2021-03-05 Cornell LD4P3 Meeting notes

Date: 05 Mar 2021

Attendees: Greg, Tim, Lynette, Steven, Jason

Regrets: Huda

Last meeting: 2021-02-26 Cornell LD4P3 Meeting notes

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
- **Strand 1: production piece**
    - Production requirements and functionality – Production decision points
    - **Discogs** data use - in production since January 2021
        - **ACTION (sometime after March D&A sprint)** - Tim to follow up on implementation and look at data from tracking use of the Discogs. There is an issue to start collection of data after the sprint by integrating the data collection in the sprint work, will need to wait long enough to have data to analyse
        - **ACTION (during March D&A sprint)**: Tim Worrall will raise usability testing for D&A queue (don't carry forward in notes as now outside of LD4P)
- **Strand 2: 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
        - 2021-01-29: Huda working to get scripts in place to populate index; bringing in period-O info; focused on locations with Wikidata URIs for consistency. Subject headings: script that takes-in components & breaks those out... and parses into timeline info. On Dave's fuseki, 34 distinct temporal terms with labels. Will finish today with actual index. Will break-down the loading to increase load speed
            - Reached out to IRB to ask about testing: if we want to disseminate results as research data, need to do IRB protocol; has a follow-up. Waiting to hear back but will submit protocol if no word. Simeon's interpretation of reply is that we are crossing line into research and the approval will likely be positive. Depending on how we describe what we're doing it either falls under research OR improving a product... but we're essentially doing research to improve a product so yes to IRB review.
                - **ACTION ITEM**: IRB did respond and say they wish us to proceed with sending in an application. Huda will work on this and reach out with any questions if needed.
                    - 2021-02-26 After some discussion and clarification there remain to submit details of consent forms, addition of potential interview and focus group questions, de-identification data, and compensation information
                    - 2021-03-05 Huda sent more replies to IRB
        - 2021-02-19 Tim has been working on entity page. Notes a number of issues with the Historopedia timeline such as items with same date being hidden, but performance is good
            - 2021-03-05 Tim resolved a number of issues. Next week will return to work on this and deal with influence-for and influenced-by presentation
    - 2021-02-05
        - What would D&A user reps favor?
            - Concern that full KP linked from  button is too much
            - Is "KP-lite" on autosuggest a good route? We think users would find this valuable. Are there options that minimize index changes?
            - What warrants a KP?
            - What is the redundancy between KP work and DASH!? Does dashboard mean a fundamental change or is just an enhanced KP?
            - We need to be aware of which options require significant indexing changes. There is already a sense that we want to add ids to the index
            - What about the open syllabus project? This relied on the open syllabus API, not sure whether it is available in LD. Essentially a mapping from domainCSIP codes ISBN, very few wikidata connections
        - What would be the smoothest next step for production?
        - Which option would give us real linked data connections via URI?
        - Steven notes that LTS authorities in FOLIO group is looking at the insertion of URIs into MARC records (resources willing)
        - ACTION - Huda Khan Tim Worrall document options and implications as preparation for user reps presentation in order to get a steer on where to continue experimentation with a view to future implementation
            - Dashboard (perhaps for some entity types only)
            - Autosuggest with KP-lite
            - Regular facet with KP
            - Open syllabus related items
            - Brainstorming notes
            - 2021-02-12 Agreement that streamlined KP is a good starting point, with possibility of later extension to a full dashboard. Autosuggest and open syllabus good alternative options.
        - ACTION - Huda Khan to line up meeting with D&A user reps
            - 2021-03-05 Understanding that we aren't going to be asking for review of anything to be deployed before the FOLIO go-live. User reps are happy to provide us with guidance for ongoing development

    - Tim on ESMIS for next weeks, Huda working on IRB and also looking at dashboard with new version of historopedia which is much faster. Huda also looking at avenues for recruitment, have found out about student worker lists for Olin and for Mann, and grad carrel users list
    - Planning for discovery work
        - Work so far has focused on authorities and what we can do in catalog
        - How might we use BF modeling and data from SVDE? At DOG meeting on Monday there was discussion, also similar discussions in DAG about specific use of modeling

Linked-Data Authority Support (WP2)

- Qa Sinopia Collaboration – Support and evolve QA+cache instance for use with QA
    - 2021-03-05
        - Discussed the status of ShareVDE. Dave is still working on cleaning up the data. He has a subset that are in Fuseki for experimentation. Action items (in order): Dave will continue to clean and index the data, Steven will compare PCC templates in Sinopia with PCC data in the cache, Steven will define the shape of data required for extended context and for a single URI dereference, Dave will create a query API based on the results of Steven's exploration, Lynette will create a QA config, Jeremy or Justin will connect the QA config to Siniopia. At that point, it will be ready for exploring search and clone in Sinopia.
        - Dave is still working on resolving issues with the new indexing scheme.
        - Performance numbers reported in the UI are getting worse over time. I believe this is related to timeouts of Dave's index. I would like to revisit the statistic collection code and have it track timeouts separate so that the response stats do not include the timeouts. This will add a column to count the number of timeouts as a separate analytic.
        - Causes of timeouts? 1) Index system down, 2) Maybe failure to find a result but not sure if Lucene still does this (was the case with SPARQL)
- Search API Best Practices for Authoritative Data working group
    - 2021-03-05:
        - There are 23 responses to the survey prioritizing the second charter's topics. All 4 potential topics are almost equally judged important. If you limit to only the first choice, linked data tooling and taking user stories to specifics recommendations are tied. If you look at only the top 2 choices, they continue to have a slight advantage over the other topics. If you take into account the top 3 choices, change management and language processing move to the top. And with all 4 choices taken into consideration, they are all roughly the same. I will send out the request for feedback one more time with the survey closing end of day Monday.
        - Some suggested topics are provenience, AI approach to selecting a term, enhanced UI for selection, and cataloging efficiency.
        - 5 respondents provided contact information.
- Cache Containerization Plan - Develop a sustainable solution that others can deploy
    - 2021-02-19 Greg completed CloudFormation template that allows someone to spin up a QA service in AWS easily. About 500 lines of template code that brings this very close to being a turnkey solution (in services-ci branch).Greg notes pre-reqs for spinning this up: S3 bucket for configs etc. which could be added to another template.
        - When complete Lynette will test, then ask Dave to test, then ask Stanford folks. Greg will also create a demo screencast.
        - What about replacing the current QA setup with this new approach? Would need to check authority configuration and correct setup for load. Lynette notes need to copy over the DB to retain history
        - Next steps
            - start to look at containerize Dave's setup. Two steps: 1) code to serve from cache, 2) indexing process
            - think about instructions for a vanilla linux server setup
    - 2021-02-26
        - Cache containerization discussion in QA-Sinopia meeting: We mostly talked about the next steps for the cache creating two containers: 1) container for API requests to retrieve cached data, 2) container to ingest data downloads and creation of the Lucene index. This is fairly straight forward in the current approach of a full-data dump and ingest. It is expected that there will be some complexities to resolve in how to update indices when change management techniques are deployed by authority providers that allow for incremental updates. We punted that discussion until later when the format of change management streams is defined. Stanford was asked their preferred deploy platform and they indicated that AWS was preferred.
        - Greg will work with Dave when he starts work on containers and tester and sounding board
        - CloudFormation - Greg has written templates and Lynette is going to test these out (will document time taken). Hope to find anything missing in template or documentation, perhaps some permissions issues will be revealed too that will allow documentation of critical permissions
        - Next Greg will look at prerequisites that need to be set up and work to template these in a helper template
    - 2021-03-05
        - Completed prerequisites template which includes S3 bucket and EFS filesystem - next step is to document instructions and how then to move to next template
        - Greg/Lynette to coordinate Lynette's testing next week - use feedback to refine documentation
        - Then create demo screencast

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
- 2021-02-05 discussion
    - 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
    - Functions of "Aggregated index, allowing pivoting & ETL"
        - Includes current functionality of Frances' indexing
        - Does it include any editing?
        - Is there interaction with CULAR?

        - Includes indexing associated with DCP
    - What interfaces or functionality do we expect for the connecting lines?
    - Do we need a diagram for now (or at least July 1, 2021 with Voyager gone)?
- 2021-03-05 Jason plans to update diagram and create narrative around it, hope to discuss next week

Other Topics

- OCLC Linked Data / Entities Advisory Group
    - 2021-03-05 See comments above
- PCC - Sinopia collaboration
    - 2021-02-05 Charge to form a new group for documentation, mentoring etc is under reviews
- PCC Task Group on Non-RDA Entities
    - 2021-01-15 PCC reviewed proposal but no decisions made yet, looking at description wrt cataloger use, discussion will continue
- 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.
    - ACTION - Huda Khan to look at changing to `main` for LD4P/discovery
- SVDE Workshop - several attended
    - Impressed by clear presentation of models and active APIs (REST and GraphQL)
    - Expecting models to be fully implemented this summer
    - At some time might want to add module to QA to query against GraphQL
- Authorities in FOLIO
    - Hope to include URIs as part of Cornell FOLIO migration, possible LD4P work

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
    - 2021-02-05 Huda submitted abstract
- code4lib - Expecting to attend: Huda, Steven, Lynette
- Lynette doing a QA presentation at Samvera partner call in June

Next Meeting(s), anyone out?:

- 2021-03-12 ...