Linked Authority Data - community notes - Hydra Connect 2016

Led by Anna Headley (CHF)

Proposed session because:

- IMLS grant application from Chemical Heritage Foundation was successful to do work on this area
- Trying to do work to leverage existing authorities or related work
    - Make cataloging easier when it comes to Authority data / applying terms (lookups, selection, managing once selected or applied)
    - Example of 1 vision:
        - want for a cataloger or curator to come to create/ingest their digital object, have to give it an author, have various authors in use already in the system, want the cataloger to be given information about those existing authors, select terms from any of a number of 'approved' vocabularies (VIAF, FAST), not have the cataloger required to go search through other websites, not know what a URI is but have the URI stored in that system.
        - What comes along with capturing URIs in system are labels for search and display, caching data from URIs so resistant to external failures
        - Wanting to create something that can easily fit into Hydra systems, work in existing codebases as much as possible, bring in other existing work

Other use cases? Examples? Tools?

- Vision sounds in line with others' needs
- Metadata people like using OpenRefine, thinking about ways to go between OpenRefine and the Hydra system, using the OR UI to perform metadata work + reconciliation + integrate with Hydra - thinking about it but not doing anything one
    - So this feels like another use case: cleaning up of existing metadata
    - Would like to use for more work
    - Could we put this forward into systems we're cataloging into? Can we use same sorts of URI calls in a reconciliation service in OpenRefine might be able to put forward in a cataloging system
    - Question: Has anyone round-tripped this stuff?
        - Yes. Right now, UC Santa Barbara repository ingest is built on a batch system, and requires URIs for any terms coming from certain controlled vocabularies.
        - (Alexandria project) Early days, but part of process is to use OpenRefine, or depending on where metadata is coming from, to sub out all of the strings with URIs before ingest. For local authorities that come into the system that happens - when they come in, the ingest process will look first to see if there is a matching string value in the local authorities table or datastore, and if there is, it will sub out URI for that string; if not, it'll go ahead and create one (URI).
        - For now, local authorities, is string / URI setup. No automatic reconciliation for things outside of local authorities, but would like to use it beyond this.
        - Problem with reconciliation in OpenRefine or MARCEdit: authors with lots of uniform titles is that requires still manually match.
- Trey: there is really good software for caching URIs, thanks to the Hydra Applied Linked Data group (Linked Data Fragments), but for reconciliation, the recommendation right now in Hydra land is put up Questioning Authority, shim for searching interfaces
    - Corey: those kind of query interfaces fit with what the LDF (linked data fragments) community want anyway

        - Want fuzzy searching
        - probabilistic matching
        - How to order these results?
    - Tom: Missing piece, all can sit down and design LDF subspec that deals with this problem, but big question is what are the local search patterns? How do you handle algorithms for matching across contexts?
    - Contextual data important for these matching algorithms:
        - E.g.: Personal names
            - In an IR they won't have an authority file, but there will be people with the same name as them that do have authority records
            - Use title, type of work
            - Use different string matching algorithms for names vs. titles
        - Context for machine matching vs. context for a human selecting the right term
- Names of tools to use:
    - Apache Marmotta, Stanbol (Amherst College)
    - Trey: big thing they discovered is that type-ahead is largely useless for catalogers' interface
    - Steven: developed LDF endpoint as standard API, use Metadata Enrichment Initiative (MEI) to see broader/narrower terms when you search, get more details back
- Is there a consortial use case? Where different members use different authorities?
- Emerging Use Case: Generating local authorities, asserting local preferences or new assertions on external authorities
- In this space, we're often ignoring a lot of work in computer science, linguistics, other domains for this work on handling strings. Lot of good practice there that we miss
    - Tom: why he's working on his PhD, going to Cornell Local Authorities IMLS Forum + doing backstory work/research
- In the Archival world, everything is a local authority file, so how can we use local context would be important for this. Linked Open Archival data tools would be helpful to share in that domain.
- Easy wins?
    - Set local preferences for where you're querying, be able to customize preferences in institutions
    - Be able to take into account records in local authority system
    - Well-defined interface for what searching authorities look like
        - One of those being setting which authorities you search through
        - Think about algorithms used for returning results
        - Then plugging into a viewer like Steven Anderson's work for plugging this into field, type-ahead
- Tom: Question to get to those end goals
    - What constitutes an "Authority" in this LD realm?
        - Anna: Is this something that Questioning Authority is doing well? You set up named authority sources, then access by name?
        - Tom: Is there something that QA relies on that's unspoken that we need to communicate to publishers.
        - Trey: When I get a record for John Smith, it should say that it is an authority record for John Smith?
        - I asked an API and it said so
        - Sometimes you get different answers from QA vs OpenRefine

Each authority has its own way of modeling

- E.g. FOAF:agent vs. MADS-RWO

- I just wanna know where to get the prefLabel! But it's in a different place in each case
- I want my data to look the same regardless of where it came from, e.g. don't want "(artist)" after the name.

LC authorities are too lossy.

Interested in wanting to serialize data in our repository to expose as LOD

- Arwen: starting to look at that, figure out how to pull in external links, on her wishlist is basic be able to have a link to a wikipedia page for a thing when there is one, and targeting where to put that locally - cache? Other? - to easily push into UI
- Trey: generally the next step. Usually you have a repository full of URIs, hard to avoid the fun of what can I do with this data now? Lets build a map out of Geonames URIs, lets get alt labels for things, don't want to configure the labels for keys for values, so get preflabel for predicate and put that in
- Comments on leveraging Wikipedia/Wikidata links
- Ryan (Oregon Digital): CSVtoBag - used for migration, put local authorities there, handles before ingest lookups. Curious about OpenRefine if it would help with cleanup or lookup data. Pointing to external sources, if archival findings with a link, something they've wanted to do for a while.
- Steve: Skepticism about all of this. Have hard time pointing to good, cool, relevant examples. These are types of things we could be doing if it is resourcing this type of work. Can we make a space for this?

Summarized threads of discussion:

1. Realtime cataloging lookup / metadata enrichment
    a. Be mindful of services created / used and how that could scale to 2nd case
2. Pre/post batch cleanup using tools like OpenRefine
3. How to represent this in Hydra once we have it so we know
    a. Provenance of reconciliation, match
    b. What reconciliation process was used
4. What do you do with this for end-users
    a. Ex Libris Primo emerging abilities of some sort, Steve Myers BibCard work

Anna: focused on real-time enrichment and lookup piece for project, but interested in all

- Arwen: when writing look up work, think about batch needs as well. Having those go through same services seems like it could be a big win

Where to put this discussion? Threads?

- Metadata Working Group for examples, practices, documentations, etc.
- Applied Linked Data Working Group too
    - Steven will clear the membership list on the wiki, send email to sign up again, send out doodle poll
    - Redefine as interest group