Applied Linked Data Call 2015-10-15

Attendees:

- Steven Anderson (Boston Public Library)
- Trey Pendragon (Oregon State University)
- Arwen Hutt (UC San Diego)
- Corey Harper (New York University)

Linked Data Fragments Update:

- No ActiveTriples progress so still have to hold off on integrating with that.
- Steven will have work towards a working "Repository" (from rdf.rb) interface as an alternative to Marmotta by a Friday standup.
- Corey has documentation on his TODO list and will work on that as time allows.

Side car indexer discussion:

- Atomic Updates:
  - When stored fields were enabled for one institution, some of their OCR was 700 MB.
  - So then you get back 700 Megabytes of full text in Solr if storing it there and no way in Solr to exclude returning a certain field.
  - Don't want to have to pick out just the fields one wants... makes it more complicated to write code.
  - Possible solution: Request the full document from Solr, then append the update, then resubmit.
    - But can't do that since the data isn't stored in Solr to do that. And this is somewhat what atomic update is since Solr does this internally if it has all stored fields.
    - Even if you can store it elsewhere, don't want to push 700 MB over HTTP back to Solr.
  - Another possible solution: One that may work is turn on field highlighting for your full text... may not return the full field in the response, only the matching part.
    - Would need to test this to see if this is true... https://wiki.apache.org/solr/HighlightingParameters
    - Steven will be the one to look into this.
  - Other possible solution: solr child objects?
    - Won't work as can't really query for the main object with that?
  - More solution option: Can you provide wild-cards to the field list selector to handle this issue?
    - https://wiki.apache.org/solr/CommonQueryParameters#fl
    - "*_ssim"
    - "full_text"
    - "fl:"*_ssim, _tesim"
    - Seems related to this issue: https://issues.apache.org/jira/browse/SOLR-6545
    - Corey will investigate this option further.
  - Other possibility: Elastic Search over Solr?
    - Elastic search can't do paging of facet results currently though?
  - Reason for Atomic Updates: don't want to have to query Fedora (or any persistence layer).
    - One possibility to reduce the amount of persistence layer calls is to have the enrichments in Fedora save method.
    - But slows down the save method.
    - And still need to update the entire document on an external source change.
  - Some work towards a side-car indexer sort of done as part of Trey's Hydra Connect 2015 talk.
    - https://github.com/terrellt/linked_curation_concerns/commit/086afe2d883990a8e1faab202da5477bd7d3b6b
    - You call an enricher, it pulls down the solr document that has the id, and then enriches it.
    - All out of Oregon Digital 2 (alt label enhancer out from Oregon Digital).
    - Could extract this out into a gem (or a Rails Engine) that the side car indexer can use.

Linked Data Fragments Standup:

- Will be on Friday, October 23rd at 11:00 AM PST / 2:00 PM EST on the same Google Hangouts link.

Next Official Meeting:

- Next meeting will be October 29th at 9:00 AM PST / Noon EST.