

Search Tools Breakout Session
BD2K AHM Friday, November 13th, 2015

Co-chairs: Henning Hermjakob, Lucila Ohno-Machado, Alison Yao

Statement of Purpose:

Explore the following questions:

1. What do we need to improve upon within a Google search?
2. (How) can we achieve a common metadata model across BD2K? - *Make sure to connect with the metadata breakout session* about this concept.
- 3.. (How) can we leverage crowdsourcing efforts to achieve our goals?
4. What are the channels to feed suggestions and improvements back into the community?

Recommended Outcomes and Activities:

1. What do we need to improve upon within a Google search?
 - a. Investigate usability of schema.org
 - b. Outline desired extensions or add-ons for rudimentary schema.org use
 - c. Develop domain specific ontologies for schema.org
 - d. Defer to working group for necessary elements list regarding schema.org
2. How can we achieve a common metadata model across BD2K?
 - a. Help metadata working group identify relevant use cases to focus the expectations of commonality across data types
 - b. Relevant metadata should be associated but separate from the data itself
 - c. This question is downstream of current efforts – proper indexing for searching and appropriate domain-specific functionality has not yet been achieved
3. (How) Can we leverage crowdsourcing efforts?
 - a. Identify and implement the keys to successful crowdsourcing
 - i. Recruitment of qualified, knowledgeable individuals
 - ii. Persistent storage mechanisms
 - iii. Basic quality control protocol
 - b. Crowdsourcing efforts should incorporate passive algorithms to capture community usage
 - c. Develop/promote a method for citing datasets
 - i. Downloading metrics can provide a rough usage statistic with very low investment
 - ii. Move towards rating of datasets and metadata while being cognizant of political challenges therein
 - d. Work with publishers (of academic articles) to educate/mitigate their concerns around data citation as a metric which will undermine their relevance
 - e. Make it clear to the ‘provider’ that it is in their best interest to provide excellent metadata
4. What are the channels to feed suggestions and improvements back into the community?
 - a. Harmonization between metadata and CDEs (templates/data and forms) will avoid dangerous disparities between metadata and CDEs
 - b. Address these issues within the CRISIS working group
 - c. The channels of communication must flow through the community and then back to the developers: this will help avoid annotation/update/version issues
 - d. The goal of feedback channels is to build a social community similar to that which exists within programming development which works collaboratively to better the available tools

Additional comprehensive notes:

1. **What do we need to improve upon a Google search?**

Would it make sense from an indexing perspective for BD2k community to have a sub-branch of schema.org? Parallel community and schema.org publishing? In this way our sanctioned tags could be integrated automatically and we (as a community) could leverage that metadata.

- If everyone is happy with google search what are we even doing?
- We're trying to search data which can be indexed by the search - can be extended to software resources tho those options are not available on schema.org currently
- Could BD2K co-opt parts of schema.org for our uses? Index github and associated pages for incorporation into the search engines.
- We can have our own 'home' route (ie **bd2k.schema.org**)
- Data, datasets, repositories, software, **absolutely everything ever** (to be searched)?
- How far 'down' are we going to be extending this indexing? PHI concerns related to "granularity of a certain degree"?
- Metadata specification concerns related to schema.org - it is under consideration and is an option for annotation
- European bioschemas currently considering specific extensions of schema.org
- The goal is not to take Schema.org as it is today but rather to take it as a basis, adapt it to our uses, and reate a new hybrid which will be useful specifically to our community.
- **Why CAN'T we just take schema.org? what is it missing? →in need of extensions for use in bioinformatics.** It's been one of the models considered for mapping and has not been chosen
- CEDAR group comment -- mapping data terms to existing ontologies -- domain specific terms can currently not be found in schema.org. Higher level terms may be co-opted for use in annotation but are not currently adequate - **domain specific terms needed**
- If we can use it, what do we gain?
- reservations: Schema.org may not be willing/able to extend to domain specific terms -- should not overlap with the work of current communities dealing with specific domains and communitys in their ontological development
- NURSA - 85-90% experiments in signalling result in relative values -- omics scale expression experiments w/ small molecule ligands: creating search terms for specific experiment types such as this **in Google are a significant difficulty -- signal-to-noise ratio for these searches makes it largely incomprehensible and nonuseful**
- Useful search tool for users would maximize signal-to-noise ratio for search terms and common experimental types
- Metastudy: efforts to make discoverable by data discovery index (DDI) - how hardened are those data elements? At this point: still in flux. Core model doesn't expect significant changes but additions are possible at this point.
- Metadata for metrics: how much has a particular dataset been downloaded?
 - o Does this include using schema.org?
 - o Is there a concept of "change management" due to cost, time, training requirements related to potential changes to a 'finalized' model?
- defer to the working group and their decisions related to this issue.

2. **(How) can we achieve a common metadata model across BD2K? - Make sure to connect with the metadata breakout session about this concept.**

- Is this the expectation? How broad can the scope be given our diversity of data types? Is universality desirable or attainable?
- There's another metadata working group -- network w/ them

- **Thinking of a metamodel the use cases must be considered.**
use cases >---< commonality across data types
- This question is way downstream of current efforts - finding them is our primary concern currently
- At this point there are simple issues with not finding any data of the type you are looking for within obtained datasets -- this is a downstream problem which will be addressed when indexing has reached sufficient complexity and depth.
- Mobile sensor data to knowledge: how to leverage. Some groups are very interested in sharing data. However individual citizen scientists collecting data which may be useful create new backwaters for data aggregation.
- Patient-powered research networks can act to aggregate and publish/coordinate data, however these groups are simply part of the data aggregation problem
- By exposing some quantity of metadata it may be possible to direct data discovery and indexing -- ie by including metadata before access to databases is obtained it may be possible to identify which databases you would like to access
- Is there a concept that a dataset can be used for multiple purposes? ie off-target searching possibilities. **No** - this will result in it not being 'as good' as a general-purpose search service, however it will result in specialized search capabilities.
- It's up to providers to be responsible about their metadata - providers influence the findability of their data. **Multiple metadata descriptor sets** can be attached to a single datasets which could for instance be related to the ways the data is processed for different use cases.
- Metadata and data -- associated but separate
- (How) can we achieve a common high level query API? - *Make sure to synchronize with the API working group that also has a breakout group planned.*
 - can should/could the API handle metadata?> Yes
- How we leverage other search methods and synergize with other search systems that are currently being used by the broader community

3. **(How) can we leverage crowdsourcing efforts?**

- In practical terms how can we benefit from crowd-source efforts (ie individuals looking at data and annotating, adding in their thoughts or curation to data) and how do we engage individuals and incentivize feedback?
- Whatever system is built, if it is created as a learning system (ie **passive algorithms** for ranking) then some community usage can be captured and feedbacked into the system.
- **In CS: crowdsourcing used to annotate datasets -- this may not be so true in health data.**
 - **recruiting qualified individuals**
 - **creating persistent storage mechanisms**
 - **quality control - does it make sense??**
- crowdsourcing requires the ability to go back to the crowd and get their feedback on changing situations
- useage, commentary on data: this can help avoid wandering down blind alleys
- potential political challenges in commenting on private datasets
- re-useable pointers -- how can you track your own data's use and receive attribution
- bioCADDIE UT how do we get user input and what do we think about rating systems?
- It'd be difficult to know without trying it, but comments are already associable (and still exists/thus isn't a 'disaster') so maybe it'd be okay
- Why should datasets be held to a higher standard than research articles? We already know that there are substandard research articles and yet nobody is talking about how

those should be rated. Citation has its flaws but is the best metric available for research articles, so perhaps we should start there and move towards dataset ratings and comments.

- Publishers have a good deal of suspicion (lack of education) about the need for citation of datasets and the possibility that it will undermine citation of the associated research article in their journal..
- Limitations: use of traffic as crowdsourcing -- may be very variable and unreliable for feedback. We aren't going to be getting searches on the order of google.
- **Ratings: rating datasets and rating metadata around datasets**
- you can lead a horse to data but you can't make it thirsty
- software developer views: we keep referencing Google's model but think about some things:
 - google incentivizes people to produce metadata -- unless there's some Reason to do it, people aren't going to do it.
 - Data set citations will encourage interconnectivity and give information about ranking of that dataset within the community. (can be used as the Reason?)
- **Make it clear that it is in the interest of the 'provider' to provide excellent metadata. ("idealistic view"?)**
- Current postdocs are growing up with this view and it may become more ingrained culturally that this metadata is a requirement for successful publication -- clarify benefits
- Database citation may create a new search metric or search function
- Protein crystallography diffraction community is developing its own search tools: how can BD2K Help?? Standardized tools? New search functions not related to keywords? Domain-specific concerns.
- Question about how to make domain-specific databases searchable and accessible to the current community??
- Search software will be available on github and can be modified for specific use. See bioCADDIE github project group.
- How to raise data in the search: download metrics may provide some information about used datasets.
- Will some information about peoples' interests provide useful information related to download info? -- this would require logins which would be a barrier to participation. Would be useful information but isn't easy to obtain and would likely cut people out of the 'feedback' loop.
- What's the general process for indexing resources which are searchable by bioCADDIE? If individual labs or groups have developed domain specific databases how can they get their data or resources included in the future bioCADDIE efforts?
- See the working group for data inclusion criteria! It's still being decided.
- Sustainability is of concern (ie is the database going to disappear at some point?)
- In principle the BioCADDIE indexing should include **everything** represented by the biomedical community however the current efforts must be prioritized. Once the system is well established, everything ideally would be included
- Getting people willing to share their metadata about their datasets should be easy! However mapping your data to BioCADDIE's common data elements may provide a challenge.

4. **What are the channels to feedback improvements?**

- ie a new release of a dataset
- If channels are only going one way then this creates **annotation/update/version** issues

- Common Data Elements represent an important requirement -- how can we conceptually build the same consensus on CDE that we're now talking about with the 'forms' created by CEDAR , etc, for metadata?
 - Metadata & CDE are different in useage and CDE disparities can be dangerous. Harmonize between CDE and templates/data forms/etc
 - See the Standards Coordination Group: this is what they do and what they should be doing
 - How do we do that? Incentives? Carotts? Sticks? Partial harmonization mechanisms.
 - Open source software projects: build social community around the project which create dynamic feedback interactions between the use community and the development community.
 - How do we develop a culture like this for the use of data in biomedical contexts?
 - How do we prevent protective silo'ing by producers (PI's)? We can't count on the level of altruism provided by the open source software community
 - Perhaps the comments and generation of comments can be used as a prestige mechanism whereby individuals gain respect for creating valuable commentary.
 - How do we send those metrics out so that researchers can take advantage of their own data responsibility?
 - **CRISIS Working group (Criteria for Repository Inclusion(standards, interoperability, sustainability))**
 - Oversubscription of experimental designs -- some of them overtake the entire field or database for citations and thus can saturate the citation schema - this is one of the potential flaws of citation-based metrics.
- How do we present the query results back to the user?
 - How do we solicit relevance feedback from the user? (e.g., usage, citation?)
 - Sustainability: How do we keep the query engine and its contents up to date?
 - how do we keep processes fresh and avoid 'stale' issue with data search mechanism
 - **What happens when bioCADDIE ends??? Who takes it over?**
 - There is a working group on this topic tho there is no concrete solution yet. In discussion.
 - How do we maintain bioinformatics resources and tools long terms?
 - **Hand it over to the NLM? Some central organizer once development is 'done'?**