2016-07-11 Trip Report -- SHARE Hackathon and Community Meeting

SHARE Hackathon and Community Meeting

July 11-14, 2016
Charlottesville, VA

Monday, July 16 – Hackathon Day 1

Jeff Spies

SHARE version 2. More specificity about the contents of the database

Need interfaces for SHARE. SHARE does not want to be an interface to the scholarly work

Data needs discovery and refinement

Rick Johnson

Exciting time to be involved with SHARE

Erin Braswell

OSF work space. Code at GitHub.

Provider -> Harvester -> raw_data -> Normalizer -> normalized_data -> changes -> change_set -> versions -> entities

The Harvester gets the data from the provider. Uses date restrictions to get "new" data. The normalizer creates the values that can go into the SHARE data models.

Title issues: Unicode, LateX, MS Word, foreign languages. Attempt to store the language provided by the provider. Joined fields for titles with multiple titles. Can be stored as a list in the extra class.

Normalizers can guess title or identifier or DOI. Usually conservative normalizers.

MC Idea: data inspectors: Write elastic searches to get percentages of populated/vacant fields, by provider, by date range. Would show the density of field values in the normalized data. Could be used to draw control charts of field values density. Mirror the values.

MC Idea: data inspectors: Identifiers are a problem, often come in "random".

MC Idea: data inspectors: feed the results back the the providers. The providers may be able to suggestions enhancers to the harvesters and normalizers.

Documents can be updated – provider's id. If the metadata comes in for a record that exists, COS versions the record and provides the most current unless the query asks for versions.

See https://staging-share.osf.io/api/

Tuesday – Hackathon Day 2

Studied the SHARE API. Learned a bit about Elastic Search. Investigated sharepa (but it was not ready to be used for version 2. By the end of the community meeting Erin had upgraded it to work with version 2. Wrote the share-data-inspector. See https://github.com/mconlon17/share-data-inspector

Wednesday – Community Meeting Day 1

Keynote Siva Vaidhyanathan, UVa – The Operating System of Our Lives: How Google, Facebook and Apple plan to manage everything

Relationships with technology and information and communication changing rapidly. Mapping a game onto reality, engaging millions of people immediately into a game – Pokemon Go. Facebook Live – mapping reality into the virtual world, immediately, effortlessly, in real-time. Facebook took the video down for an hour, did not anticipate the incident of violence. 1.6B users, leading source of news for many millions of people. Facebook matches content to people. Facebook denying its level of power in the world. Google has the same position – constantly underplaying its role in pointing people at information.

We are collectively dependent on Google.

"The web is dead" – flows of data are not open docs loosely joined. Most data is moving through proprietary devices and formats. Our concept of the Internet is flawed/primitive. We have never been comfortable with the concepts of radical openness. Internet described in terms of place based metaphors "cyberspace" "Internet superhighway." Mobile devices changed that.

Apple sells boxes. Microsoft sells software. Amazon a retailer, largest source of revenue is AWS. Facebook sells connectivity to people. Google sells connectivity to information. Compete for labor, political power, advertising revenue, attention. Each has a plan to "win the game" – to become the operating system of our lives. Put things on our bodies, drive our cars, fully imbedded in our bodies. Data flows must be proprietary and controlled. Can not be open/standard.

Internet of Things – forget it. Seems helpful. The important thing is the monitoring and managing of people. Us. Companies must have a lot of knowledge about us. Difficult to enter the market – these companies have 18 years of data on us.

Edward Snowden showed us the data the government is collecting, and the purposes they have for the data. State actors are not benign, and often result in violence. Chinese government in full association with its social media companies. All states are excited by Modi, Putin, Erdogan, and the work they are doing on surveillance. Surveillance will increase.

We have voices as citizens. The Googlization of Everything.

Breakout – SHARE Notify Atom Feed

https://osf.io/share 117 providers, 7 million records (but how many unique – some feeds a re completely duplicative – Dryad and DataCite, for example). Clinical Trials.gov Zenodo, PLoS, Arxiv.org, Figshare, and 50 instititional providers.

How might we use the data:

1. The VIVO Use Case – showcase the work of the people at an institution. All the work.
2. Track work over time – increase/decrease of various kinds of work.
3. Check work over time – what do we know, what does SHARE know?
4. Understand the social network of scholarship – who works with who across institutions across the world
5. Understand the trajectory of scholarship – what areas are emerging, what areas are receding?

Atom Query String

http://osf.io/share/atom/?q=(shareProperties.source:asu)AND(title:"fish")

http://osf.io/share/atom/?q="maryann martone"OR"maryann e martone"

http://osf.io/share/atom/?q="m conlon"OR"michael Conlon"OR"Mike Conlon"

http://Blogtrottr.com for sending a feed digest to a mail address on a regular schedule.

Breakout session – related projects

Gary Price, Infodocket. Find more users for SHARE – high school students. Include press references to research. Semantic Scholar.

Karen Hanson, Portico, Ithaka, RMap. DiSCO – distributed scholarly compound object. Linked Open Data. Very cool. Discos can related to each other. Each disco has an immutable identifier (URI) that points at the Disco. Assertions about the resources. No ontology restrictions. Discos have a status. Using known ontological elements for connections. OSF Person to OSF Project to Datacite URI, linked. Plug in a DOI and see a graph of what RMap knows of that resource. IEEE was a sponsor, used IEEE data on publications to help validate RMAP. Has RDF representation of each DiSCO. End of grant, all tools will be open source. http://rmap-project.info

Lisa Johnson, University of Minnesota. Data Curation Network. Rise of Data Sharing Culture. Role of librarians – discipline specific expertise, technology expertise. Data curation network: Minn, Cornell, PennState, Illinois, Michigan, WUSTL. Collecting and reporting data curation experiences, metrics for results. http://sites.google.com/DataCurationNetwork

Hackathon Report back

Institutional Dashboard

Data Inspector

Metadata documentation

Research Data Discovery in Share

Data is coming from DataCite. Is there a data type for datasets? Yes, but perhaps not in the API yet?

Quality of data? Depends on the provider. Level of curation varies.

Sharing and discovering artifacts of the research process? Some artifacts can not be shared – proposals before funded. Data management plans before funding.

Does DataCite totally duplicate Dryad for data set consumption? Metadata might be different. Similar questions apply to other overlapping services – Dataverse and DataCite.

Alexander Garcia Castro VIVO and SHARE

SHARE is chaotic and promiscuous. VIVO is chaste, great precision.

Research Hub concept to engage faculty. Claiming, feedback to metadata providers.

SHARE Scopus Mendeley GitHub

Match and claim

Search -> Claim -> Add -> Connect Research Objects -> Social Connections -> Done

VIVO needs an engagement strategy. Beautiful, clear models, open, reusable semantic data.

ORCiD has minimal faculty engagement. Sign up, disambiguate, No reason for faculty member to go back to ORCiD after initial set-up. Identifier only. OpenVIVO could be, should be, more engaging.

SHARE is big, but messy. Also needs an engagement strategy.

Mendeley, ResearchGate. Giving researchers something. OpenVIVO has a bit more, but still very little.

Thursday, Community Meeting Day 2

Jeff Spies, Scholarly Workflow

OSF as a platform for scholarly workflow. Slides available here: http://osf.io/9kcd3

MC: OSF Needs:

1. Identity – benefit for faculty: collaborators around the world
2. Extensible/local workflow – benefit for faculty: reduce regulatory/administrative burden
3. Github issues – benefit for faculty: improve ability to manage team/project

Prue Adler, Brandon Butler, Metadata Copyright Legal Guidance

Copyright protects the original expression of the authors. Modicum of creativity, independent creation. Copyright does not protect facts, ideas, discoveries, systems. Effort, time, expertise are irrelevant in the US, not in the UK and EU.

Merger doctrine – if the idea can be expressed in only a limited number of ways, the expression merges w/ fact and is unprotected.

Selection and arrangement of facts can be protected if creative and original.

No copyright in words, titles, and short phrases.

No copyright in blank forms (psychometrics – perhaps this is a patentable method)

MC: VIVO Project was able to work with Web of Science and SCOPUS to clarify which facts in their databases were public domain and which were not. Public Domain facts can be harvested from these systems and used in VIVO systems, effectively making the facts open and reusable by others.

Contracts can restrict reuse regardless of contract.

Copyright applies for 70 years after the author(s) death.

Brian Nosek – Research Integrity

Signals – open data, open materials, preregistered. Badges are stupid, but signals helpful.

3% of articles had recognition of open data, two years later 40% have open data. PSCI journal.

http://cos.io/top Top guidelines. 713 journals. 62 organizations in the process of review and adoption of the guidelines.

Two modes of research: exploratory, confirmation

Preregistration challenge: http://cos.io/prereg

Registered reports: Design -> Peer Review -> Collect and Analyze -> Report -> Publish Aligns all incentives

Tyler Walters – Closing Comments

VIVO is a fundamental part of the university's infrastructure

Some Observations

The SHARE information environment is maturing rapidly. They have a large staff of talented young developers augmented by a very large group of talented interns from UVA. They have a tremendous amount of data, a growing set of providers, sophisticated version control, and a change set architecture for all their data which allows them to reharvest data to improve the quality of the current version. The API is easy to use. ElasticSearch works very well. They recognize the need to disambiguate their data and are planning heuristics, use of identifiers, and use of curation associations at libraries to merge entities.

Their data model is shallow, and driven by the fields that are commonly available from the providers. VIVO is very interesting to SHARE as it provides missing semantics and a very deep data model. The two efforts – VIVO and SHARE – are strikingly complimentary. VIVO would benefit tremendously from "world" metadata collected and curated by SHARE. SHARE qould benefit tremendously from VIVO's institutional footprint and the potential to feedback metadata improvements to SHARE.

Additional Conversations

I spoke with numerous COS staffers regarding data specification, semantics, use of identifiers, and the "golden query": "Return all the metadata for institution x" Given a reasonable result for this query, institutions could contribute to metadata improvement and use SHARE metadata for VIVO.

Next Steps

There is great interest in creating a "virtual cycle" of data movement between VIVO and SHARE. The VIVO existing VIVO Harvester for SHARE will be upgraded by SHARE to their version 2 data model. VIVO will enable harvesting of OpenVIVO for SHARE. SHARE2VIVO will be upgraded to allow users of OpenVIVO to update their profiles from SHARE data – this will require additional specificity in the SHARE data.

Tyler Walters – Closing Comments

VIVO is a fundamental part of the university's infrastructure

Some Observations

The SHARE information environment is maturing rapidly. They have a large staff of talented young developers augmented by a very large group of talented interns from UVA. They have a sophisticated version control, change set architecture for all their data which allows them to reharvest data to improve the quality of the current version. The API is easy to use. ElasticSearch works very well. They recognize the need to disambiguate their data and are planning heuristics, use of identifiers, and use of curation associations at libraries to merge entities.

Their data model is shallow, and driven by the fields that are commonly available from the providers. VIVO is very interesting to SHARE as it provides missing semantics and a very deep data model. The two efforts – VIVO and SHARE – are strikingly complimentary. VIVO would benefit tremendously from "world" metadata collected and curated by SHARE. SHARE qould benefit tremendously from VIVO's institutional footprint and the potential to feedback metadata improvements to SHARE.

Additional Conversations

Numerous COS staff regarding data specification, semantics, use of identifiers, and the "golden query": "Return all the metadata for institution x" Given a reasonable result for this query, institutions could contribute to metadata improvement and use SHARE metadata for VIVO.

Next Steps

There is great interest in creating a "virtual cycle" of data movement between VIVO and SHARE. The VIVO existing VIVO Harvester for SHARE will be upgraded by SHARE to their version 2 data model. VIVO will enable harvesting of OpenVIVO for SHARE. SHARE2VIVO will be upgraded to allow users of OpenVIVO to update their profiles from SHARE data – this will require additional specificity in the SHARE data.