

2021-05-28 Cornell LD4P3 Meeting notes

Date: 28 May 2021

Attendees: Greg, Tim, Lynette, Huda, Simeon, Jason, Steven

Regrets:

Meeting time after July 9 (or some after change)

- Agree keep to 9:30am Friday

Discovery (WP3)

- <https://github.com/LD4P/discovery/projects/2> for issues etc.
- Draft of a discovery plan: https://docs.google.com/document/d/1zKYW7FQVVNvyd0XjW0qWznX9PC3jbmOE6Kz_yygPjs/edit?usp=sharing
- **Research: how to go from knowledge graph to an index**
 - [Research decision points](#), [Use cases](#)
- DASH! (Displaying Authorities Seamlessly Here)
 - Dashboard design meeting kickoff [notes](#)
 - User reps D&A meeting: Expect next follow-up in August ([Slides](#): from user reps meeting 2021-04-09 and result was "not no")
 - https://docs.google.com/document/d/1PgQi3xobsPhr9DUHU_YGeimL1OjNiiTdkinWb36r3Gg/edit
 - Usability testing and followup for DASH: [Usability results](#)
 - 2021-05-21
 - [Usability results](#), a few little things to finish up
 - [GitHub issues](#)
 - 2021-05-28
 - Tim looking at how to handle pages for entities with little or no information, required significant page reworking
 - Tim also looking at definitive list for timeline
 - From slack
 - I haven't worked on Bang! this week since focused more on Dash!. (a) merged in latest from D&A catalog code as of earlier this week into dev branch of our fork. Will merge into dash work branch later (b) Started looking at knowledge panel not showing view full record link/styling. Panel gets the full record link from Ajax request to catalog subject browse (d&a version) and then replaces with link. Display logic for D&A browse does not include full record link so it isn't shown in our version. Relates to broader approach for subject knowledge panel i.e. using browse page to get content. Worked fine for demo approach but will rework to either use more discrete queries or include more control over content. (c) started discussion with Tim on other usability related issues. Resolution of some issues, such as clarifying role of data sources like repositories and digital collections, also relates to making design of author and subject pages consistent. Tim took on mockup creation to address possible designs. One main difference between current author and subject dash pages is that author page only shows result numbers for various sources while subject page shows first set of results in tabs for sources. Mockups would look at option for subject page more aligned with author page design as well as a mockup in the other direction i.e. showing results in the page for authors as well. Larger philosophical discussions in this area as well. Will also be revisiting timeline and map coordination and what else we may show for related subjects on timeline.
 - How long to continue working on DASH! ? Tim and Huda to discuss what could be done by mid-July, discuss next week
 - Video for DASH!, theme?
 - Sonic? Roadrunner?
- BANG! (Bibliographic Aspects Newly GUI'd)
 - [Jamboard link](#)
 - Expect to include Works. Need to do something beyond what we already have live from the OCLC concordance data.
 - Full OCLC concordance us 343M rows, and gzipped the file is 3.3GB
 - SVDE Works
 - 2021-02-26 Have to develop SPARQL queries to pull out certain sorts of connected Work. Don't expect data to be very dense but do expect that we would get useful connections between print and electronic for example. We already have a link based on the OCLC concordance file from several years ago.
 - ACTION - [Steven Folsom](#) and [Huda Khan](#) to work on building an equivalent of the OCLC concordance file based on SVDE data and then do a comparison to see how they are similar and different
 - 2021-04-02 Steven and Huda met to think about putting together queries to extract a similar dataset. ([Document for recording queries](#)). Open questions about the counts – got 16k works from one view, got about 8k where limited to case with at least one instance. These numbers are much much lower than expected
 - 2021-04-16 Steven working with Dave on how to pull our SVDE data. Dave still working through some errors in ingest of SVDE data – this needs to be resolved before looking for concordance. Has asked Frances for 2015 concordance
 - 2021-04-23 Waiting on indexing of PCC data, have learnt more about the basis for the old OCLC concordance file
 - 2021-05-07 Steven didn't have much luck getting data from SVDE, learning GraphQL endpoint but also problems with timeouts there (HTTP 503)
 - What is the space of Work ids that we might use and their affordances?
 - OCLC Work ids, SVDE Opus (Work), LC Hubs (more than Hubs), what else?
 - Connections to instances, how to query, number
 - 2021-05-07 ACTION - Huda to start analysis
 - Other SVDE entities
 - 2021-05-07 ACTION - Huda will reach out to Jim Hahn about entities other than Works represented in SVDE - DONE
 - Summarized here: [Jamboard link](#) - U Penn Enriched Marc: Work Ids in 996 Field. 1.2 million with OCLC Work IDs in > 1 description. ~3.9 million with OCLC Work IDs in only one record.

- Publisher authorities/ids
 - At Cornell we haven't tried to connect authorities with publishes
 - LC working on connecting to publisher identifiers - utility is things also published by a publisher
 - Also possible interest in series and awards
 - 2021-04-23 Might be able to use LC publisher ids in BANG!, Steven will look at whether there is a dump available
- 2021-05-21 - Continuing to work on all the above to find out what works data will allow us to develop in BANG! something more than we already have from the concordance data
 - ACTION - Steven to look into OCLC work data. Perhaps ask rep for ways to access/query - 2021-05-28 DONE: Steven found that WorldCat has removed the links to the RDF because the data was getting stale. Wasn't being used. Expectation that any work on Works would be folded into Entities project
- DAG Calls
 - 2021-05-28 Had high level topics overview discussion. Interesting comments and philosophical discussions about the benefits of linked data, demonstrations/examples that are useful to cite, BIBFRAME value proposition for discovery. Going forward plan to look more at user research and then work through high level areas

Linked-Data Authority Support (WP2)

- [Qa Sinopia Collaboration](#) – Support and evolve QA+cache instance for use with Sinopia
 - 2021-05-28
 - We began discussing the document describing interaction patterns between Sinopia-QA/cache-ShareVDE. The key take away from the discussion is that there needs to be clarity about how the PCC data and Stanford Institution data are expected to be used. There are some key questions to be answered that will drive the technological solutions in Sinopia, QA/cache, and ShareVDE. They include:
 - How is data initially ingested into ShareVDE?
 - Once initial ingest is complete, how is new data added (e.g. incremental ingests from original ingest source, newly created entities in Sinopia)?
 - Where is the Source of Truth (e.g. ShareVDE, Sinopia)?
 - How are updates made to entities? The short answer is that edits are expected to happen in Sinopia. But the details of how editing will happen is highly dependent on the answer the Source of Truth question.
 - **We have to recognize what is possible now or soon, with shapes/connections/etc..** Sinopia and SVDE data shapes are too different for it to be realistic to edit SVDE data in Sinopia. This leaves us with related but only weakly connected pools of PCC data in Sinopia and SVDE – what does it then mean to search PCC data? One would need to search both Sinopia and SVDE. One can derive a new description in Sinopia based on a "starter record" from SVDE but it will not be possible to edit SVDE data in Sinopia
 - Do we need to start distinguishing between Sinopia PCC data and SVDE PCC data?
- [Best Practices for Authoritative Data working group \(focus on Change Management\)](#)
 - 2021-05-28
 - Second meeting this past Monday. We continued discussing [type of changes](#). No new types were added. We began the process of discussing what information is needed in the change management stream for each type.
 - We talked some about NEW entities identifying two options. We did not make a decision on which approach is preferred. (*NOT E: Format here is to convey data and is not necessarily the final recommended format.*)
 - { 'type': 'NEW', 'URI': 'https://uri.for.new.entity' } with that, the downstream consumer can dereference the URI and use the results to add the entity to the cache
 - { 'type': 'NEW', 'URI': 'https://uri.for.new.entity', 'entity': { json-ld for new entity } } with that, the downstream consumer can use the entity in the change management stream to add the entity to the cache
 - We spent a good bit of time on CHANGE LABEL which turned out to be more complex than expected. The purpose of this type is to facilitate applications that cache labels for quick display in the application or for indexing labels to facilitate search. Again two options were identified.
 - { 'type': 'CHANGE_LABEL', 'URI': 'https://uri.for.new.entity', 'predicate': 'skos:primaryLabel', 'OLD_LABEL': 'old literal'@en, 'NEW_LABEL': 'new literal'@en } with OLD_LABEL being optional. Without OLD_LABEL, this is a new label. With it, the OLD_LABEL is being replaced with the new label. Applications can search their caches for the OLD_LABEL triple and update it to the NEW_LABEL.
 - { 'type': 'REMOVE_LABEL', 'URI': 'https://uri.for.new.entity', 'predicate': 'skos:primaryLabel', 'LABEL': 'old literal'@en } { 'type': 'ADD_LABEL', 'URI': 'https://uri.for.new.entity', 'predicate': 'skos:primaryLabel', 'LABEL': 'new literal'@en } I have a question about how the application will process the change management stream. It will need to know that these two change management documents are related. This is fine for a full cache, but may not help applications that are only caching the labels.
- [Cache Containerization Plan](#) - Develop a sustainable solution that others can deploy
 - 2021-04-30
 - Have worked on permissions issues and documented how to implement in AWS
 - Greg now running out of things to do without more input from Dave. Can document existing work and develop presentation for conference
 - Consider moving live QA instance from EBS to container version? Need to consider update mechanisms CI/CD. Agree that this is a good direction and Greg/Lynette will discuss
 - 2021-05-28
 - Last week Greg solved some issues. Still working on documentation and still working on CI/CD. Have abandoned idea of using github actions so working with Jenkins instead
 - No progress on work to containerize DAVE, Dave is focused on the SVDE work

Developing Cornell's functional requirements in order to move toward linked data

- C.f. Stanford functional requirements document: https://docs.google.com/document/d/18H6zYGwKuG3SZqm9Q_cxkZThcdmBjknE6HdtQ-RRzk/edit#heading=h.4fu64x8jzm6e
- What does success look like? And then how do we get there?
- Miro board (diagramming): https://miro.com/app/board/o9J_ifXUJ8=/
- Notes space: <https://docs.google.com/document/d/1TVPBFak7DkfjBptKl-pCMWQnOaiWHB0XCHswiB3Fr9g/edit?usp=sharing>
- Purpose: Vision for mid-term (3-5 years) transition to support linked-data at Cornell. May include things we don't yet have or cannot yet do, but not long-term vision of post-MARC environment

- Important to understand sources of truth (primary data) and where there is derivative data
- Imagine landscape with items described in multiple formats including at least MARC, BF, DC (eCommons), JSTOR
- Imagine all items indexed and discoverable via D&A
- 2021-04-23 Working in Miro board

Other Topics

- PCC/Sinopia and SVDE shape analysis
 - 2021-03-19 Steven has been working through a spreadsheet of 400+ lines to compare the shape of SVDE data with the PCC/Sinopia profile. He is finding that there are many many differences which will severely limit how well Sinopia will be able to consume and edit SVDE data. For the purposes of QA/Sinopia cloning, Steven could come up with some ldpaths but not sure whether the amount of data will be useful. Steven expects to be able to share the spreadsheet at the next Sinopia/SVDE meeting. Going forward we need to consider the role of versioning/documenting shape changes and validation at both scale and single descriptions. Justin's validation scripts: <https://github.com/LD4P/dctap>. Tom Baker's csv2shex: <https://github.com/tombaker/csv2shex>
 - 2021-03-26 Steven finished working through the spreadsheet comparing SVDE data with the PCC profile. Notes that he is looking only from the side of the PCC profile and would thus miss other things in SVDE data. Patterns around different types of work in SVDE data (e.g. Opus and other higher level works have very different shapes). Difficult pattern of double-reified relationships between works. Steven will let SVDE/QA folks know about completion of the work. Need to find a way toward alignment.
 - 2021-04-23 Write-up complete: [SVDE PCC/Sinopia PCC Template Analysis, March 2021](#)
- OCLC Linked Data / Entities Advisory Group
 - 2021-04-02 Michelle asked about connecting QA to the OCLC Entity Backbone as part of updates for partner meeting, Lynette has reached out about API
 - 2021-04-23 Still nothing back from OCLC, need to reach out again
- PCC Task Group on Non-RDA Entities
 - 2021-03-19 Group headed by standing committee on standards will formally propose a list of non-RDA entity types. Steven will join. Deliverables by June
 - 2021-04-02 Many participants involved in ILS/LSP migrations so work delayed until July
- Default branch name - Working through repositories in [Renaming of LD4P Repositories](#)
 - Created [Renaming of LD4P Repositories](#) page to identify Cornell repos, provide instructions, and track progress
 - 2021-04-23 Waiting on <https://github.com/cul-it/blacklight-cornell> to change, not yet planned
- Authorities in FOLIO
 - Hope to include URIs as part of Cornell FOLIO migration, possible LD4P work
 - 2021-04-23 Likely going ahead in August

Upcoming meetings

- <https://kula.uvic.ca/index.php/kula/announcement/view/1> . Call for Proposals - Special Issue: "The Metadata Issue: Metadata as Knowledge". Due January 31, 2021 (abstract 300-500 words). Includes "The use of linked open data to facilitate the interaction between metadata and bodies of knowledge" and "Cultural heritage organization (libraries, archives, galleries, and museums) and academic projects that contribute to or leverage open knowledge platforms such as Wikidata"
 - Folder [Link, CFP + Brainstorming](#)
 - 2021-03-26 Abstract accepted, paper due June 25
 - 2021-05-07 Meeting scheduled for 5/13 with group. [Document to capture discussion/references](#).
 - Research paper format: "**Word count**: Research articles should be between 6,000 and 9,000 words (including notes but not references)". Peer reviewed. Have emailed KULA folks just to confirm type of article.
- [LD4 Conference 2021](#) - response on proposals will be announced April 30; conference is July 12-23
 - Discovery - suggestion of discussion form
 - Huda submitted status of discovery update, BOF for DAG, and discussion of opportunities for effects and use of linked data (60min discussion)
 - 5/7: Received word that affinity groups will be getting reserved time they can use as desired, so may go ahead with BOF format (will need to discuss with co-chair/DAG members)
 - Lynette/Greg/Dave - containerization, should have documented product by then
 - 4/9: Submitted abstract for presentation
 - 5/21: Accepted
 - Lynette – possibly something about the working group, perhaps updated version of code4lib
 - 4/9: Submitted as a lightning talk
 - 5/21: Accepted
 - [Document for brainstorming](#) (in case anyone wants to use it)
- BIBFRAME in Europe workshop - September 21-23 15:00–18:00 CEST = 9am-12noon EST
 - Call is out <https://www.casalini.it/bfwe2021/>, proposals due June 30
 - PCC Workflow Simplified view?
 - FOLIO integration?
- SWIB virtual again this year, call for proposals out, due June 28
- Lynette doing a QA presentation at Samvera partner call in June

Next Meeting(s), anyone out?:

- 2021-06-04 Tim out ...