# 2020-01-31 Cornell LD4P2 Meeting notes

Date: 31 Jan 2020

Attendees: Tim, Lynette, John, Huda, Simeon

Regrets: Jason, Steven

# Agenda & Notes

## Review actions from 2020-01-24 Cornell LD4P2 Meeting notes

- Huda Khan to discuss with Astrid and David possible collaboration with U Chicago over usability (and maybe others in DOG team)
  - Notes: https://docs.google.com/document/d/1_V7JfAKqSn63-G0zupLZigt8o7R3v_WXC7hWlt8w744/edit
  - 2020-01-10 Chicago report will be discussed in DAG meeting on Jan 21 (expect to happen as scheduled). Follow up with David and Astrid to look at what we could apply from what they learned and what additional user studies they could help us with
  - DAG presentation did occur and it was very interesting (link to meeting) – in depth interviews to understand research needs with open-ended questions, need more thought to understand how we might apply lessons from this
- E. Lynette Rayle QA performance
  - Dave made some changes to address issues with 50x responses, still possible issues under high load (possibly something on the Sinopia side but waiting to look at IP address to logging to identify whether same source is hitting us with same request)
  - Ongoing discussion about the need build more complex indexes to deal with slowness of complex queries
  - 2020-01-31 - Dave, Steven, Lynette had a meeting this week. Created a list (HERE) . Dave is still optimistic about perf improvements in SPARQL with change from CONSTRUCT to SELECT, but not sure when Dave will be able to try this. Also looking at indexing approaches with smaller sources than SVDE, will try LOC which is expected to take 3 days. Will also work to cache context when needed and not to request it when it isn't needed. Steven noted this on #authorities channel. Three categories of approaches: 1) amount of extended context, 2) efficiency of queries, 3) scalability of requests. Longer term there are questions of lookup vs. autocomplete modes
- See items under travel/conferences

## Status updates and planning

- John has attended to D&A meetings with user reps, showed video searching Google Books to deal with the zero results case which they found interesting. Adam notes that much D&A work is quite low-level and we don't have a good way to think about the big ideas.
  - Could we have a server for labs/beta? Might want to have a carefully chosen selection of the most promising features we develop
  - Maybe bring up idea of beta server at March meeting? Or do it ahead of time so people can play. ACTION - Simeon to ask Adam to work out cost of a similar server, and devops work we would need
- John had discussion with HT about API which is on their to-do but not scheduled. May be open to providing index access
  - There is investigation but not sure whether it will result in something we can use
- Cataloging Sinatra and other 45's (Discogs data, https://github.com/ld4p/qa_server/issues?q=is%3Aissue+is%3Aopen+label%3ADiscogs)
  - Lookups for place not usable, relies on work from Dave to fix: https://github.com/LD4P/qa_server/issues/248 & https://github.com/LD4P/qa_server/issues/240
  - Work is being done but places are not being recorded
  - Steven: Revised the Cornell Sinopia documentation (https://docs.google.com/document/d/1vatjjuDOAy5Qzi-Jj-JmH9zytjyUd6c3X4DzMWvybL4/edit#heading=h.3tc5xzitbah8) to reflect the Discogs lookup and UI changes
  - Steven: Investigated why not all entities created by the catalogers aren't showing up in the Search Tab. Turns out… because embedded templates don't produce URIs for those entities, they are not indexed as separate entities for reuse. Only the "primary" enitity gets a URI and is a result in the search. Eventually we can decide whether to de-embed templates and as catalogers to copy and paste URIs between Sinopia templates so everything is searchable, or wait for Sinopia to improve on this.
- Enhanced Discovery (see also https://wiki.duraspace.org/x/sJI7Bg and https://github.com/LD4P/discovery/projects/1)
  - BAM!: (Still) Finishing up lessons learned
  - SMASH! (to run through end of January)
    - John looking at the cases of queries that return small numbers of bad search results. Should "a secret history" return "the secret history", especially if there a few matches from the default search? John will continue to explore...
    - John: Followed up with HathiTrust and they may have Solr results and/or access experiment in a few weeks
    - ANNIF update: Annif requires a vocabulary and then training data to enable suggestion retrieval based on input text or an input document. Worked through their tutorial/documentation (https://github.com/NatLibFi/Annif-tutorial/ and https://github.com/NatLibFi/Annif) to setup LCSH vocab and training documents based on our Solr index. Vocab: Retrieved all LCSH pref labels to URI matches from Dave's LCSH SPARQL endpoint (excluding any blank nodes). Training documents: First queried solr index for 10000 documents looking for full title display and subject display fields, set up script to go through documents and query text of subject field against id.loc.gov to retrieve URIs. Resulting training set had 8432 rows (each row is tab delimited title then followed by whatever subject URIs correspond). Loading in vocab and training documents, annif can be asked through command line or through REST API for subject suggestions based on input text/query. Tried that out with a few keywords and could see some results. Plan next on (a) integrating REST API with data as it stands into the subject/person suggestions UI, (b) increasing size of training document set and (c) looking into what it would take to set up the ensemble option which allows for integrating multiple text analysis/classification strategies.
      - Additionally, Annif's own site includes Wikidata (English) suggestions. John may look at these.
    - Tim demonstrates auto-suggest based on a small specialized index. Separates authors, locations, subjects, etc.. Would probably require a specialized index to support at scale
    - John working on getting synonyms from wikidata with the hope of addressing the zero-results scenario. Feel that current wikidata data is rather limited in its understanding of synonyms
    - Huda demonstrates related subject and people. Does search in LCSH, facet values based on the query and also calling annif (set up locally with 10k records from our catalog with title and subjects, and entire LCSH vocab. Will work to add more data into annif, will try to get FAST info from Dave, and also avoid empty boxes showing with no query
    - Plan to wrap up work next week, then move on to user tests, write up and video
  - Open meeting March 3, 2-3:30pm in Mann 102 and should Zoom it too

    - How will we decide what to take forward from KAPOW!, BAM! and SMASH!? (or as Tim put it, "what happens in late February?")
- Authority Lookups for Sinopia (Lookup infrastructure: https://github.com/LD4P/qa_server/projects/2, Authority requests: https://github.com/LD4P/qa_server/projects/1)
    - See above
    - Monitoring tests have been adjusted to run at night, required work with Robbie to adjust pingdom
- Travel and meetings (see LD4P2 Cornell Meeting Attendances)
    - 5th International LODLAM SUMMIT at the The Getty Center in Los Angeles. February 3-4, 2020
        - Steven, Lynette and Simeon will attend
    - code4lib - March 8-11, 2020. proposals done so no opportunity to share LD4P work, although registration still open
    - Knowledge Graph Conference (Columbia University), May 6-7, workshops 5/4-5/5
        - Workshops due 1/15 (not doing this), proposals due 2/28, talks 20mins
        - Huda Khan will lead effort to consider proposal
            - Preliminary brainstorming
    - LD4 Conference at College Station, TX (TAMU) - May 13/14, 2020
        - Proposals due 2020-01-31 – https://stanforduniversity.qualtrics.com/jfe/form/SV_ebmQF48Gn9bag6x
        - Lynette will consider lookup best practices and discuss with OCLC
        - Steven/Tim to think about discogs "supervised conversion"
        - Per slack message, Steven considering workshop on RDF
        - Huda/John/Tim/Steven to think about KPAOW! BAM! SMASH!
            - Preliminary brainstorming
                - Briefly discussed how there is probably enough content between user research, larger questions, and experiments/development and prototyping to fit into two presentations. Tim suggested using the larger questions to frame discussion of experiments/prototyping.
    - rdfs:seeAlso Conferences Related to Linked Data in Libraries
- Next meetings: