

Repository metadata contain ontological ambiguities that fold into flat hierarchies

Martin Paul Eve

Professor of Literature, Technology and Publishing
Birkbeck, University of London

Associate Director of Knowledge Commons
Michigan State University

28 August 2026



Image: A series of tiles spelling metadata (Markus Winkler on Unsplash)

Cite as: Martin Paul Eve, "Repository metadata contain ontological ambiguities that fold into flat hierarchies", *eve.gd: Martin Paul Eve*, August 28, 2026, <https://doi.org/10.59348/mjvdw-w0051>.

DOI: <https://doi.org/10.59348/mjvdw-w0051>

Available at: <https://eve.gd/2026/08/28/repository-metadata-contain-ontological-ambiguities-that-fold-into-flat-hierarchies/>

Metadata are commonly referred to as “data about data”; they are considered as data that describe and refer to something else. Institutional repositories collect metadata about scholarly items that have been added to the repository. But to what do those metadata refer?

Are they references to external publications (say I put details of a book published by a publisher in there)? Or are they descriptions of the specific files within the repository, in which case, why do I put the details of the external publication, like the publisher name, even when the publisher did not produce these files? What if it is a metadata-only deposit and there are no files to describe? To what are the metadata then even referring? Because a metadata-only deposit that is published solely in the repository, without any files, makes no sense; they are not pointing to a real object but generating a phantom, triangulated scholarly non-object. Such metadata, with no files and no external manifestations, would refer solely to themselves.

These questions can be boiled down to ask whether metadata records in repositories are:

1. Descriptions of potentially external items? I would often think of them as such. When I put in the details of a book or journal article, I am describing another object out in the world.
2. Descriptions of items in the repository? So is the metadata record referring to this thing we just published on the repo and *not* the external thing? (Or both/neither?) In which case, why might it refer to an external publisher or give details of the item elsewhere etc?
3. Separate published items describing themselves? If I put a published date of today on my new metadata record with no files, because I have no publication date, to what does that date refer? OK, the date that “this item” was published on the repository. But “this item” is just the metadata record. It’s not actually “the publication”, which might subsequently appear elsewhere, invalidating the referential accuracy of the repository record.

Put otherwise: do metadata in repositories *refer externally* or *describe locally*? If the latter, what are they locally describing in the case of a metadata-only record?

I had always thought the external reference model was at least part of it and that then local file descriptions (“This is an AAM” / “This is a preprint”) had the “of” suffix. “This is an AAM of the publication with these details”.

But it’s entirely possible and desirable to have unique content in a repository with no external reference point. This can lead to the absurd scenario of the phantom self-reference.

OAI-PMH: Resource, Item, Record

So what’s the bibliographic model used in repositories? The OAI-PMH spec ([Lagoze et al., 2008](#)) makes quite a good stab at this.

Here is how they separate out the components involved in a repository metadata record:

resource - *A resource is the object or “stuff” that metadata is “about”. The nature of a resource, whether it is physical or digital, or whether it is stored in the repository or is a constituent of another database, is outside the scope of the OAI-PMH.*

item - *An item is a constituent of a repository from which metadata about a resource can be disseminated. That metadata may be disseminated on-the-fly from the associated resource, cross-walked from some canonical form, actually stored in the repository, etc.*

record - *A record is metadata in a specific metadata format. A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item.*

So, we then have the following type of structure: repository **item** → metadata **record** → describes scholarly **resource**.

FRBR: Functional Requirements for Bibliographic Records

The other classic is IFLA's FRBR ([Salaba and Zhang, 2007](#)), which subsequently became The Library Reference Model ([Riva et al., 2017](#)).

Under this system, you have **Work** → **Expression** → **Manifestation** → **Item** (WEMI)

The current Library Reference Model ([Riva et al., 2017](#)) defines these as:

Work: *The intellectual or artistic content of a distinct creation.*

Expression: *A distinct combination of signs conveying intellectual or artistic content.*

Manifestation: *A set of all carriers that are assumed to share the same characteristics as to intellectual or artistic content and aspects of physical form. That set is defined by both the overall content and the production plan for its carrier or carriers.*

Item: *An object or objects carrying signs intended to convey intellectual or artistic content.*

I am also laughing as I re-read this model and remember that the highest level entity specified is “*res*” (thing), defined as “Any entity in the universe of discourse”. Well that’s cleared things up, then!

(I did also find an interesting article on how the FRBR has wormed its way outside traditional library structures ([Coyle, 2022](#)), that helped me remember lots of this stuff and that I just wanted to credit.)

So in a repository scenario, we can actually see, from WEMI, whence some of the strangeness. Consider this list of WEMI objects:

- **Work:** A scholarly argument/article or research output
- **Expression A:** My submitted manuscript
- **Expression B:** My accepted manuscript (AAM)
- **Manifestation 1:** The journal publication on the publisher’s platform
- **Manifestation 2:** The repository dissemination/publication of the accepted manuscript
- ◦ **File:** eve-article-aam.pdf

This might be represented in the repository as:

Author: Eve, Martin Paul
 Title: A Hypothetical Article
 Journal: The Journal of Psychoceramics
 Publisher: Open Library of Humanities
 Volume: 12
 Pages: 100–120
 DOI: <publisher DOI>
 File: eve-article-aam.pdf

This record has combined the properties of different entities. The publisher, volume, pagination and DOI describe the publisher’s manifestation. But the PDF attached to this repository record is not that manifestation! It’s the accepted manuscript, which is an expression or copy related to it.

This is why repository systems end up with relational phrases, such as “version of record,” “accepted manuscript,” “published version,” “isVersionOf,” “hasVersion,” “citation,” etc. Indeed, research by Lisa Gonzalez on repository metadata ([Gonzalez, 2016](#)) has explicitly identified this versioning and relationship problem. For example, she notes MIT’s use of DC.relation.isversionof pointing to the publisher DOI, with a second metadata set describing the version actually held in the repository ([Gonzalez, 2016, p. 256](#)).

This is the first instance where we can see, clearly, though, how these metadata records, which are themselves often partial, patchy, and incomplete in repositories, fold complex ontologies into flat hierarchies of representation.

The Scholarly Works Application Profile (SWAP)

But we're not done with the representations yet! The Scholarly Works Application Profile ([Allinson and Powell, 2009](#)) gives a Dublin Core Profile for the FRBR-based SWAP Model ([Allinson and Powell, 2008](#)). This work was, as I understand it, undertaken (or at least most thoroughly written up) by Julie Allinson at the University of York ([Allinson, 2008](#)).

In SWAP, you have the following (which looks a bit like an extension of WEMI): **ScholarlyWork** → **Expression** → **Manifestation** → **Copy**, and **Agent**. So we know what **ScholarlyWork**, **Expression**, and **Manifestation** are here from before. But we now also have **Copy** and **Agent**. **Copies** are duplicated instances of specific “**Manifestations of Expressions of the ScholarlyWork**”. **Agents** are the people or organizations who did things for the publication. The DC Profile describes and demonstrates these ([Allinson and Powell, 2009](#)) as part of a description set:

```

Description Set (
  Description (
    # description of the eprint as a ScholarlyWork
    ...
  )
  Description (
    # description of an Expression of the eprint
    ...
  )
  Description (
    # description of a Manifestation of an Expression of the eprint
    ...
  )
  Description (
    # description of a Copy of a Manifestation of an Expression of the eprint
    ...
  )
  Description (
    # description of an author, funder, supervisor of the eprint or an affiliated
institution
    ...
  )
  Description (
    # description of an editor of an Expression of the eprint
    ...
  )
  Description (
    # description of the publisher of a Manifestation of an Expression of the eprint
    ...
  )
  ...
)

```

SWAP has been used in anger (I mean, production) at major institutional repositories worldwide. MIT, for instance, commissioned a report into a metadata collision analysis for SWAP usage in their DSpace repository ([Wolfe, 2009](#)). To do this, though, they had to make a decision about what a DSpace “item” actually *is*, ontologically. The mapping they decided on was that a DSpace **item** is

equivalent to a SWAP **Expression**. So an item might be describing an AAM or a preprint or a submitted manuscript or any other **Expression**. As Robert Wolfe puts it in the report, the most important mapping is “SWAP:Expression equals Dspace:Item” ([Wolfe, 2009, p. 4](#)).

Metadata Mangled

These profiles and models can seem endless. Indeed, there is a substantial-looking systematic review of the literature on repositories and metadata ([Cerrão and Castro, 2020](#)), but it is in Portuguese, where my sadly limited polyglotism hits its limit.

Hence, the conceptual problem in a lot of real-world repository metadata cases is not that it's wrong to record information such as the publisher even when you are referring to a local copy. It is more the fact that the schema gives you one record, usually, that represents an implicit graph of multiple entities and relations. We then often serialize the whole graph as though every single statement were a predicate of one single undifferentiated `dc:resource`.

Institutional repositories, then, ambiguously flatten and merge an ontology of an archive with a bibliographic ontology. Sometimes, repository records describe the objects held directly in the repository and affiliated with the item in question. But sometimes they describe scholarly works merely known to the repository and that do not even have to be digital publications. Commonly, these metadata describe externally published manifestations, even while they attach a totally different locally held manifestation or expression in the record. Metadata-only records make these ambiguities unusually visible and clear, particularly when they exist in total isolation, with no known actual work attached.

It seems important, then, that repositories (such as [our excellent KC Works](#)) should make conscious decisions about what their metadata describe. Signposting these ontological decisions explicitly and designing around them will have real-world consequences for the use of the platform and what people deposit and how.

Say you want to encourage people to move away from more traditional publishing structures, we could strongly signal that what the repository describes is what is in the repository. (“What goes in Works stays in Works”, as they (don’t) say.) You then build a description of a local archive.

Alternatively, you could have a strong bibliographic model where you insist that anything in the repository has an external publication “validated”, so to speak, ScholarlyWork reference.

Or you exist chaotically, allowing a mixture of these types of descriptive format. It is perfectly possible to remain in a state of ontological ambiguity here and to use a repository for multiple purposes. The benefit in such a case is that people from both walks, the radical (publish-in-repo) and the more conservative (publish-elsewhere-with-publisher-and-deposit), can use the repository and become accustomed to its operation. It can serve multiple use cases. But the sacrifice you make in that case is the substantial loss of metadata accuracy/clarity. Your metadata will, confusingly, and with the same appearance, represent different objects in different ontological contexts.

Acknowledgements: Thanks to Ian Scott for prompting (and participating in) the discussion that generated this post.

As a closing note, I would like to mark that this is my 1000th blog post on this site, since 2007.

References

This reference list has been generated, in many cases, through an automated process and may therefore contain errors or incomplete metadata. It is provided purely as assistance for finding the items mentioned within this post.

Allinson, Julie, and Andy Powell, ‘Scholarly Works Application Profile’, *DigiRepWiki*, 5 March 2009

<https://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile>

Allinson, Julie, and Andy Powell, 'Model', *DigiRepWiki*, 14 May 2008

<<https://www.ukoln.ac.uk/repositories/digirep/index/Model>>

Allinson, Julie, 'Describing Scholarly Works with Dublin Core: A Functional Approach', *Library Trends*, Project MUSE, September 2008 <<https://doi.org/10.1353/lib.0.0034>>

CERRAO, Natalia Gallo, and Fabiano Ferreira de CASTRO, 'Aplicações de metadados baseadas em FRBR e RDA em repositórios institucionais digitais: uma revisão sistemática da literatura', *Transinformação*, FapUNIFESP (SciELO), 2020 <<https://doi.org/10.1590/2318-0889202032e190080>>

Coyle, Karen, 'Works, Expressions, Manifestations, Items: An Ontology', *The Code4Lib Journal*, 9 May 2022 <<https://journal.code4lib.org/articles/16491>>

Gonzalez, Lisa, 'Representing Serials Metadata in Institutional Repositories', *The Serials Librarian*, Informa UK Limited, 18 May 2016
<<https://doi.org/10.1080/0361526X.2016.1160308>>

KCWorks <<https://works.hcommons.org/>>

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner, 'The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0', Open Archives Initiative, 2 December 2008
<<https://www.openarchives.org/OAI/2.0/openarchivesprotocol.2008-12-02.htm>>

Riva, Pat, Patrick Le Boeuf, and Maja Žumer, *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*, IFLA, 2017 <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf>

Salaba, Athena, and Yin Zhang, 'Functional requirements for bibliographic records: From a conceptual model to application and system development', *Bulletin of the American Society for Information Science and Technology*, Wiley, August 2007
<<https://doi.org/10.1002/bult.2007.1720330606>>

Wolfe, Robert, *Implementing the Scholarly Works Application Profile in DSpace: A Metadata Collision Analysis for the MIT Open Access Initiative*, MIT Libraries, 4 September 2009
<<https://wikis.mit.edu/confluence/download/attachments/49220071/2009-09-04OAMDanalysisv2p2.pdf?api=v2>>