Consulting report

for Simon Fraser University Archives

May 29, 2015

1. Background

In February 2015, SFU Archives approached Artefactual Systems to propose a series of enhancements to Archivematica and AtoM, open-source software tools for digital preservation and online access developed by Artefactual and used by the Archives. Over the following weeks several meetings were held to start to refine the requirements and develop an understanding of the scope of development that would be needed. In April the Archives approved a small consulting contract in which Artefactual would analyze the features in greater detail, formulate a basic plan for development and provide cost estimates for each enhancement. This report provides summary information about that analysis and a table of time and cost estimates. The report will be the basis for a development contract to be approved by June 30, with work expected to begin in early July.

2. Proposed development approach

Artefactual recommends using an agile development methodology to develop the requested features. This is an iterative, incremental approach to software development which incorporates repeated, relatively rapid rounds of requirements analysis, coding and client acceptance testing. This stands in contrast to more traditional approaches, which involve long periods of development followed by deploying completed or nearly-completed features ready for production use. The agile method allows software developers and users to work together to iteratively refine requirements, introducing new requirements and re-prioritizing existing ones if the need arises. The method recognizes that it is difficult to foresee all the functionality that might be required to meet a client's needs at the beginning of a project, and that hands-on testing of new features should happen frequently throughout the development process.

The agile development methodology is well-suited to hourly contracts as opposed to fixed-fee contracts. A fixed fee contract specifies that the developers will deliver x features in x amount of time for x amount of money. It is relatively inflexible and accommodates change requests only with difficulty. For hourly contracts, Artefactual provides an estimate of the amount of time required to complete a feature or given set of features: if the amount of time required is less than estimated, the client is invoiced only for the hours used; if the amount of time required is greater than expected, Artefactual works with the client to adjust the scope of the

feature or to estimate how many more hours will be required to bring the feature to completion. Hourly contracts accommodate change requests more easily and help mitigate risks for both sides posed by under- or over-estimating the complexity of a given feature. Because Artefactual prefers agile development methods and hourly contracts, it offers a discounted hourly rate for such contracts ($130 per hour as opposed to $150 for fixed fee contracts).

3. Proposed development plan

The analysis identified 12 main areas of development, as follows:

1. **Archivematica-AtoM DIP upload**: Implement improvements to current handling of DIP upload from Archivematica to AtoM.
2. **AtoM rights management:** Build on recently completed work to implement actionable PREMIS rights, focusing on supporting management of copyright- and statue-based restrictions.
3. **Metadata-only DIP upload and AtoM CSV import**: For digital objects that are restricted or under review, upload metadata about the objects to AtoM to enable discovery. Upload the digital objects at a later date when restrictions have been removed.
4. **Ingest email accounts into Archivematica**: Ingest SFU email accounts into Archivematica for long-term preservation and access purposes.
5. **Manage sets of records in AtoM**: Save search sets; export and print saved sets.
6. **Web archiving**: Ingest WARC (Web Archiving format) files, extract metadata as needed and make extracted metadata searchable in the Archivematica archival storage tab.
7. **Fixity checking and reporting:** Run regular collection-wide fixity checks and provide information on results.
8. **Transfer backlog search and retrieval enhancements**: Improve ability to find transfers and individual directories and files within transfers; add ability to download copies of transfer content via the dashboard (similar to AIP retrieval)
9. **Atom accessions module enhancements**: Link accessions to repositories; create repository-specific accession number masks.
10. **AtoM permalink URLs**: Implement improvements to how AtoM manages the permalink URLs that function as web addresses for records.

11. **Archivematica OCR micro-service**: Add Tesseract micro-service to Archivematica to create a PDF with an OCRed text layer as preservation layer
12. **AtoM treeview enhancements**: Review AtoM's current treeview functionality, get user community input, identify improvements or alternatives

This proposed plan addresses each of these main areas and the development tasks proposed for each one. Note that the phases in this plan, although numbered, are not necessarily linear; some phases will overlap other phases, and the tasks can and likely will be re-ordered as needed, as part of an agile development process. Because development costs are difficult to estimate, and because user requirements often evolve over the course of a project, SFU Archives may choose to prioritize some features over others, moving resources from one task or feature to another. This may result in excluding some tasks or features if they become a lower priority; in particular, features designated as medium priority in the proposed plan may be removed if higher priority features take longer or include more enhancements than originally estimated.

The costs provided in this plan are calculated on an hourly contract rate of $130 per hour. The estimates cover project management, requirements analysis, development, internal QA and test deployment for client acceptance testing. Production deployment may be included if desired and feasible (to be determined on a feature-by-feature basis).

Phase 1: 30 hours Development area: **Archivematica-AtoM DIP upload** Priority: critical
Estimated development cost: $4,000
Notes: ● Archivematica 1.4 and AtoM 2.2 implement a number of improvements and enhancements to the current DIP upload functionality. These are described for information purposes in **Appendix A**.
Proposed development tasks: ● Add filter to AtoM to strip extensions from filenames ● Use placeholder title (e.g. “Untitled 1”) instead of filename for item-level description title in AtoM

Optional development task and estimated cost:

- Allow the user to manually change title of description instead of using filename ($3,500)

Phase 2: 180 hours
Development area: **AtoM rights management**
Priority: critical

Estimated development cost: $23,400

Note:

- SFU Archives has provided a table describing automation rules for actions based on statute and copyright restrictions. See **Appendix B**.

Proposed development tasks:

- Add a new taxonomy (PREMIS Statute) to database. Change Rights > Statute field to auto-complete and allow selecting existing statute or entering a new one. New statutes entered via auto-complete would automatically be added to taxonomy.
- Add “Basis” to Admin > Rights setting page as per "SFU rules" table (See **Appendix B**)
- Update PREMIS ACL checks to include "Basis" conditions
- Show copyright pop-up on every download attempt when restriction: “conditional” and right: “allow”
- Add admin setting to toggle pop-up on/off
- Add admin field to define text to show in copyright pop-up

Phase 3: 160 hours
Development area: **Metadata-only DIP upload and AtoM CSV import**
Priority: high

Estimated development cost: $20,800

Proposed development tasks:

- Add AtoM REST API endpoints to GET archival hierarchy and PUT archival description
- Add Archivematica REST API calls to AtoM endpoints to GET archival hierarchy and PUT archival description

- Show AtoM level of description in Archivematica appraisal Tab
- Investigate uploading small amount of digital object technical metadata from Archivematica to AtoM

Phase 4: 84 hours
Development area: **Ingest email accounts into Archivematica**
Priority: high

Estimated development cost: $10,920

Notes:

- Email preservation is highly complex and to date has not been adequately addressed by the archives and library community. The proposed development tasks detailed below will allow SFU Archives to ingest Zimbra backups and mbox files into Archivematica for basic processing and storage. The optional tasks are designed to explore tools and processes for more complex tasks such as attachment extraction and provision of access. A proposed workflow for preserving email accounts incorporating attachment extraction is provided in a diagram in **Appendix C**.

Proposed development tasks:

- Prepare acquisition and ingest plan for SFU email accounts
- Test capture of Zimbra as maildir and conversion to mbox
- Establish descriptive metadata requirements for Zimbra and maildir/mbox AIPs
- Assist SFU Archives to acquire Zimbra backups and prepare them for ingest

Optional additional development tasks and estimated costs:

- Test mbox attachment extraction tools ($4,160)
- Research and test incorporation of attachment extraction tools into Archivematica micro-services chain; incorporate selected tool(s) if feasible ($8,320)
- Test ePADD open-source email management tool functionality using Artefactual mbox samples ($2,080)
- Test ePADD functionality using SFU mbox samples ($2,080)

Phase 5: 112 hours
Development area: **Manage sets of records in AtoM**
Priority: high

Estimated development cost: $14,560

Development tasks:

- Add user interface to select ad-hoc set of archival descriptions from search and advanced search results page
- Add user interface to select ad-hoc set of archival descriptions from browse page
- Save selected archival descriptions in browser session
- Add page to view, sort and remove selected descriptions
- Print selected descriptions (using HTML/CSS print template)

Optional additional development task and estimated cost:

- Export and download a CSV file for selected descriptions ($10,400)
- Print selected descriptions as PDF ($20,800)

Phase 6: 72 hours
Development area: **Web archiving**
Priority: high

Estimated development cost: $9,360

Notes:

- WARC (Web Archiving format) files are plain text files with binary content (such as raster images) embedded as base-64 encoded files. The plain text includes detailed metadata about the crawl and embedded content. Archivematica can currently recognize the WARC format and link it correctly to the relevant PRONOM format entry, but no meaningful information is extracted about the contents. The proposed development is designed to extract meaningful information so that some of the contents of the WARC file are searchable in the Archivematica archival storage tab, and to provide a means of identifying embedded filetypes for preservation planning purposes. **Appendix D** provides an example of how this information could be parsed to the Archivematica AIP METS file.
- Successful completion of this work will be dependent on testing WARC ingest using sample WARC files provided by SFU Archives

Development tasks:

- Analyze WARC header information and prepare metadata mapping to Archivematica AIP METS file
- Parse WARC header information to Archivematica METS file

Optional additional development task and estimated cost:

- Adapt existing ingest automation script for automated ingest from Archive-It ($4,680)

Phase 7: 34 hours
Development area: **Fixity checking and reporting**
Priority: high

Estimated development cost: $4,420

Development tasks:

- Document how to run collection wide fixity check "on demand" using fixity CLI python application
- Add script to send email alerts to administrator(s) when a fixity check fails
- Modify Storage Service to record time and results of fixity checks; add column to packages tab

Optional alternative development task and cost:

- Develop web application to report results of fixity checks ($15,600)

Phase 8: 70 hours
Development area: **Transfer backlog search and retrieval enhancements**
Priority: high

Estimated cost: $9,100

Development tasks:

- Add ability to search transfers from archival storage tab
- Provide gui access to transfer directory_tree.txt files and bag manifests
- Add ability to download copies of transfers or selected files from archival storage tab

Optional additional development tasks and estimated cost:

- Add enhanced transfer backlog reporting functionality (such as size, format counts, accession reports) ($5,000 - $10,000)
- Add ability to perform transfer deletion requests (similar to AIP deletion) ($5,000)

Phase 9: 80 hours
Development area: **Atom accessions module enhancements**
Priority: medium

Estimated development cost: $10,400
Development tasks: ● Add accession "repository" field, and copy repository to archival descriptions created from that accession record ● Allow administrator to set different accession mask and counter for each repository
Phase 10: 80 hours Development area: **Atom permalink URLs** Priority: medium
Estimated development cost: $10,400
Development tasks: ● Allow administrator to select option for default permalinks (title, reference code) ● In multi-repository system, allow different settings for different repositories
Optional additional development task and estimated cost: ● Allow administrator to change a permalink (e.g. when title or reference code changes) ($10,400)
Phase 11: 50 hours Development area: **Archivematica OCR micro-service** Priority: medium
Estimated development cost: $6,500
Development tasks: ● Add Tesseract micro-service to Archivematica to create a PDF with an OCRed text layer as preservation layer
Phase 12: 4 hours Development area: **AtoM treeview enhancements** Priority: medium
Estimated cost: $520
Task:

• Deploy World Bank inventory list feature and Canadian Museum of Human Rights full-width treeview (both in qa/2.3.x) to SFU Archives AtoM development/testing instance for assessment
TOTAL ESTIMATED DEVELOPMENT COSTS: $124,280
ESTIMATED COMMUNITY SUPPORT FEE: $10,000
TOTAL ESTIMATED PROJECT COSTS: $134,280

Appendix A: Archivematica DIP Upload to AtoM workflows

DIP upload from Archivematica to AtoM has a number of enhancements in the upcoming 2.2.0 AtoM release. The Archivematica 1.5.0 release will add further enhancements.

The DIP Upload workflow is as follows (atom 2.2 and am 1.4):

1) The DIP is copied to the AtoM server from Archivematica
2) Archivematica POSTs a message to AtoM to say a deposit is ready
3) AtoM parses the DIP's METS file. If there is descriptive metadata for the DIP i.e. a dmdSec in the METS file for the objects directory, then an archival object with level of description = file is created and title = dc.title. For each file in the DIP, AtoM creates an archival object with level of description = item. The title of that description is derived from dc.title if there is a corresponding dmdSec for the object in the METS file, and uses filename if there is no dmdSec for that object.
4) Digital objects are attached to the item-level information object in AtoM.

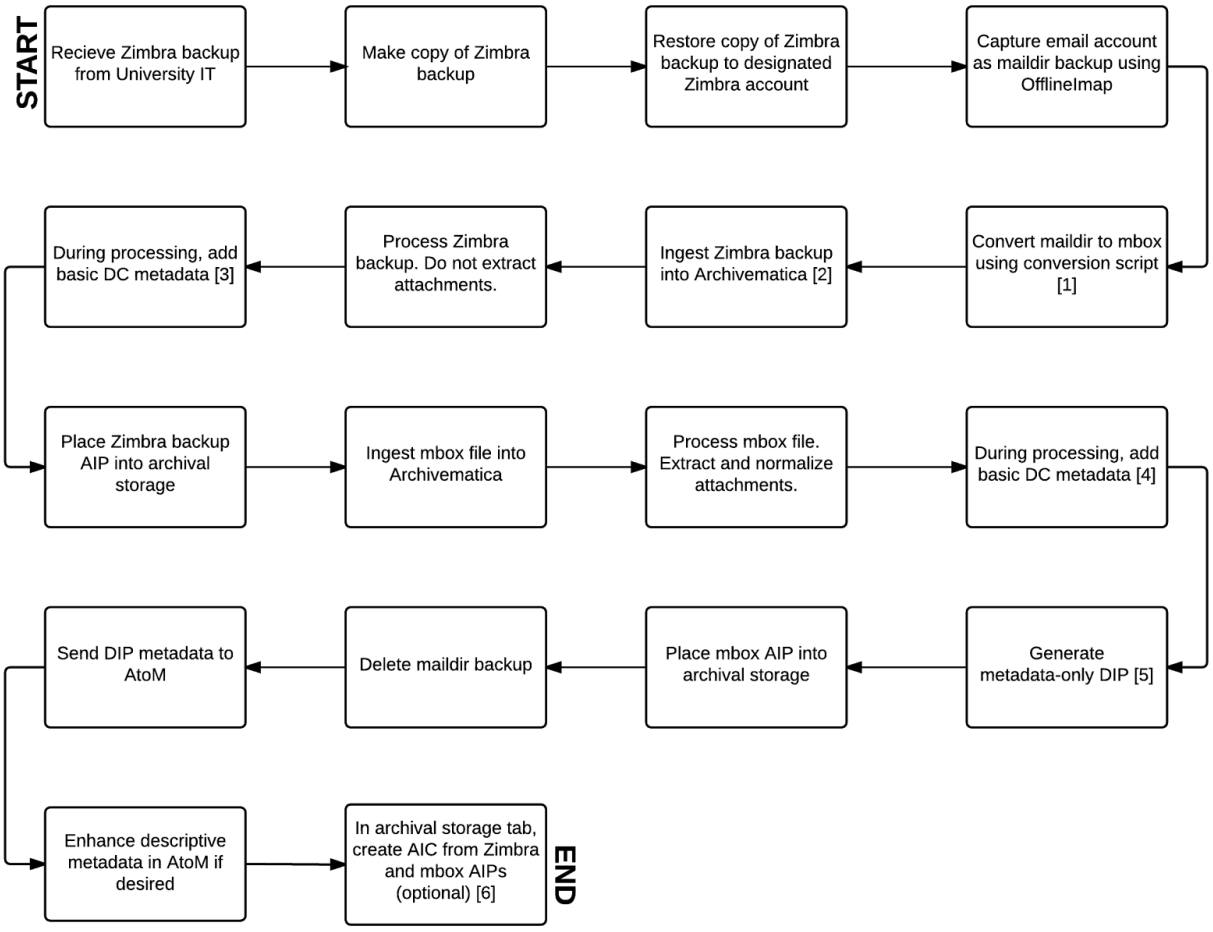
In Archivematica 1.5, the system will automatically create a second METS structMap incorporating archival levels of description added during SIP arrangement. If this structMap is present during DIP upload, a descriptive hierarchy corresponding to those levels of description will be created automatically in AtoM.

Appendix B: Rules for managing copyright restrictions in AtoM

Restriction	Basis = Statute (FOI)	Basis = Copyright
Allow	Means: material has been reviewed, no personal / confidential info identified, no restrictions apply Action: allow display	Means: work is public domain Action: allow display
Disallow	Means: material has been reviewed, personal / confidential info identified, restrictions apply Action: do not allow display	Means: work is copyright-protected, copyright is owned by third-party (not SFU) and Archives considers work to be “high risk” Action: do not allow display
Conditional	Means: material has not yet been reviewed (“pending review”), it is not known if restrictions apply Action: do not allow display	Means: work is copyright-protected but one of the following applies: copyright is owned by third party but judged low-risk; or is made available under CC license; or is owned by SFU; or owner has given permission to disseminate Action: allow display

Appendix C: Proposed workflow for preserving email accounts

In the proposed workflow, the Archives acquires Zimbra email account backups from SFU IT, creates mbox versions, and ingests both the original backups and the mbox versions into Archivematica. The mbox version is treated as the preservation format, and its attachments are extracted and normalized. Although not included in the diagram, the mbox version would also be considered an access format. A potential workflow for providing access would be to load the mbox files into ePADD, redact restricted content and provide a redacted version of the mbox file to researchers.



[1] Artefactual will explore methods and tools for extracting Zimbra contents as mbox, rather than using maildir as an intermediate format. However, the Zimbra/maildir/mbox conversion path is currently known to work and will be used if direct Zimbra to mbox conversion is not possible. Tools that could possibly used for direct conversion are available at https://addons.mozilla.org/en-US/thunderbird/addon/importexporttools/ and http://www.whatan00b.com/migrating-mail-from-zimbra-desktop-to-imap-server/.

[2] Note that the Zimbra and mbox versions are processed separately and placed into different AIPS. Reasons:

- Different workflow decisions during processing;
- Complexities and errors during processing may be multiplied when the two formats are being processed within the same SIP;
- Copies of entire AIPs may be delivered to future researchers in response to access requests, which would be simplified by delivery of the mbox AIP only.

[3] Dublin Core metadata should be sufficient to identify the email account and link the Zimbra version to the mbox version.

[4] Dublin Core metadata should be sufficient to identify the email account and link the mbox version to the Zimbra version.

[5] An assumption is being made that SFU Archives will need to review and possibly redact the mbox file before disseminating it.

[6] The Dublin Core metadata alone may be sufficient to indicate the relationship between AIPs containing MBOX and Zimbra versions of the same email account. However, if desired, the archivist could choose to create an Archival Information Collection (AIC) consisting of the two AIPs. For more information about AICs in Archivematica, see https://www.archivematica.org/wiki/AIC.

Appendix D: Draft AIP METS output for ingested WARC files

WARC files are plain text files with binary content (such as raster images) embedded as base-64 encoded files. The plain text content includes detailed metadata about the crawl and embedded content. Archivematica can recognize the WARC format and link it correctly to the relevant PRONOM format entry, but no meaningful information is extracted about the contents. This draft mapping proposes placing WARC header information and information about embedded content into the sourceMD section of the AIP METS file (as transfer metadata) to make information about the crawl searchable in Archivematica's archival storage tab. It also includes summary information on embedded file mimetypes, in order to support support format risk monitoring for the WARC file contents over time.

```
<?xml version='1.0' encoding='ASCII'?>
<mets:mets xmlns:mets="http://www.loc.gov/METS/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/version18/mets.xsd">
  <mets:metsHdr CREATEDATE="2015-05-13T19:34:58"/>
  [...]
  <mets:amdSec ID="amdSec_6">
    <mets:sourceMD ID="sourceMD_1">
      <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="BagIt">
        <mets:xmlData>
         <transfer_metadata>
            <WARC-Type>warcinfo</WARC-Type>
            <WARC-Date>2008-04-30T20:48:25Z</WARC-Date>
            <WARC-Filename>IAH-2008043[...]-blackbook.warc.gz</WARC-Filename>
             <WARC-Record-ID>urn:uuid:35f02b38[...]</WARC-Record-ID>
            <Content-Type>application/warc-fields</Content-Type>
            <Content-Length>483</Content-Length>
            <software>Heritrix/@VERSION@ http://crawler.archive.org</software>
            <ip>192.168.1.13</ip>
            <hostname>blackbook</hostname>
            <format>WARC File Format 0.17</format>
             <conformsTo>http://crawler.archive.org/warc/0.17/
                 WARC0.17ISO.doc</conformsTo>
            <operator>Admin</operator>
            <isPartOf>archive.org-shallow</isPartOf>
            <created>2008-04-30T20:48:24Z</created>
```

```
            <description>archive.org shallow</description>
            <robots>classic</robots>
            <http-header-user-agent>Mozilla/5.0 (compatible; heritrix/1.14.0
                +http://crawler.archive.org)</http-header-user-agent>
            <http-header-from>archive-crawler-agent@lists.sourceforge.
                net</http-header-from>
            <response>
              <Content-Type>text/html; charset=UTF-8</Content-Type>
              <Content-Type>image/jpeg</Content-Type>
              <Content-Type>image/gif</Content-Type>
              <Content-Type>image/png</Content-Type>
              <Content-Type>application/x-javascript</Content-Type>
              <Content-Type>text/html; charset=ISO-8859-1</Content-Type>
              <Content-Type>application/x-shockwave-flash</Content-Type>
              <Content-Type>text/plain</Content-Type>
            </response>
          </transfer_metadata>
        </mets:xmlData>
      </mets:mdWrap>
    </mets:sourceMD>
  </mets:amdSec>
</mets:mets>
```