

# LIFESPAN METADATA SCHEMA NOTES

Author: Alison Pope  
 Date: 19 March 2010

Royal Holloway currently has two live collections in its Digital Object Repository: Royal Holloway Research Online containing research publications and a digitised exam paper collection.

In designing the early metadata schemas for the repository we have considered the following factors:

- Building on reference models and best practice guidelines to inform the schema
- The need to design a schema that will work with the functionality of our chosen repository solution, Equella, to deposit, display and discover content.
- The need to expose and share metadata in standard formats for re-use elsewhere.
- We were conscious of the need to curate and preserve objects in the collection but this was not a primary influence and so had not really been a key concern of earlier metadata design efforts. The Lifespan RADAR project was a good opportunity to take the first steps towards informing strategy and practice in this area.

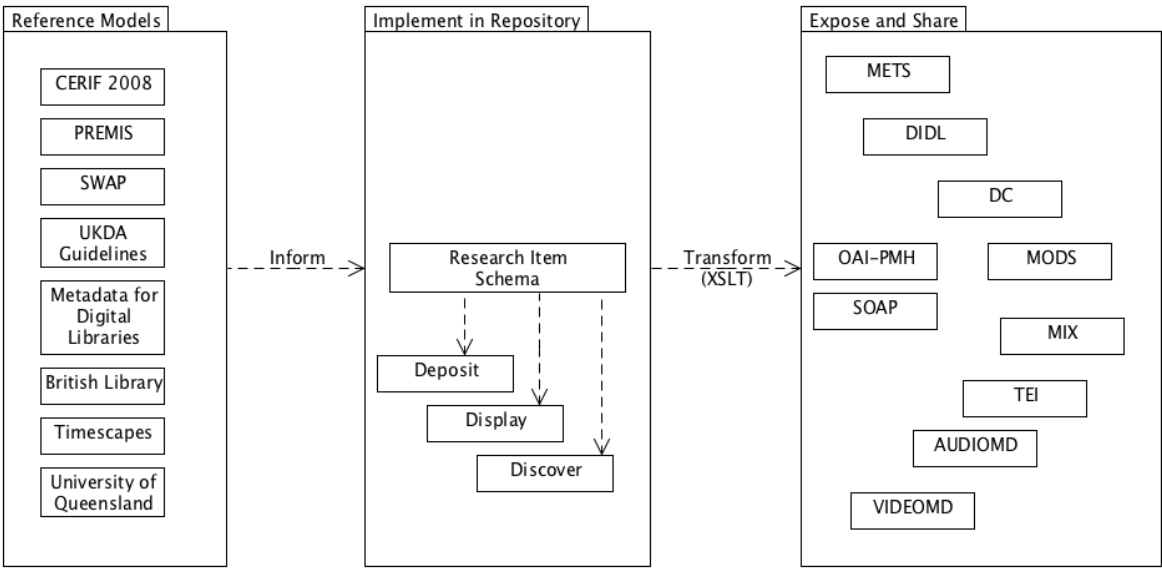


Figure 1: Schema Context Diagram

Royal Holloway Research Online (RHRO) uses a custom schema that was strongly informed by the Scholarly Works Application Profile, (SWAP), which itself builds upon the Functional Requirements for Bibliographic Records (FRBR). Early versions of our metadata schema have already been shared with other Equella users as the basis for their research output collections. When presenting metadata for harvesting via OAI-PMH this custom schema is transformed into

the oai\_dc schema based on unqualified Dublin Core. We have not yet considered transforming this schema into other standard formats. The exam paper collection uses a simple Dublin Core based schema and is not exposed outside of the institution.

The need to host collections to store primary research data presented a new requirement for the repository. The first step was to draw up an ontology of the Lifespan data collection to understand the entities and attributes involved.

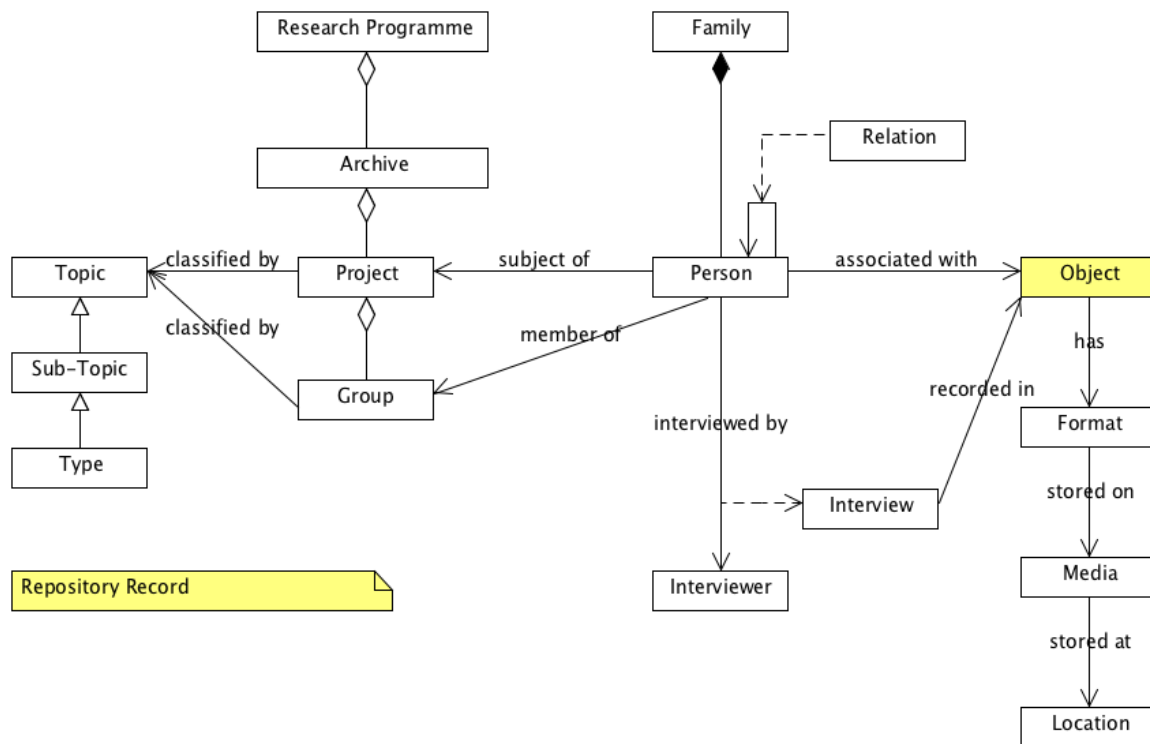


Figure 2: Lifespan Research Ontology Model

Whilst we initially explored having two records, one for person entities and one for object entities, we decided that the curation and preservation focus was on the objects created from the interviews (the research activity), with everything else being potential descriptive metadata about the object in the collection catalogue. The initial draft schema was still too oriented towards this ontology to be re-usable as metadata elements were named specifically around the concepts and attributes needed to represent this ontology in the repository.

We therefore began to think of the object as a generic research item with curation and preservation needs, so we used existing knowledge, along with the research done as part of the Lifespan project to think about how a new schema for a Research Item might work.

Much of the schema work done for Royal Holloway Research Online had concentrated on research outputs as publications but the SWAP based schema used here did not translate so

well to primary data, partially because the versions problem is not so important for primary data and this layer seemed to add unnecessary complexity to the schema. Instead we also considered the CERIF 2008 model for describing research information and the PREMIS Data Dictionary for specifying digital object preservation metadata. We considered research primary data as a variant of research result that is distinct from a publication and decided to simplify the SWAP/FRBR three tier model into the PREMIS two tier model of intellectual entity and object (which can be sub-classed in to three types: representation, file and bitstream).

It should be noted that the PREMIS dictionary specifies semantic units: these need not be in the metadata schema itself if they can be derived within the implementing system, rather than recorded as metadata. By using semantic units rather than defined metadata nodes we are also able to make no presumption as to the research structures, classification taxonomies, identifier types, format types etc that may be used and so create a schema that is extensible and re-usable beyond the initial Lifespan RADAR use case. We have a second proposed research data collection that can be used to test how re-usable the schema is.

Additionally, the PREMIS dictionary is primarily concerned with the preservation of digital objects and a research item collection that is intended for other researchers to use for secondary research will be concerned not just with the digital objects contained within, but perhaps also knowing something about the research context for the intellectual entity and also metadata and access to physical objects that have not yet been digitised. For example, in the PREMIS model events are only associated with objects and describe preservation acts, whereas we can also extend events to the research result itself to capture any research activities that may be of interest.

The resulting schema will therefore not be a pure implementation of any one approach but will adopt aspects drawn from research information management, research publication and preservation models to create a schema for a repository collection that is both interactive and accessible, and aware of the ethical and preservation issues around such a collection.

Therefore we have several reference models and metadata standards to draw upon, but none of which fully encompass our objectives with this collection. As such the aim of our proposed schema is to describe, curate and preserve records about a Research Item which can be understood in CERIF terms as a sub-class of Research Result and in PREMIS terms as an intellectual entity. The Research Item is a primary research data asset, created as a result of research activity and that is manifested in one of more digital or physical objects. We are beginning to form the ontological basis for such a Research Item, and the metadata schema, taxonomies and related repository functions necessary to implement this as a collection in Equella.

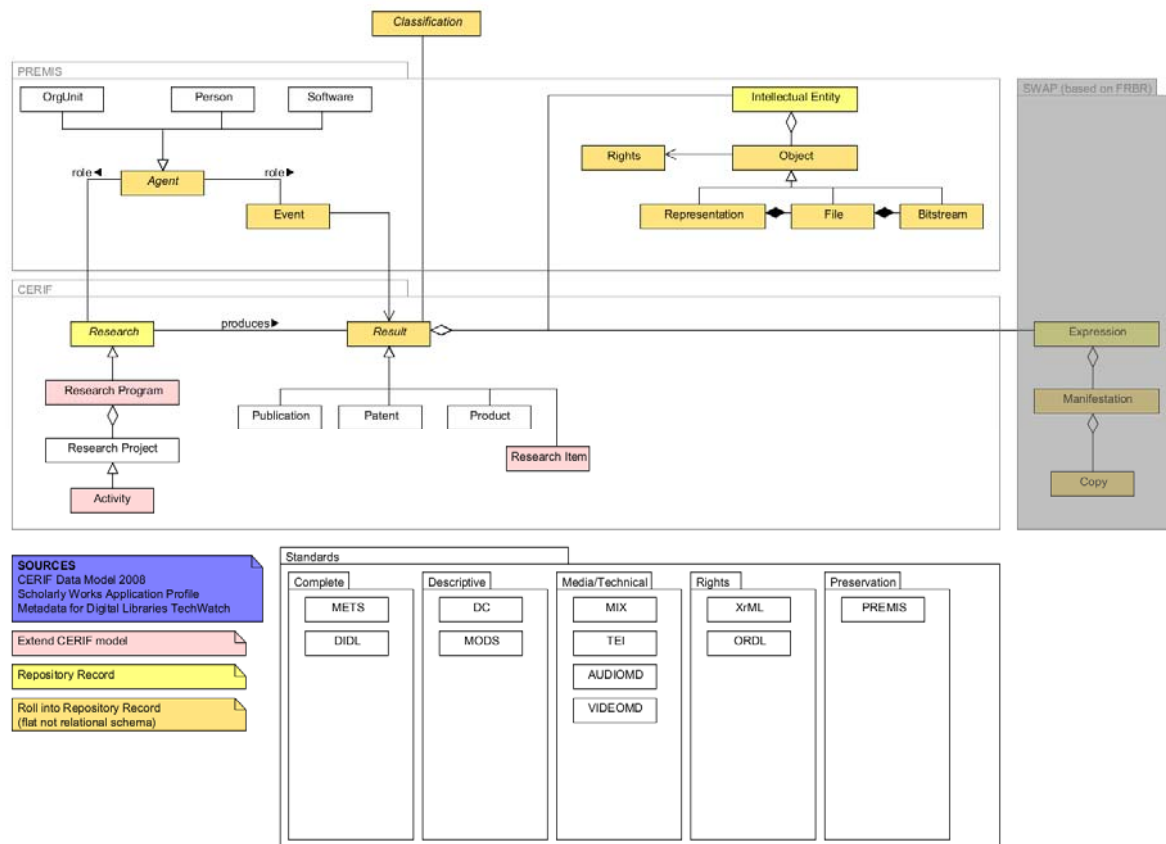


Figure 3: Research Item Ontology Model Draft 1

So far we have:

- An initial draft of the Research Item ontology
- An initial attempt at translating this into an Equella Metadata Schema

These will need to be refined as we learn more about the reference data models, metadata standards and best practice we are drawing on for this work and as we try out implementations of the schema using the Lifespan data and other research data collections.

Our experience has shown that a few iterations are normally required to find the appropriate balance between best practice and pragmatic implementation decisions and the level of detail required in the schema but we are now in a position to start building a Lifespan RADAR Proof of concept collection in the repository to begin the first iteration of this.

## References

- <http://www.loc.gov/standards/premis/>
- [http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly\\_Works\\_Application\\_Profile](http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile)
- <http://www.ifla.org/publications/functional-requirements-for-bibliographic-records>
- <http://www.eurocris.org/cerif/cerif-releases/cerif-2008/>