Sinopia Profiles Task Group Interim Report

October 1, 2020

Preface

October 1 was the original deadline for the final report of the Task Group, but has since been moved to November 1 at the request of the TG and in agreement with the Chair of the Policy Committee. The reasons for the delay are as follows:

- Sinopia is in the middle of a development work cycle. Among several other enhancements, this work cycle includes a complete re-working of how resource templates are created, accessed, and stored. Not surprisingly, major development work means that various components are temporarily broken, which is the case with template creation. This current work cycle was originally due to end October 2 and has now been extended to October 9. We will not be able to fully input the new profiles and test them thoroughly until after this date

- One of our task group consultants is currently out on medical leave. This has created a delay in working with the serials template as she is the coordinator of the LD4P serials profile and has the most experience there. This has slowed down things considerably

We would like to thank the Chair for recognizing our predicament and extending the date for the final report.

This current report will summarize our work thus far; much of its content will be repeated in the final report. The TG also is providing a separate spreadsheet which provides an overall view of the modeling of the monograph resource templates.

We are not expecting a great deal of feedback from this particular report. It is intended primarily to give PoCo a headstart in looking at the application profile for monographs and to learn some of the issues we are confronting in its development.

Introduction

The PCC Sinopia resource templates are built on the experiences and active participation of the Library of Congress catalogers who have participated in the LC Pilot over the past three years, and who have been active in providing feedback on the LC profiles and of the members of the

LD4 Profiles and Serials Affinity Groups, who worked over two years to develop the LD4P profiles and experimented with their own specialized templates. The Task Group would like to acknowledge their hard work and contributions to template development. Also, a big thank you to the LD4P developers, particularly Jeremy Nelson and Justin Littman, and to Michelle Futornick, the LD4P project owner. The TG charge is available on the PCC web site at https://www.loc.gov/aba/pcc/taskgroup/Sinopia-Profiles-charge.pdf.

The PCC Sinopia resource templates are based on:

1. The LD4P profiles developed by the Profiles Affinity Group and Serials Affinity Group during LD4P Phase 2. These were both intended to reflect PCC requirements, though sometimes diverge according to the needs of the LD4P cohort. These profiles, in turn, were based on the then current LC profiles, taking into account any changes that took place over the time of the grant.

2. Vocabularies used by MARC catalogers in their daily work, vocabularies developed by LC for BF cataloging and stored on id.loc.gov, and non-library vocabularies requested by LD4P cohort profile wranglers, and that were made available by Questioning Authority (QA) during the grant period. It should be noted that while Sinopia resource templates favor certain vocabularies through providing lookups in QA, other vocabularies may be added manually.

3. Requirements for the BIBCO & CONSER Metadata Application Profiles (MAPs) as laid out by PCC. This includes the use of RDA as a content standard and interpretation of MARC fields required by the MAPs.

The intention of these templates is to provide a structured core of resource templates that allow catalogers to create PCC-level descriptions with uniform modeling and a basic set of vocabularies. It is hoped that they serve as the basis for a formal PCC standard (as an extension to the current BSR and CSR) at some point and that in feeding the PCC data pool, serve as a pool of well-structured data to share, and provide vendors and developers data with which to experiment.

A separate spreadsheet provides the modeling for each resource template required for cataloging. This includes “primary” templates for work and instance and “referred” templates that model other entities, such as contribution, identifiers, etc. The monograph template is also available in Sinopia in the “stage” environment (stage.sinopia.io), though at the time of writing this report, it is incomplete. Sinopia development is in an active work cycle which includes a wholescale reworking of the template creation system and we are not able to add lookups. This should be fixed soon.

The report that follows is in four parts. First, the Task Group explains our choice of creating un-nested templates vs nested templates. Second, the TG highlights some modeling in the

templates that have been subjects of long term discussion at LC, in the LD4 Profiles Affinity Group, and in this TG. Third, we list areas in which Sinopia is lacking capability or creates modeling problems that will affect PCC use of the templates. (This will likely change as the current work cycle puts us in a dynamic environment!). Fourth, the TG lists desired changes to the BF ontology. These have been culled, not only from the LD4 Profiles Affinity Group experiences, but from the long term experiences of some TG members in LD4P and local projects.

Thus far, we have a complete (barring work cycle issues) template only for monographs. The template for serials is delayed by the medical leave of its primary author and our only person with good serials knowledge. We also are still discussing the Administrative Metadata template. The TG will provide these templates as part of the final report.

Nested vs Un-nested Templates

When the LD4 Profiles Affinity Group first started work on profiles, it was creating what are now known as “nested” profiles, meaning that either the instance template was embedded in the work template or the work template in the instance. This had the advantage that the cataloger dealt with just one template for cataloging. There are, however, some serious disadvantages in the current Sinopia environment.

1. The embedded template serializes as a blank node, and in the Sinopia environment, this currently means that it does not have a URI associated with it as an entity. It is possible to program Sinopia to create a URI for a blank node, but this has not been implemented. If PCC is interested in nested templates, a formal request to the developers via github for an enhancement is necessary.

2. The relationship between the primary entity of the template and the embedded template is unidirectional. If a cataloger starts with a work template, they can link from the work to the instance, but not the inverse without creating a regressive loop. This is a major weakness of the nested template. Resolving it would require inferencing of inverse properties, something that is not currently possible in the Sinopia environment.

Partly because of these problems, the TG decided to emphasize un-nested templates which avoid these issues. Also, in what the TG feels is an added benefit, un-nested templates emphasize that we are now describing entities, not creating “records”. Un-nested templates also encourage and facilitate re-use because of the entity-based approach they create.

The TG has said it will look into how Sinopia might better support nested templates. The new template environment, however, emphasizes individual templates over nested ones, particularly since the “profile” level has been removed from the templates. The TG needs to see the final result of the current work cycle to further comment on this.

Modeling Issues

Use of rdfs:label vs. rdf:value

The profiles created by the Library of Congress use a mixture of rdfs:label and rdf:value for literal values. The literal rdfs:label is used for notes and descriptive fields like bf:note and bf:statementOfResponsibility. The literal rdf:value is used for identifiers and other fields with structured values, following the W3C recommendation 5.4.3 (https://www.w3.org/TR/rdf-schema/#ch_value) in the RDF schema. For example, the bf:identifiedBy classes (bf:Issn, bf:Lccn, bf:Barcode, etc.) use rdf:value as do the bf:classification classes (bf:ClassificationLcc, bf:ClassificationNlm).

The use of rdfs:label vs. rdf:value is a wider question that remains unresolved in the RDF community at large. Because of this, the TG decided to follow the LC lead in these templates to provide a consistent practice that could be retained or easily changed in the future as best practices develop. The TG still needs to check that we are consistent with this application of rdfs:label vs. rdf:value.

Primary Contribution & Contribution

PrimaryContribution

The class bflc:PrimaryContribution is not part of core BF, but from the Library of Congress BF extension (bflc). Its primary function is to provide the equivalent of a MARC 1XX field both for discovery and for conversion to MARC. It is not necessary in BF itself, nor is it an RDA necessity, except of course for use in the MARC environment.

General Modeling issues

There has been a fair amount of discussion about whether to provide individual reference templates for the subclasses of bf:Agent (bf:Person, etc.) within the Sinopia template or whether to rely on the authority file itself to provide this on the fly when we have such capability. A direct lookup is a simpler model, especially if we restrict ourselves to LC NAF, but it also avoids unilaterally stating that URIs from other authority files are bf:Agents. On the other hand, it restricts possible search refinements by type of agent within Sinopia, prevents catalogers from creating local authorities with the type of agent defined, and makes it much more difficult to convert the BF data to MARC. While creating local authorities is less of a problem for PCC data, which requires established names, the other problems still remain. And given it would be best for bf:Contribution templates to be based on the same basic modeling, the local authority issue is still a problem.

Another problem with the modeling is the matter of repeatability. Sinopia does not allow a property to be repeated on the same resource template, since it makes it very difficult to recall a description back into a template or to convert the description. Because of this, both Primary Contribution and Contribution are accessed through the same property in the template (bf:contribution) and this property is marked as “repeatable”. This is problematic because a Primary Contribution should not be repeatable, given its role in MARC conversion. Some solutions to this problem are being discussed as part of the current Sinopia work cycle.

The TG has left the modeling as is, while acknowledging the problems with it. This would be a good conversation for PCC, in tandem with the comments on RWOs below.

Vocabulary

Thus far, both the Profiles Affinity Group and the Library of Congress have used the LC NAF authority record URI as the object of the bf:agent property. This is not correct, since a bf:Agent is an RWO, not an authority record. If PCC wishes to reference Wikidata or ISNI, both RWOs, an alternate model of bf:Contribution is needed and should be made available to the Wikidata Pilot members working with Wikidata and Sinopia. The TG is working on an experimental version, which will be provided in the final report.

It should be noted that the Library of Congress, in developing the LC “Hubs”, has developed behind-the-scenes programming that provides both RWO and authority record URIs.

Series

Series relationships appear in both the bf:Instance and the bf:Work.The bf:Instance template contains the property bf:seriesStatement. It is intended to include all the metadata that might be expressed in a more granular fashion with the properties bf:seriesStatement, bf:seriesEnumeration, bf:subseriesStatement and bf:subseriesEnumeration.

The reason for keeping them together is that there is no other way to logically group together multiple series statements with multiple enumerations since they are all expressed as literals.This same modeling was adopted by both the Profiles Affinity Group and the Library of Congress.

There are, however, a few drawbacks:

1. In order for the statement to read easily, the cataloger needs to add ISBD punctuation to separate the sections. This is not exactly a drawback, but a bit out of place in a linked data environment. It could potentially help, however, in converting these statements to

the MARC 490 field and its subfields

2. If an ISSN is present, it is to be transcribed as part of the bf:seriesStatement or bf:subseriesStatement. This makes sense logically, since the ISSN is for the series, not the individual monograph. However, since it is transcribed as part of a string, it is not readily available as a searchable identifier. This needs some more thought.

The bf:Work template contains the property bf:hasSeries. This is a work-to-work relationship, as indicated in both FRBR & IFLA-LRM. Because it is a work-to-work relationship there is no enumeration, since that is an instance-level property.

This relationship, while in RDA, is not actually used currently by PCC. We instead use the 8XX, which gives the series plus its enumeration for that particular volume. The PCC Aggregates in Beta RDA Task Group recently recommended that this particular relationship not be used and to instead continue with the 490/830 combination, and that recommendation was accepted by PoCo. This makes sense in MARC but is more difficult in BF, since without it there is no mention of the series in the bf:Work, except perhaps an identifier for the ISSN.

Sinopia Issues

- ***Support for resource template version control***
  Until this current work cycle, there was no support at all for version control of template; not even the creation date could be updated in the index. Both the LD4 Profiles Affinity Group and LD4P cohort in general felt that version control was important and should be on the development list. The Affinity Group did experiment with adding a version property to the JSON-LD, but it was not recognized by Sinopia and caused the template to fail. Further testing is needed to identify a template version control strategy that works in the new template creation and management context.

- ***Support for Library of Congress complex subject headings***
  If PCC wants to make full use of LCSH combinations of headings, there will need to be development to support it. Sinopia is currently only able to provide subject components through lookups to the LCSH vocabulary. It is not able to combine headings or put them in a particular order. Complex subjects in id.loc.gov and the LC BF editor use a madsrdf:componentlist property to keep the subheadings order, but Sinopia Resource templates don't support a property type list. Alternative approaches to this are currently being explored (e.g. https://github.com/LD4P/sinopia_editor/issues/2358). There is also a new "ordered" attribute in the Sinopia template creator which needs to be explored.

- ***Support for repeatable properties when they point to different Resource Templates or one points to a resource template, but the other is preferably a direct lookup***

The primary use case for this right now is to allow a property to be repeatable when pointing to one referred resource template and non-repeatable when pointing to another, as in the case of the Contribution (repeatable) and Primary Contribution (ideally non-repeatable) templates. It would also be useful when the modeling between the two predicates is different,something that seems fairly likely when we get to music and a-v resource description, but not currently possible.

- ***Controlled vocabularies missing from QA Service***
  While Sinopia and its lookups via Questioning Authority were being developed, the Library of Congress began adding multiple vocabularies to id.loc.gov. Many of these are not yet available to Sinopia template creators. While all of these vocabularies will eventually be required, for monographs we are missing the following:
    - Script (https://id.loc.gov/vocabulary/mscript.html)

Bibframe Issues

Below are listed changes to the BF ontology that the TG suggests to support PCC cataloging (and cataloging in general).

- **Remove ranges from numerous BF properties**
  The addition of ranges to BF properties introduces challenges regarding inferencing and linked data best practices. At a practical level, a range may restrict the breadth of possible vocabularies that are usable according to best practice. For example, catalogers might use AAT terms to describe a form or genre. These terms, however, are not bf:GenreForms.

    - Range of bf:GenreForm should be removed from bf:genreForm
    - Range of bf:Language should be removed from bf:language
    - Range of bf:Place should be removed from bf:place
    - Range of bf:Content should be removed from bf:content
    - Range of bf:Media should be removed from bf:media
    - Range of bf:Carrier should be removed from bf:carrier
    - Range of bf:GeographicCoverage should be removed from bf:geographicCoverage
    - Range of bf:IntendedAudience should be removed from bf:intendedAudience

- **Changing certain data properties into object properties**
  The data property bf:temporalCoverage should be changed to an object property (range of rdfs:literal should be removed from bf:temporalCoverage, and add a class for bf:TemporalCoverage)
    - The creators of BF originally made this a literal, probably because of a seeming lack of a controlled vocabulary. Such vocabularies do exist however, at least

inpart (e.g., FAST provides temporal vocabulary), and BF should allow for their use while still permitting use of strings.