Minimal Effort Ingest

Bolette Ammitzbøll Jurik
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2320
baj@statsbiblioteket.dk

Asger Askov Blekinge
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2100
abr@statsbiblioteket.dk

Kåre Fiedler Christiansen
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2036
kfc@statsbiblioteket.dk

ABSTRACT

In this poster we present the concept of *Minimal Effort Ingest* into a digital repository and discuss benefits and disadvantages of this approach.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Digital Preservation, Digital Repositories, Minimal Effort Ingest, Ingest Workflow, Quality Assurance, OAIS

1. MINIMAL EFFORT INGEST

An expensive part of ingesting digital collections into digital repositories is the quality assurance (QA) phase. Traditionally, data and metadata are quality assured before ingest, to ensure that only data which complies with the repository data formatting and documentation standards is preserved. In Minimal Effort Ingest, we postpone the QA of data and metadata until after the data has been ingested and even further, if resources are not available. This approach makes it possible to secure the incoming data quickly.

There are benefits and disadvantages to this approach, as detailed below. At the State and University Library, Denmark, we have implemented Minimal Effort Ingest as the workflow for our *Newspaper Digitization Project* [4]. About 30 million newspaper pages are being scanned, and we receive about 50,000 scanned pages per day. We have built a workflow which first ingests the data and metadata into our repository and then performs QA on the ingested data. If a delivery is found to be invalid, a new scanning is requested. When the new delivery is received and approved, the old delivery is purged from the system.

It has proven easy to continually add additional checks to the QA, and to run these checks on both the new deliveries and the already approved content.

2. OAIS COMPLIANCE

It has long been standard to establish trustworthiness of a digital repository by a more or less strict compliance with the *Open Archival Information System (OAIS)* Reference Model [2].

In the OAIS model a *Submission Information Package (SIP)* is received into temporary storage, where QA is performed, then an *Archival Information Package (AIP)* which complies with the archive's data formatting and documentation standards is generated, and *Archival Storage* is updated.

In the Minimal Effort Ingest model the SIP is transformed into a minimal AIP and ingested directly into *Archival Storage*. QA is performed from the *Data Management Functional Entity* on data in *Archival Storage*. That means we have moved the QA step from the *Ingest Functional Entity*, where it is performed on SIPs, to the *Data Management Functional Entity*, where it is performed on the minimal AIPs.

Ingesting the SIPs into *Archival Storage* directly as described above appears to be in contradiction with the OAIS reference model. The QA is however still performed, and we thus claim that a repository implementing the Minimal Effort Ingest model will be, content- and preservation-wise, *eventually consistent* with a repository implemented in strict compliance with the OAIS model.

The State and University Library, Denmark has incorporated Minimal Effort Ingest into both it's *Digital Preservation Policy* [5] and *Strategy* [6]:

> "As soon as possible after a collection has been received, all data and metadata are ingested into the library's Repository to preserve the functionality of the digital collection. Once a collection has been ingested into the Repository, a number of preservation actions can be carried out. The owner of the collections and the system owner coordinate the activities."[5]

We audit the State and University Library, Denmark as a trustworthy digital repository using the ISO 16363 Audit

and Certification of Trustworthy Digital Repositories Standard [3]. While this standard uses the common conceptual framework provided by OAIS, it does not require strict compliance with OAIS.

3. BENEFITS

Ingesting content early into the repository has a number of advantages.

3.1 Preserving as Early as Possible

By adding the content to the repository as early as possible, we ensure that the content is preserved, at the very least in it's binary representation.

3.2 A Consistent Platform

By ingesting the data and metadata into the repository system, we have a consistent platform for doing QA and normalization.

Instead of developing tools specific to the ingest workflow for a given collection, we create tools that work on the repository. This gives us a unified platform for the development process, and it also makes it easier to reuse the tools for different collections.

3.3 Repository Tools instead of Ingest Tools

QA and normalization tools can be used in other phases of the information flow than ingest. By making the tools into repository tools, we can run the tools whenever it is relevant.

This also ensures that any QA actions are performed on the same data we preserve. This is in contrast with an OAIS ingest workflow, where content conceivably might change in the interval between the QA step and the actual ingest step.

We can also update the QA tools, and rerun them on the collections, whether they are recently ingested or approved a long time ago.

3.4 Recording Preservation Events

Since all preservation actions are performed on content within the repository, it becomes natural to save information about the actions as metadata in the repository. In the newspaper digitization project [4], we use PREMIS [1] to store this metadata as preservation events.

3.5 Empowering Repository Managers

Since all tools now work on the repository content, it is much easier to empower repository managers to work with the digital preservation tools without involving IT resources.

In that way repository managers without special IT background can take responsibility for preservation actions.

4. DISADVANTAGES

Minimal Effort Ingest does have drawbacks.

4.1 Normalization

When working with normalization in an ingest workflow, it may result in having both pre-normalization and post-normalization copies of content in the repository. This requires, depending on policy, either twice the space, or a method for cleaning up in the repository.

4.2 Content Failing QA

If content is not approved by the QA process, it may be necessary to either delete content or replace content with a new version from the content provider if possible. This can be a problem, since repositories often have policies that content should never or rarely be deleted.

4.3 Malicious content

By moving QA process from the ingest phase to a process within the repository, we risk ingesting content that has not been analysed or filtered for malicious content.

This could lead to vulnerabilities, if the content is accessed before such a check can be made, or if the analysis software itself is vulnerable. Moving the QA process from the *Ingest Functional Entity* to the *Data Management Functional Entity* can be seen as an increased security risk. Extra care should be taken that malicious content in the repository cannot compromise the security of the repository.

5. CONCLUSION

All things considered, performing preservation actions post-ingest on the repository content, rather than during ingest provides benefits in both development effort and preservation liability.

6. ACKNOWLEDGMENTS

Thanks to the National Library Division and the IT Preservation Section at the State and University Library, Denmark, who took part in formulating and implementing the Minimal Effort Ingest approach in general and particularly for the newspaper digitization project. A special thanks go to Gry Vindelev Elstrøm, Knud Åge Hansen and Jette G. Junge. You know why!

7. REFERENCES

[1] Library of Congress. `http://www.loc.gov/standards/premis`, 2015. Accessed June, 2015.

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

[3] Space Data and Information Transfer Systems. *ISO 16363:2012 Audit and Certification of Trustworthy Digital Repositories*. The International Organization of Standardization, 2012.

[4] `http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/newspaper-digitization`, 2015. Accessed June, 2015.

[5] `http://en.statsbiblioteket.dk/about-the-library/DigitalPreservationPolicy_2014.pdf`, 2015. Accessed June, 2015.

[6] `http://en.statsbiblioteket.dk/about-the-library/DigitalPreservationStrategy_v3.pdf`, 2015. Accessed June, 2015.