

2020-01-24 Cornell LD4P2 Meeting notes

Date: 24 Jan 2020

Attendees: Tim, John, Huda

Regrets: Simeon, Jason, Steven, Lynette

Agenda & Notes

Review actions from 2020-01-17 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](#))
- **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
 - 2020-01-17 - no recent changes, Dave is out
- **Simeon Warner** to check on travel budget
 - DONE - Funding is sufficient for 3 at LODLAM, 1 at Knowledge Graph and potentially all at LD4
- 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?
- John had discussion with HT about API which is on their to-do but not scheduled. May be open to providing index access
- Cataloging Sinatra and other 45's (Discogs data, https://github.com/ld4p/qa_server/issues?q=is%3Aissue+is%3Aopen+label%3ADiscogs)
 - Usability - suffering from long vertical display with expanding sub-forms, lots of scrolling
 - Discogs QA module input is only using some data based on the profile being used, trouble filling in only work and not instance data when working from a work form with embedded instance. Steven will investigate whether template or profile changes can address this – 2020-01-17 Steven did some work to see whether filling in of multiple forms work differently depending on where you start. Found that starting with master doesn't add instance, starting with instance does add master too. Tim reports that this is intentional in the design of the Discogs authority, one could use the Discogs notion of "main release" to add one release when a master is loaded but this isn't done and might not be right ==> **conclusion is that cataloging should start with a release/instance**
 - Lookup performance still an issue but not even or even consistent within the same lookup. Seems also that there is an effect from overall load on the service. Steven/Lynette/Huda to work out where the problem lies – 2020-01-17 being scheduled next week (as I write)
 - Scheduled for 1/28 now
 - 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
- Enhanced Discovery (see also <https://wiki.duraspace.org/x/sJl7Bg> 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.
 - 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 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
 - **E. Lynette Rayle** will put QA in for the "tool challenge": CFP due 1/17
 - code4lib - March 8-11, 2020. proposals done, 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, discuss next week 2020-01-17
 - 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: