate staff resources to address them
- flexibly and rapidly address major and urgent content defects
- eliminate redundant interactions with publishers over already-corrected content, or content for which publishers are unable to provide corrections
- minimize or eliminate manual interventions to correct defective content

B. Process and Procedural Changes

We now "grade" archival units as they move into the archive. These "grades" will enable us to prioritize resolution of the biggest problems of "broken" content. Some of those problems are with the content itself, and are of varying severity for the "community of use" of preserved content. For example, if we have descriptive metadata in an XML file that references a missing figure graphic in the article abstract, but we also have a PDF file with all the content of the article, including that figure, the article is completely usable. That is, a reader has enough information, through the PDF, to understand the full intellectual content of the article. That article would be ingested into the archive with a grade of B, along with information about the file referenced, the fact that it was missing, and the rationale for giving the article less than an A grade.

Another benefit of "grading" is that it surfaces an accumulation of errors in a particular content stream which seems to indicate a dramatic shift in the regular practices of the publisher, that has not been communicated to Portico, but which perhaps mean we need to make changes to our profiles, transforms, or tools.

We updated publisher profiles to indicate, on a per-stream basis, the minimum grade allowable for content from that stream to be ingested into the archive. Also, in addition to ConPrep workflow changes to categorize errors and grade content, we updated our preservation metadata schema to include grade information about each archival unit. This grade information is displayed when the archival unit is accessed.



Content Unit Information

Content Unit Events	Generate Descriptive Metadata | 2019-05-23T02:14:28.536-04:00
Grade	B
Grade Rationale	Rendition available but one or more referenced files missing

Figure 5

In addition, for any archival unit with a grade less than 'A,' a new functional component has been added to the package containing the preserved object – a JSON file containing detailed error tracking information about the nature of the detected defects. A long standing requirement of the Portico archive is that it is "bootstrap-able" – everything one needs to reconstitute the archive is contained in the archival units themselves. This means we must capture all these errors in a machine (and human) readable way within the archival units. All error information is also cached in our analytics system.

We have developed new reports, both for production staff and for communicating to publishers about problematic content. These reports enable us to manage and report problems by publisher and by defect type. Additionally, we have developed new workflows to capture the "feedback loop" of publisher responses (or of updates to problematic content without accompanying feedback from the publisher). This enables us to eliminate duplicate or out-of-date reporting, as well as to detect publisher remediation, or, should it be the case, to record a publisher response indicating they are unable to repair defective content.

C. Early Outcomes

The first use case of problematic content implemented in the project is "Grade B" content that is missing referenced ancillary files, but has a component that provides the full "intellectual content" of the archival unit. In the first few weeks of processing, we are finding, as we hoped, that batches containing such problems are spending less time in ConPrep. Additionally, previously blocked content without problems, but in the same batch as problem content, is now going into archive, where it is being replicated and is available for access

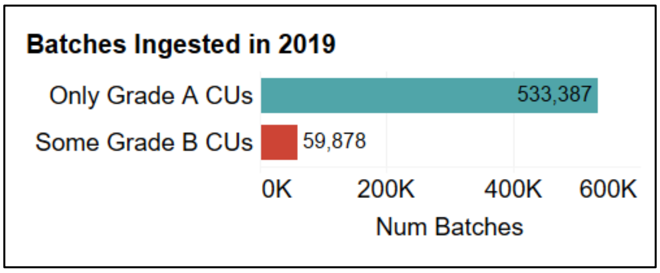


Figure 6

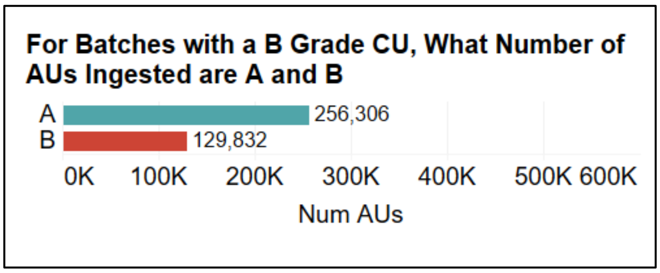


Figure 7

We will be looking over the next several months to see if our new reports and automated feedback loops, along with streamlined processes for communication with publishers, is significantly reducing the management burden on our production staff.

D. Next Steps

Over the coming months, we will expand the grading scheme to include other categories of errors. A next likely use case is using alternative methods (regular expression parsing, natural language processing) for extracting at least minimal descriptive metadata for archival units where publisher-supplied XML is not well-formed and valid.

Additionally, we will be building improved "dashboard" reporting and alerts that, based on accumulated process analytics and grade information, indicate some uncommunicated change in a publisher's content stream – or, perhaps, a previously undetected error in Portico's workflow configuration, tools, or transforms.

V. Implications for Best Practice

As noted by Jurik et al in their description of minimal effort ingest at the State and University Library, Denmark [7], moving content that, according the policies of a given archive or the "submission agreement" worked out with a contract provider, is somehow incomplete or deficient into the archive implies a change to the function model of OAIS [9]. It moves a function from the "Ingest Functional Entity" to the "Archive Function Entity". OAIS assumes only wholly perfect or complete content is allowed into the archive; it also assumes on-going, as well as tractable, communication with content providers. Nor does OAIS provide a very rich model for what Caron et al referred to as enrichment and enhancement of digital content after ingest [6]. Realistically, we feel that not only must we find a way to accommodate what we would consider to be less than "perfect" data, but also that we must make pragmatic decisions to ensure the overall sustainability of the archive, and of digital preservation as a whole.

Ideally, Portico would like all content coming into the archive to be "born preservable" – complete, correct, and supported by ample contextual information. We feel however, as Stephen Abrams has suggested [8], that while OAIS provides a useful model for measuring the trustworthiness of archival processes and procedures, this must be balanced against, and completed by, a measure of the effectiveness of communication with a future user – a robust topic to be addressed by the next iteration of preservation standards.

References

[1] H. Uffen and T. Kinkel, "Controlling the Cost of Digital Preservation," Dutch Digital Heritage Network, 2018 https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2019/02/Article_DDHN_Controlling-the-cost-of-digital-preservation.pdf.

[2] Charles Beagrie LTD, "Costs Factsheet," CESSDA Cost-Benefit Advocacy Toolkit, DOI: 10.18448/16.0003.

[3] Mosely, S., Moran, J., McKinney, P., Gattuso, J., 2016. Conceptualising Optimal Digital Preservation and Effort, iPRES2016 *13th International Conference on Digital Preservation,* Bern, Switzerland.

[4] Digital Preservation Coalition, *Non-Print Legal Deposit Digital Preservation Review: Final Report*, Dec 2017, https://www.bl.uk/britishlibrary/~/media/bl/global/digital%20preservation/non-print-lega-deposit-digital-preservation-review.pdf.

[5] K. Wittenberg, S. Morrissey, "The Evolution of Infrastructure: Making a Renewed Investment in Preservation at Portico," *The Scholarly Kitchen,* Oct 2018, https://scholarlykitchen.sspnet.org/2018/10/17/guest-post-the-evolution-of-infrastructure-making-a-renewed-investment-in-preservation-at-portico/

[6] B. Caron, J. De La Houssaye, T. Ledoux, S. Reecht. "Life and Death of an Information Package: Implementing the Lifecycle in a Multi- Purpose Preservation System," iPRES2017 *14th International Conference on Digital Preservation,* Sep 2017, Kyoto, Japan.

[7] B. Jurik, A. Blekinge, K. Christiansen, "Minimal Effort Ingest," iPRES2015 *12th International Conference on Digital Preservation*, Nov 2015, Chapel Hill, NC.

[8] S. Abrams, "Nothing succeeds like success: An approach for evaluating digital preservation efficacy," iPRES2018 *15th International Conference on Digital Preservation*, Sep 2018, Boston, MA. https://escholarship.org/uc/item/8cp160c9

[9] Space Data and Information Transfer Systems. ISO 14721:2012 Open Archival Information System (OAIS) - Reference Model. The International Organization of Standardization, 2012.