2020-06-08 Meeting: Introductions and Brainstorming

ACTION ITEMS:

Agenda
- Introductions
- Review of Working Group Charter and Logistics
- Brainstorming
- Review documents completed prior to this meeting: Current Approaches to Providing Search APIs and Needs of Consumers and Applications

Meeting Materials
- Working Group Charter
- Current Approaches to Providing Search APIs
- Needs of Consumers and Applications

Notes

Introductions
Name, Organization, 1-2 sentences on what you hope to get from the group. Introductions need to remain short to avoid having them take up the majority of the meeting. Please share the critical aspects that draw you to participate in the group.

- Lynette Rayle (Cornell) - Two pain points: 1) Differing methods for accessing authorities requiring one-off coding, 2) Getting back enough information to be able to display results to end users in a meaningful way.
- Justin Littman (Stanford) - Desire to code as little as possible and be able to access authorities.
- Christine Fernsebner Eslao (Harvard) - Continue with work begun in LD4P2. Would like to see various lookups standardized with a goal of better usability. Would like to see reconciliation automated.
- Nancy Fallgren (NIH) - Hoping to create more authorities. Want to make more available through RDF. Want them to be usable.
- Tiziana Possemato (Casalini) - Make connections with other groups and be on positive terms with others. My invite technician to join when needed.
- Kirk Hess (OCLC) - Worked in research previously with LOC. Worked on high availability and throughput. Looking to bring experience with LOC and new connections with OCLC.
- Jeremy Nelson (Stanford) - Works on Sinopia Editor. Hopes to minimize code and interested in seeing how this supports a diverse set of authorities.
- Steven Folsom (Cornell) - Coordinator of metadata design; As a consumer, want any tool creating metadata to be able to tie into everything you want where data looks similar and easier to co-index.
- Kevin Ford (LOC) - Has served in roles as a producer, publisher, and consumer.
- John Graybeal (Biportal): consumer (Cedar), producer (Biportal); Cedar is a consumer of Biportal
- John Chapman (OCLC): entity management infrastructure project; producer/consumer; possible implications for VIAF or other projects.
- Rob Sanderson (currently independent, Yale in September): vocabulary management (currently working Arches); integrations that are not system-specific
Aspects of interest generally fall into these categories:

- Publishers want to make consumption easier
- Consumers want meaningful display and search of data
- Software developers and engineers want lighter/more maintainable code.

Working Group Charter and Logistics

Reference: Working Group Charter

- What does it mean for linked data to be an "approach" rather than a part of a system?
  - Having a REST API return LOD
  - Sanderson: Are we concerned with update/delete/write/etc?
  - Rayle: Those are part of the larger picture, but for the purposes of this group's deliverables, we are focused on search and return of data. Everything is fair game in brainstorming, and a subsequent working group is possible.
  - Sanderson: Previous attempts at standardizing search APIs have failed. We need to be very specific about details.
  - Rayle: Add links to outside documentation under "References".

Logistics:

- access to Slack channel
- access to Wiki
- note takers

Brainstorming

Terminology used in the brainstorming and laying the foundation for a common understanding...

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconciliation</td>
<td>Things-to-Things - The process of identifying that two things represented by different URIs are actually the same thing.</td>
</tr>
<tr>
<td>Entity Resolution</td>
<td>Strings-to-Things - The process of identifying that a String label is the label for a thing identified by a URI.</td>
</tr>
<tr>
<td>Disambiguation</td>
<td>When presented with multiple very closely labeled options, this is the process of determining which option is the correct entity.</td>
</tr>
</tbody>
</table>
| Caching               | 1. Storing local copies of entire dataset.  
                             2. Caching of a single label or small pieces of data about a single term.                                                               |
| Accuracy              | 1. The ability of an API to return relevant data  
                             2. The ability of a user to select a term from multiple similar terms given a set of search results                                   |
| URI Derefencing (aka Term Fetch) | Accessing a URI through CURL or a web browser shows data related to the entity the URI represents. The data can vary for different formats (e.g. JSON-LD, n-triples, HTML, etc.), for example the data displayed on a webpage using HTML may be different than the data retrieved through CURL when requesting an RDF format. And the amount of data returned can vary between authorities with some only returning data where the URI is the subject and others returning data in the wider graph. |
| Search (general)      | Given a string query, search a dataset for occurrences of the string. The search is generally an index over the dataset to provide acceptable performance times. |
| Keyword Search        | Search for a string query typically across multiple fields where the query can be anywhere in the field.                                    |
| Left-anchor Search    | Search where the query is an exact match of a label starting with with the first character of the label.                                   |
| Browse (general)      | Approaching a dataset from a specific perspective and navigating to a desired (known or yet known) entity. A “perspective” may or may not be appropriate or supported by a given data set. |
| Hierarchical Browse   | Starting at a term in a hierarchy and moving to broader or narrower terms.                                                                   |
| A-Z Browse | Navigating through all terms in alphabetical order. This may involve canned queries for groupings (e.g. A-D, E-H, etc.) |
| Incremental Search | aka real-time suggestions or a typeahead, as the user types text the query is run in real time and matches are found and immediately displayed. See also Left-anchor Search |

Instructions: All topics related to accessing authoritative data. This can include topics that are not directly related to search APIs. Topics of interest to the group, but not directly related to search APIs will be considered for the tail end of the working group if there is time or considered for a new working group if there is enough interest. So let all your thoughts flow.

- **Caching**
  - NOTE: There is some overlap in the ideas listed under the two areas of caching, for example, there are questions for both types of caching around versioning of authorities.
  - Caching 1 (Storing local copies of entire dataset)
    - Downloads - cache management synchronization
  - Caching 2 (Caching of a single label or small pieces of data about a single term)
    - Notification of updates when entity descriptions change, or at least ability to search by dates/types of changes
    - Deprecations - mechanism to state that this term is replaced by another term - how do end users know that the term is no longer valid
    - Versioning of authorities and impact on labels that have been deprecated, changed, or deleted
    - Learning how something has changed and knowing what to do in response

- **Reconciliation**
  - entity reconciliation
  - reconciliation - reconcile more than just adjunct works and references to outside data and across languages
  - concerns with sameas connections between data that may not be accurate
  - reconciliation - why isn't this just the opendrefine API; implemented and well understood

- **Accuracy**
  - Accuracy 1 (The ability of an API to return relevant data), Accuracy 2 (The ability of a user to select a term from multiple similar terms given a set of search results), and Entity Resolution (Strings-to-Things - The process of identifying that a String Label is the label for a thing identified by a URI)
    - NOTE: I put these all together because they are highly interrelated. What and how well the API returns relevant data greatly affects the user's ability to accurately select the correct term from the results.
    - NOTE: Authorities bear the primary weight of entity resolution by assigning a primary label to a URI. Once a term is selected, the label comes from the authority. It is in this section since the end user has a string in mind when they type a query, and the search/selection process turns the search query string into a selected URI with a label.
    - How do humans choose between two similar but distinct entities?
    - identify you have the authority you want - got the right john smith

- **Extended Context**
  - right information to search and display - need an easy way to define
  - moving to linked data - which attributes to include for each entity
  - May need more context some times
  - Can users personalize how the data comes back?
  - being able to manage and choose the context of your query where context is provenance, community recommendations, relationships, etc
  - Presence or absence of connections to other entities; making choices based on interconnectedness and fullness of data

- **Order of Presentation**
  - Rank ordering so results are displayed with first results as best result
  - Listed alphabetically
  - Option for left anchored search
  - Pagination
    - expanding set of results
    - turning off pagination to get all results
    - server side pagination vs client side

- **Other**
  - Which labels to display when multiple labels - across languages and scripts and kinds of name

- **API approach**
  - browsing with context when know what you are looking for and a good amount of time, catalogers know what they are looking for
  - searching to discover when you don't know what you are looking for
  - One service doesn't fit all needs. Suggest, Search, Browse - each serve different needs
  - How to surface versioning of authorities through API
  - what is the impact on APIs with respect to deprecations
  - to enable local/specialized authorities to participate by implementing the API - allow local to define a narrower term to a broader
  - API - focus first on retrieval by REST before search and browse
  - more consistent and granular error reporting to determine the source of the error
  - handling different data models (the organization of the returned data)
  - handling different syntaxes

- **Data Related**
  - Which data elements are unique or intended to be unique?
  - Which data elements are intended for end users (such as library users browsing a collection) and which are intended primarily for internal use
  - Suggest SKOS ontology (Avoid talking about modeling for years) and
  - Suggest JSON-LD for format
  - handling different data models (the organization of the returned data) handling different syntaxes

- **Scalability**
  - responsivity for autocomplete (< 10ms)
Other
- indexing: connections of relationships between entities
- suggest 2.0 document
- Learning from users
- Drive recommendations from Use cases
- discovery of (and enrollment in) new authorities
- standards-based

Review documents completed prior to this meeting

Reference: Current Approaches to Providing Search APIs

Reference: Needs of Consumers and Applications

Attendees
- Kevin Ford (LOC)
- Kirk Hess (OCLC)
- Rick Bennett (OCLC)
- John Chapman (OCLC)
- Nancy Fallgren (NIH)
- John Graybeal (Bioportal)
- Tiziana Possemato (Casalini)
- Lydia Pintscher (Wikidata)
- Hetty van Zutphen (ISNI)
- Steven Folsom (Cornell)
- Christine Fernsebner Eslao (Harvard)
- Jeremy Nelson (Stanford)
- Justin Littman (Stanford)
- Lynette Rayle (Cornell)
- Rob Sanderson (independent)

Absent
- Nate Trail (LOC)
- Jens Ohlig (Wikidata)