Data Migrations

Some Considerations when Preparing to Migrate to AtoM

http://boingboing.net/2016/11/08/heres-the-unexpected-origin.html
Overview

- Assess what data will be part of the migration
- Do any in-system clean up prior to export
- Review AtoM data formats and available fields
- Establish a crosswalk between your data and AtoM’s fields
- Export your data
- Perform any additional clean up as needed
- Transform your data to an AtoM compatible import format
- Import
- Review your work
- Revise and reimport if needed
- Make small clean up edits in AtoM directly
Expect this to take time

- Our average length for a client data migration project is around 4-6 months. Even for a simple project, there will be a lot of time needed for data clean-up, quality assurance review, and reimports.

Expect to do your import more than once

- It’s unlikely that everything will go perfectly on the first attempt. You’ll discover some records don’t quite match the same pattern as the rest, or one field didn’t import, etc. Don’t be discouraged, and do budget your time with this assumption in mind.
Before Starting

Develop a data management plan while you migrate

• How will you ensure you are not stranding data during the time of your migration? Will you freeze data entry entirely for the length? Manage your data in a spreadsheet? Run a small migration for new data at the end? Make sure everyone knows the plan.

Clarify roles, deadlines, and communication channels

• Ensure everyone involved knows what is expected of them throughout the project, and when. Clearly identify those responsible for key roles, and where to go for support.
Data assessment

Questions to ask in the data assessment phase:

• How many descriptions do you have? How many top-level records?

• Have the records been described based on any content standards? (e.g. ISAD(G), RAD, DACS, MAD, MODS, etc.?)

• Are there custom fields with data in your system? How many? Do they readily map to known standards or not?

• What export formats does your system support?

• Is all record data captured in the exports?

• Are some descriptions “draft” or non-public? Is this information captured in the export?
Data assessment

Questions to ask in the data assessment phase:

- How many digital objects do you have to migrate? What types (images, text, video, etc.) and formats (e.g. JPG, mp4, etc.) are represented?
- Are authority records maintained separately from descriptions? What about other entities? Accession records?
- Is the relationship between these entities and descriptions captured in the export formats available?
- Do these other record types have their own export formats? (e.g. EAC-CPF XML, SKOS XML, CSV, etc.)
- How are hierarchical relationships captured in the export?
AtoM Data Formats

Archival descriptions
- CSV, EAD 2002 XML, MODS XML

Authority records
- EAC-CPF XML, CSV

Accessions
- CSV

Terms (Subjects, Places, Genres, etc.)
- SKOS - many serializations supported

Repository records
- CSV
AtoM CSV Templates

CSV templates

Tip
Did you know you can also find all the CSV import templates inside AtoM? They are located in
lib/task/import/example.

For more information on CSV import, see the CSV section of our User manual. Links directly to each major
release's CSV import documentation are included in the relevant sections below.

Page organization
This page has been divided first by type - archival description, or other CSV import type (as the archival
description CSV templates are the most commonly used, and also the ones most subject to further
development and versioning), and then in each section, by release. If the template has not changed for a
version, it will not be added to the newest release section. You can simply download the link in the
previous version.

https://wiki.accesstomemory.org/Resources/CSV_templates
CSV import will be the best way to get data into AtoM - because the CSV import template is a format specific to AtoM, there is no data loss and all fields are represented.

If you are able to get your data out of your legacy system and transform it into an AtoM-compatible CSV format, we recommend using this method for your migration project.

https://wiki.accesstomemory.org/Resources/CSV_templates
In the example CSV files from v2.2 on, we have included the relevant content standard name and number in the sample data field. This means you can import the CSV template to produce a sort of “crosswalk” or key, showing you how fields in AtoM map to the column headers.

[Table]

<table>
<thead>
<tr>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>repository</td>
<td>archivalHistory</td>
<td>acquisition</td>
<td>scopeAndContent</td>
</tr>
<tr>
<td>Example Repository</td>
<td>Example fonds Archival history (ISAD 3.2.3)</td>
<td>Immediate source of acquisition or transfer (ISAD 3.2.4)</td>
<td>Example fonds Scope and content</td>
</tr>
<tr>
<td>Example Repository</td>
<td>Example item Archival history</td>
<td>Example item Immediate source of acquisition or transfer</td>
<td>Example item Scope and content</td>
</tr>
</tbody>
</table>

[URL] https://wiki.accesstomemory.org/Resources/CSV_templates
AtoM CSVTemplates

Fonds F1 - Example fonds

1. repository
2. Example Repository
3. Example Repository
4. Example Item

Identification area
- Reference code: F1
- Type: Example fonds
- Dates:
  - 2016 (Accumulation)
  - January 1, 2015 - December 31, 2016 (Creation)
- Level of description: Fonds
- Extent and medium: Example fonds extent and medium (ISAD 3.1.3)

Content area
- Name of creator: Creator (ISAD 3.2.1)
- Biographical History: Example fonds Creator History (ISAD 3.2.3)
- Repository: Example Repository
- Archival History: Example fonds Archival history (ISAD 3.2.3)
- Immediate source of acquisition or transfer: (ISAD 3.2.4)

Scope and content
- Example fonds Scope and content (ISAD 3.3.1)
- Appraisal, destruction and scheduling: Example fonds Appraisal, destruction and scheduling (ISAD 3.3.3)
- System of arrangement: Example fonds System of arrangement (ISAD 3.3.4)

Related people and organizations
- Subject access point
- Place access point
- Related places

 CSV Templates for AtoM
Similarly, if you ensure all data entry fields in AtoM are filled in with related content standard names and numbers, you can now export as EAD XML to generate an EAD - ISAD(G) crosswalk from AtoM.
Similarly, if you ensure all data entry fields in AtoM are filled in with related content standard names and numbers, you can now export as EAD XML to generate an EAD – ISAD(G) crosswalk from AtoM.
EAD 2002 XML is a flexible standard with many possible valid but different implementations. For this reason, your locally generated EAD, while valid, may still not import perfectly into AtoM. This is why we prefer working with CSV imports whenever possible.

We recommend running a test import of a representative sample from your source system into AtoM, and using the crosswalk method discussed above to evaluate if you will need to make changes to how your EAD XML is encoded for a successful import into AtoM.
Crosswalking is the process of mapping your source data fields to equivalent AtoM ones.

To do so, you must understand how AtoM handles some data (such as authority records, terms, etc.) first.

There will be cases where there are no 1:1 equivalencies either – you will have to make decisions about how to combine or split apart your existing data to make it work with what is available.

Crosswalking

AtoM is standards-based.

This means you can focus on crosswalking to the content standard you know best. Use the guidance provided in the relevant standard to help inform your mapping.
A(n incomplete) list of the main entity types around which AtoM was built.

- **Term**
  - belongs to a Taxonomy
- **Archival Institution**
  - ISDIAH record
- **Actor**
  - ISAAR (CPF) record
- **Archival material**
  - ISAD(G) record (or RAD, DC, MODS)
- **Event**
  - e.g. Creates, Takes custody of, etc.

https://www.accesstomemory.org/docs/latest/user-manual/overview/entity-types/
Accessions

Accession records have their own CSV import format. As there is currently no international accessions standard, you will need to review the available fields in AtoM closely and determine where to map your data.

Descriptions can be linked to Accessions via the accessionNumber column in the description CSV templates. We recommend importing your Accessions first, then your descriptions with the corresponding accession number, to establish links.

See:
Actors

In AtoM, creators and name access points are maintained separately as **authority records**, so they can be re-used and linked to multiple descriptions.

This means **any creator name or name access point you import with your descriptions will create an authority record, or link to an existing match!**

Make sure that names are consistent in your data, and the biographical/administrative history is about the actor only — not specific to the description.

See:

- [https://www.accesstomemory.org/docs/latest/user-manual/add-edit-content/authority-records/#authority-bioghist-access](https://www.accesstomemory.org/docs/latest/user-manual/add-edit-content/authority-records/#authority-bioghist-access)
Actors

The Actor data you can add to a description CSV is minimal - if you do maintain authority records, then you may want to import them separately via AtoM’s authority record CSV templates.

There are 3 actor CSV templates - the main actors template, 1 to supplement relationship data (between actors and/or resources) and 1 to supplement alternative forms of name.

We recommend importing authority records before descriptions, so you can link them on description import.

See:
Event Dates

- Display date
- Start date
- End date

Description edit templates have 3 date fields. The Display date is what the end user will see - it is free text. The start and end dates must follow ISO 8601 (YYYY-MM-DD, etc) formatting. These fields are used to support AtoM's date range search.
Event Dates

During CSV import, Creators and Dates are paired (as Events – see Entity types diagram).

Use the | pipe character to add multiple creators/dates.

You can use a literal NULL value in your CSV file to keep the spacing correct for dates without actors or vice versa:

See:
Access Points

In AtoM, access points on a description (e.g. subjects, places, genre terms) are maintained separately as terms in a taxonomy so they can be controlled and reused.

This means that access point data in your description imports will either create new terms or link to existing ones. Make sure your data is consistent so you don’t have near-duplicates later! (e.g. “cars” vs “car” vs “automobiles”)

The exception is name access points - these are authority records!

See:

- [https://www.accessstomemory.org/docs/latest/user-manual/add-edit-content/terms/#term-name-vs-subject](https://www.accessstomemory.org/docs/latest/user-manual/add-edit-content/terms/#term-name-vs-subject)
Hierarchies are managed in the description CSV templates via the `legacyId` and `parentId` columns. Parent records must import in a row above child records. The children should have the `legacyId` value of the parent record in the `parentId` column.

See:
# Digital Objects

Can be imported with descriptions using the `digitalObjectURI` or `digitalObjectPath` columns.

**URIs** point to external, web-accessible resources - must end in file extension!

**Paths** point to a local directory added to your server prior to import.

See:

<table>
<thead>
<tr>
<th>AU</th>
<th>AV</th>
<th>AW</th>
<th>AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>publicationStatus</td>
<td>digitalObjectPath</td>
<td>digitalObjectURI</td>
<td>physicalObjectName</td>
</tr>
<tr>
<td>Published</td>
<td>/path/to/my/local/exampleDigitalObject.pdf</td>
<td><a href="http://www.example.com/example-external-image.jpg">http://www.example.com/example-external-image.jpg</a></td>
<td>Example Shelf</td>
</tr>
</tbody>
</table>
You have 3 main options when it comes to transforming your data into an AtoM-compatible format:

- **Manual data transformation**
- **Tools such as OpenRefine**
- **Transformation script**

OpenRefine is "a free, open source power tool for working with messy data and improving it."

There are many great free resources to help you get started.

See:

- [http://openrefine.org/](http://openrefine.org/)
- [https://github.com/OpenRefine/OpenRefine](https://github.com/OpenRefine/OpenRefine)
Data Transformation

Use OpenRefine to:

- Add AtoM column headers
- Normalize names and terms
- Standardize identifiers or accession numbers
- Split source data into separate columns
- Combine data into a single column
- Delete unnecessary rows
- Global search/replace
- etc

You can also use OpenRefine to clean up XML data!
A *transformation script* is generally a script prepared by a developer that takes an input (your source data) and runs a series of operations to transform the data into the desired output (an AtoM-compatible file).

These can be prepared in many programming languages (e.g. PHP, Python, etc).
Import Ordering

If you are working with several different types of data, you may need to perform multiple imports, possibly in different formats. If so, we recommend proceeding in this order to link entities together as your imports proceed.

1. Terms
2. Repositories
3. Actors
4. Accessions
5. Descriptions

We also recommend running a smaller sample test first!