SVDE PCC/Sinopia PCC Template Analysis, March 2021

Here you will find documentation on analysis comparing and contrasting shapes found in the SVDE PCC data with decisions made in the Sinopia PCC Templates in March 2021.

Context and Caveats on the Analysis

1. A spreadsheet was programmatically created by Justin Littman (Stanford) from the Sinopia PCC Templates according to the DCMI Tabular Application Profile specification. A copy of the spreadsheet (March 2021 Comparison of PCC Sinopia Templates and SVDE Data) was used to evaluate row by row if SVDE PCC data has triples conforming to the template.
    a. No systematic attempt has been made to perform analysis from the other direction (to see if every pattern in SVDE is reflected in the Sinopia PCC templates), but there are undoubtably SVDE patterns absent from the Sinopia templates.
2. Analysis of the SVDE PCC data was performed using SPARQL queries on the LD4P cache available at: http://services.ld4l.org/fuseki/dataset.html?tab=query&ds=/PCC.
3. Every attempt was made to account for different patterns for bf:Works in the SVDE data, but the patterns may be too different to reconcile neatly with the Sinopia PCC Works patterns. This reflects the need for the community to decide on similar patterns for reflecting the WEMI model in BF, assuming this is a goal.
4. The bflc:Relationship pattern to capture relationships (between Works and other works, and Instances and other Instances) in the templates and in the SVDE data are not intuitive. There should be further analysis and community discussion about the extent to which the bflc:Relationship pattern should be used and/or avoided.
5. The property bf:issuance seems to not be consistently applied for Serials in the SVDE data, and perhaps Monographs. This made it difficult to confidently say the SVDE data complied with the PCC templates usage of http://id.loc.gov/vocabulary/issuance/mono and http://id.loc.gov/vocabulary/issuance/serl for monographs and serial respectively as bf:issuance value. We need to solidify what pattern the community should use to confidently signal this information.

How SVDE Data and the Sinopia PCC Templates Differ

There are significant differences between the Sinopia PCC templates and SVDE data. In March 2021 Comparison of PCC Sinopia Templates and SVDE Data, Column L (with the header "SVDE Differences") captures compatibility and differences between the data and templates. Areas of compatibility are highlighted green, and differences are highlighted yellow. It may be that not all differences require action.

The type of differences include, but are not limited to:

- SVDE data uses URIs for entities that Sinopia template creates blank nodes for.
- Sinopia template allows for URIs or literal, while SVDE only uses URIs.
- BIBFRAME class/subclass usage doesn't line up exactly (e.g. bf:Titles and bf:Instance subclasses)
- SVDE data does not use http://sinopia.io/vocabulary/hasResourceTemplate to connect entities to templates.
- Properties uses in templates are not found in the SVDE data.
- Properties used in the SVDE data are not found in templates.
- SVDE data links to values (often from id.loc.gov) seemingly not permitted by the sinopia:LabeledResource template.

SVDE/Sinopia Dataflow Questions

We need to clearly define what actions we expect to be able to performed on the SVDE data in the Sinopia environment and vice versa. Depending on the answers to the following questions, we may need to consider changes to the tooling, data outputs, and/or templates.

1. Are we only committed to linking to or deriving new Sinopia descriptions from SVDE data?
    a. If so, is it ok to only map in the SVDE data that aligns with the Sinopia templates? Data that doesn't map neatly could also be returned and visible to the cataloger so that values can be copied and pasted to applicable Sinopia fields.
        i. Alternatively, data that doesn't map neatly could be transformed to the Sinopia template shape where possible, but this would require QA and other potential lookup tools to understand both the source data and the template shapes. This would be a remarkable amount of work.
2. Do we expect to be able to edit SVDE data from within Sinopia?
    a. If so, is it ok if not all parts of the SVDE data is editable from within Sinopia? Are we ok with open shapes ("extra" data beyond what the PCC templates address)?
        i. If so, we need to be able to identify acceptable differences.
    b. If so, (and assuming changes will be sent back to SVDE) is it ok if Sinopia templates afford additional patterns not represented in the SVDE data? Is SVDE ok with open shapes ("extra" data beyond what their tooling may be set up to interpret)?
        i. If so, we need to be able to identify acceptable differences.

More Broadly

In the absence of community shared shapes, it is understandable that tools attempting to interact with external data sources will struggle to account for differences in data shapes. The attempt to consume SVDE data in Sinopia according to a first pass at PCC templates is a great opportunity to test both the templates and modeling decisions being made in the SVDE community. Ideally, PCC's nascent attempt to create and maintain a PCC MAP will be informed by this learning opportunity. As the templates are further vetted, changes should be made through official processes, and communicated through official channels so that data providers can fold these decisions into their outputs.

Looking to the near future, SVDE intends to make changes to their data based on feedback from the SVDE community this summer. Another round of analysis will need to be performed when those changes are made. Rather than relying on bespoke SPARQL queries that require human interpretation for this analysis, resources should be allocated to develop more expertise in validation and reporting tools using ShEx or SHACL. We should consider whether Justin Littman's (Stanford) proof-of-concept Sinopia RDF validator using DCTAP profiles is fit for this purpose, or if other tooling is needed. Further, if the PCC is committed to a spreadsheet representation of MAPs, it may want to consider conforming to the DCTAP specification to more easily generate ShEx and/or SHACL for validation.