2021-05-07 Cornell LD4P3 Meeting notes

Date: 07 May 2021

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

Regrets: Steven

Discovery (WP3)

- https://github.com/LD4P/discovery/projects/2 for issues etc.
- Draft of a discovery plan: https://docs.google.com/document/d/1zKYW7FQVVNvyd0XjjW0qWznX9PC3jbmOE6Kz_yygPjs/edit?usp=sharing
- **Research: how to go from knowledge graph to an index**
    - Research decision points, Use cases
    - First goal: DASH! dashboard (full page for entity) that extends on the idea of an embedded knowledge panel, aim to have functional prototype for end of year
- DASH! (Displaying Authorities Seamlessly Here)
    - Dashboard design meeting kickoff notes - will also try to understand what our data will support or connections to other data sources
    - https://docs.google.com/document/d/1PgQi3xobsPhr9DUHU_YGeimL1OjNiiTdkiNWb36r3Gg/edit
    - Usability testing for DASH
        - 2021-04-16 - Usability testing is underway, will finish next week. Useful input on what users see/understand and what they struggle with. Users don't know what a repository is, can guess what digital collections means. Users tend to think that the author name link will get more information, don't see/understand info button – could mean adding tool tips and/or making button more obvious or even flip button/link or one link for info+search
        - 2021-04-23 - Five people helped with usability tests. Setting up final results document. High level:
            - Uncertainty of info button vs author link (or lack of understanding if same/separate). Tim has created issue in D&A
            - Good understanding of author page in general, opportunities for some clarification in labeling etc., influences tab well understood
            - On subject page the timeline scroll/zoom fast creating interaction problems, broader/narrower mixed understanding (difficult when subject is an event), "show all subject" checkbox not found easily, timeline/map connection was understood
        - 2021-04-30
            - Writing up usability test results
        - 2021-05-07
            - Nearly done! Tabulation across participants
            - Will be adding recommendations as github issues: https://github.com/LD4P/discovery/projects/2 and work on straightforward issues in the next month
    - User reps D&A meeting: Expect next follow-up in August
        - Slides: from user reps meeting 2021-04-09 and result was "not no"
            - Positive feedback on linking back to catalog results
            - Questions about links and display from wikidata, other sources
            - End was to discuss more in August with possible discussion of how to move into production
            - User reps have slides and wanted demo link
            - Could have lite KP might to start, possible later entity pages – or perhaps both together
        - 2021-04-23 How much more work to do before August? Candidates:
            - Follow up on results from usability tests
            - Work on how smoothly the prototype works
            - Think about more consistent look between pages - Tim thinking of making some mockups of redesign
        - 2021-04-30
            - Working on a list of tasks for the final refinements stage. Includes aligning look & feel of the entity pages, Tim is looking into this. Current subject pages optimized for things with date ranges, what do we do for subjects without date ranges (e.g. microbiology)?
    - Video for DASH!, theme?
        - Sonic? Roadrunner?
- BANG! (Bibliographic Aspects Newly GUI'd)
    - 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
  - 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
  - DAG Calls
    - 2021-05-07 Next week talking about visualization but then planning out sequence of topics: Framework for thinking about discovery questions, possibilities within data, etc.

Linked-Data Authority Support (WP2)

- Qa Sinopia Collaboration – Support and evolve QA+cache instance for use with Sinopia
  - 2021-04-30
    - Discussed the potential to include non-RDF authorities. Not difficult on the QA side, but Sinopia team is concerned about the difficulty in supporting on their side. The earliest they could work on it is Fall 2021 and even then it would have to be prioritized.
      - The non-RDF authorities that people want are ISO language and script (e.g. ISO 639-3, ISO 15924). There is a data property in BF that LC use with literal values (not URIs). Sinopia doesn't have a facility for lookups from a set of literal values, QA could support this but Sinopia wants RDF from QA only.
      - For rare materials there will be a similar issue with RBMS vocabs that aren't RDF
    - Discussed general priorities for Sinopia. Used the priorities spreadsheet (under working group section) as the starting point. Then looked at issues already in Sinopia which represent their expected priorities.
      - Key takeaways:
        - Sinopia fully supports 4/7 of Priority 1 user stories, partially supports 1, the remaining 2 are performance /timeout related which are at the cache/QA level and both are improved by the latest indexing approach.
        - Sinopia fully supports 2/7 of the Priority 2.
        - There are 4 others that are either fully or partially supported.
        - There are 14 not supported.
      - Potential work we've discussed in the past:
        - Support for left anchor search, exact value search, and allowing user to select the approach (including existing keyword search).
          - Jason notes that Laura Daniels has done lots of work around inventory search in FOLIO. Can do both left anchor and keyword based on special syntax.
          - In D&A one can do search on subject or subject a-z
        - Improved functioning for broader/narrower where the user can select a broader/narrower term or use one of them as a new search.
        - Advanced search with passing in specific fieldname:value to find results and filtering of results by field value or date range.
        - ACTION - Ask Sinopia team for explicit feedback on priorities – 2021-05-07 still to do
    - Dave's new indexing is there, accuracy and performance is improved though we aren't sure about timeouts
- Search API Best Practices for Authoritative Data working group
  - 2021-04-30
    - I put together a spreadsheet that looked at which priorities from the first charter are in Sinopia now.
  - 2021-05-07
    - First meeting May 10, same time slot as old group. Will look at use cases first (keeping cache in sync, keeping cached labels in sync in application, URIs cached in application)
- 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
  - 2021-05-07
    - Greg spent some time learning about github actions for CI/CD. Has some permissions issue with github repo. Expect to continue working on this today and hope to be able to redeploy container when content changes

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

- C.f. Stanford functional requirements document: https://docs.google.com/document/d/18H6zYGwKuCg3SZqm9Q_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_lfXUUj8=/
- 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
  - 2031-05-07 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
  - Lynette – possibly something about the working group, perhaps updated version of code4lib
    - 4/9: Submitted as a lightning talk
  - 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?
- Lynette doing a QA presentation at Samvera partner call in June

Next Meeting(s), anyone out?:

- 2021-05-14 Cancel - too many out