Briefing Document for "Changing the Conduct of Science in the Information Age" NSF Workshop on April 26, 2010

Katy Börner, Indiana University, katy@indiana.edu

(1) Openly accessible data

Replicability is a hallmark of science. Those communities of science which embrace shared data (and software) repositories thrive. SciSIP researchers frequently work with "for-pay" data from Thomson Reuters or Elsevier or proprietary download data, e.g., Mesur, and hence cannot share their data. Many also use proprietary tools and few share code (it takes about 5-10 times more effort/time/funding to generate and document code that can be shared/reused). Consequently, it is impossible for others to replicate, seriously review, or improve results.

There are a number of efforts that support

- federated search over one or multiple, de-identified or not de-identified data holdings and
- raw data download as dump in structured formats.

Among them are

- CiteSeer, http://citeseerx.ist.psu.edu
- NanoBank, http://www.nanobank.org
- Scholarly Database, http://sdb.slis.indiana.edu
- NSF awards, http://www.nsf.gov/awardsearch
- NIH awards, http://projectreporter.nih.gov
- Data .gov, http://www.data.gov

The Scholarly Database (SDB) (http://sdb.slis.indiana.edu) at Indiana University supports federated search over 23,000,000 MEDLINE papers, USPTO patents, NSF awards, and NIH awards. Matching records can be downloaded as csv file but also in network format, e.g., co-author, co-investigator, co-inventor, patent citation networks, and formats suitable for burst analysis in the NWB Tool (http://nwb.slis.indiana.edu) and Sci2 Tool (http://sci.slis.indiana.edu). In April 2010, SDB has more than 220 registered users from 26 countries on 5 continents. We are in the process of exposing the SDB datasets to the Semantic Web as Linked Open Data.

The VIVO National Researcher Network (http://vivoweb.org) will soon expose high quality people data from systems of record (Human Resource, Sponsored Research, Course databases from academic and government institutions) to Linked Open Data.

Linked Open Data (LOD) (http://linkeddata.org) is relevant to this NSF workshop as it makes many datasets openly available in a structured and interlinked form. However, before LOD can be used for SciSIP studies, it needs to be known who is exposing what data semantically, exactly what data is exposed, and what linkages exist. The provenance trail, i.e., what data came from what source and what changes/unifications/linkages were made, needs to be documented in an audible fashion. Below I provide a listing of the kinds of data I/others need to understand together with sample data formats.

People that serve LOD
Name | Institution | Contact info/email | Geolocation (ZIP if in US, city+country otherwise)

Datasets
Dataset Name | Original Source | URL | # Records | Link to raw data sample | Ontology/structure/data dictionary | topic coverage, e.g., medicine, CS | Type, e.g., patents, funding, genes | Available in LOD since when?
Are there also derivative datasets in LOD? For example datasets that add additional (calculated) values or unify names, geolocations, etc?

Services
The number of services that use a LOD dataset is a major indictor of its quality, reliability, and utility.
What tools/services use what datasets?
Service Name | URL | Type of functionality | Available since when?

People—Data Linkages
A listing of
People Name | Dataset Name
This will show who contributes how many datasets but also what datasets are served by multiple parties.

Data—Data Linkages
Dataset Name 1 | Dataset Name 2 | Mapped classes/attributes/linkages, etc. | #matches | # records in Dataset 1 | # records in Dataset 2
One row per mapping.

Data—Services Linkages
Dataset Name | Service Name

If this information can be acquired for LOD and non-LOD data then we can make informed decisions on what data to use for what type of SciSIP study.

(2) Electronically accessible publications

Please see (1) but I believe we need more than publications for SciSIP research. We need information on the input (e.g., funding, policies, jobs) and the output (e.g., publications, patents, datasets, software, tools, resources, expertise) of science. This information is commonly stored in data silos. However, we need to know, e.g., what students/Postdocs/staff and funding one faculty member attracted and what output s/he produced. Hence, the linkage of funding to publications (as provided in NIH's RePORTER and soon available for NSF data), the ARRA required reporting of jobs in academe (http://www.recovery.gov), or individual level data on people soon available via VIVO to name just a few relevant datasets, are essential.

(3) Digitally identifiable scientists

A recent NIH Workshop on *Identifiers and Disambiguation in Scholarly Work* at the University of Florida, Gainesville, Florida on March 18-19, 2010 (http://scimaps.org/flat/meeting/100318/) identified an impressive set of different identifiers and data structures that are currently used or planned to describe "people".

General and scholarly identifiers comprise: FOAF, OpenID, Federated identity management, InCommon, Medpedia, ORCID, VIAF, Marc Authority 12, CV Lattes, ISNI, national identification number scheme, Concept Wiki/WikiPeople, Amazon, Repee, ResearcherID (ISI), Scopus ID, Google Scholar, Citeseer, arxiv, VIVO, PubMedUMLS. Federal identifiers such as SSN, TIN, EIN, VISA numbers, PIV cards also exist but are less relevant here.

It seems impossible that all institutions, publishers, service providers will agree on one identifier. However, it is very possible that each researcher is assigned one ID whenever s/he publishes the first paper—analogous to getting a computer account, the author might go to the local/institutional library with his driver's license or other identification to receive this author ID. In addition, authors (using VIVO, WikiPeople, etc.) would provide "see also" links that interconnect IDs across data silos. For example, a researcher might add to his/her cv that his/her ID at IU is *aaa*, see also ID *xxx* in Scopus, ID *yyy* in ISI database, *zzz* in VIVO, etc.

Again, it will be important to know who added what data/link, i.e., the full provenance trail needs to be known.