Prioritizing Metadata Output Formats for Thoth

Vincent W.J. van Gerven Oei

Published on: Feb 03, 2021
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
Open book metadata management platform Thoth recently moved to its new home on thoth.pub, and is getting ready to ingest the catalogs of ScholarLed member presses Mattering Press and meson press, complementing those of Open Book Publishers and punctum books. With a new user management system in place, Thoth will have become an operational metadata management system for four fully open access presses, which will continue to provide feedback to improve the open source software.

COPIM Work Package 5, which has managed the creation of Thoth, will then focus itself on the next phase of development, namely the expansion of the suite of metadata output formats facilitated by Thoth. In order to map out the most efficient development path, the WP5 has created an ever evolving wiki covering the multitude to data and metadata formats, persistent identifiers, distributors, content platforms, sales channels, catalogs and indices, and end user interfaces that touch upon the design architecture of Thoth.

As there are many metadata output formats tailoring to a broad variety of stakeholders according to both open and closed standards, below we suggest a two-step approach to metadata-output development for Thoth aimed at maximizing the reach and compatibility of Thoth metadata with other stakeholders in the open access book publishing ecosystem.

**Step 1: Content platform/distributor-targeted Formats**

Although there is a multitude of metadata outputs currently in circulation, a recent report gathering input from many different stakeholders from the OA book supply chain (Clark & Ricci 2020) singles out four specific, largely complementary formats: ONIX, KBART, MARC, and OAI-PMH. This report confirms the stakeholder recommendations of the WP5 Cambridge workshop. As a proper implementation of OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) would require not only the implementation of a particular XML schema, but also a standardized URL structure and special server architecture, we will not include an implementation of OAI-PMH in our codebase, though we will reuse an existing software solution to enable OAI-PMH in Thoth.

**ONIX** is one of the most widely adopted metadata formats currently in use, managed by non-profit EDItEUR. Rather than a specific format, it is an open XML schema, which has led different platforms to implement different flavors of ONIX, with distinct ideas about what a minimally viable record is. Thoth can currently already output ONIX 3.0 files for records in its catalog, which have been successfully ingested by Project MUSE. In a recent blog post (Snijder 2021), Ronald Snijder discussed some of the issues that OAPEN faces ingesting multiple “flavors” of ONIX from different sources and the pipeline they created to normalize the records, basically harmonizing them into an OAPEN ONIX
file. We are currently working to make the Thoth ONIX output as compatible as possible with a multitude of stakeholder platforms.

**KBART** (Knowledge Bases and Related Tools) is a standard facilitating collection-level interchange of metadata managed by **NISO**. This output format is particularly relevant in view of the revenue platform currently being developed in COPIM WP2, which will need to showcase OA collections to potential library (and other) funders.

**MARC** remains the primary format through which libraries ingest metadata into their Library Management Systems. As MARC 21 is slowly being phased out, WP5 will focus on developing MARC XML output (the standard of which is maintained by the [Library of Congress](https://www.loc.gov/)), which is interoperable with other XML-based standards.

### Step 2: Other Open Formats

Besides the four metadata output formats that are content platform/distributor-targeted and essential for the discoverability of OA books, there are also a number of metadata output formats that serve different stakeholders.

- A tab-delimited format such as CSV (Comma-Separated Values) is useful for publishers who want to create human-readable spreadsheets that are used an internal or external communication. CSV is defined by the [RFC 4180](https://tools.ietf.org/html/rfc4180) standard.
- Thoth's native API outputs metadata in JSON (JavaScript Object Notation), which is an increasingly popular format in software development. JSON syntax is defined by the [ECMA-404 JSON Data Interchange Standard](https://www.ecma-international.org/standards/publication/404/). JSON-LD (Linked Data) output will allow publishers to embed book metadata on their websites and to expose these to web crawlers.
- Many scholars use bibliographic reference management software such as Zotero, EndNote, or BibDesk. There is only one open standard accepted by most available software packages, **BibTeX**, which is native to the open source LaTeX typesetting software.
- **WikiProject Books** is a community-driven project to integrate book metadata into Wikidata, an open knowledge base serving the Wikipedia community and Wikisource, a free library. WikiProject Books has developed its own open metadata scheme.

The main criteria for selecting these four output formats for Step 2 are again complementariness, interoperability, and openness. CSV can be easily converted to human-readable spreadsheet formats such as Google Sheets and LibreOffice ODS files, while they can also be easily ingested in for example MySQL databases. The JSON-LD format is complementary to XML-based formats such as ONIX and MARCXML, allowing for the ingest of Thoth metadata into a new generation of online platforms that operate with a different syntax. BibTeX is an open format serving communities of scholars and readers while at the same time being ingestible by most commercially available reference management software.
software. Finally, an open, community-driven project such as WikiProject Books aligns with the values of the COPIM project and has the potential to disrupt the oligopoly of for-profit knowledge bases currently operating in the market.

References

https://drive.google.com/file/d/1iCu5OlT1cFDyq6r6yBC6orgjoEMjtor/view.


Citations
