Towards a Research Profiling Ecosystem Weaving Scholarly, Linked Open and Big Data

David Eichmann
School of Library and Information Science &
Iowa Graduate Program in Informatics
The University of Iowa

Research Profiling / Networking

Research profiling systems provide programmatic support for discovery and use of research and scholarly information regarding people and resources – essentially serving as special purpose institutional knowledge management systems. They have also achieved notable adoption by research institutions.[1] A number of systems have been developed, including open source (e.g., VIVO and Harvard Profiles), commercial (e.g., Elsevier Pure) and local institutional systems (e.g., Iowa's Loki and Stanford's CAP).

Multi-site search of research profiling systems has substantially evolved since the first deployment of systems such as DIRECT2Experts.[2] CTSAsearch is a federated search engine using VIVO-compliant Linked Open Data (LOD) published by members of the NIH-funded Clinical and Translational Science (CTSA) consortium and other interested parties. Eighty-seven institutions are currently included, spanning eight distinct platforms and three continents (North America, Europe and Australia). CTSAsearch has data on 174-421 thousand unique researchers (depending upon how you count) and their 10 million publications. The public interface is available at http://research.icts.uiowa.edu/polyglot.

Linked Open Data (LOD) holds substantial promise for tools supporting collaborative and translational science. The NIH-funded Clinical and Translational Science (CTSA) program has already proven to be a significant catalyst for tools supporting research discovery. Our work on extending Loki, the University of Iowa research profiling system, into the Semantic Web serves as a substantial case study in modular architectures extending into LOD.

The Loki Research Profiling System

Loki was developed as a component of the University of Iowa CTSA to support researcher discovery and collaboration. Comprised of investigator-authored research narratives coupled with publication data from MEDLINE and the Web of Knowledge, Loki's functionality expanded to include NIH funding opportunity awareness, demographics data from Human Resources and grant data from the Division of Sponsored Programs. Loki is investigator- rather than institutionally-focused, supporting multiple phases of the research life cycle, from funding opportunity identification (through NIH announcements), to team formation (through expertise search), to proposal creation (through biosketch management) to outcome dissemination (through automated inclusion of investigator publications). The modular nature of Loki's architecture has been a key element of this approach, consisting of

- A database layer, where each source is managed by a separate connector;
- A tag library layer, where each source is mapped into a suite of semantics-based tags; and
- A Java Server Page (JSP) presentation layer, where the semantic tags are woven into HTML and CSS elements to comprise a browser page.

From Tags to Triples

Subsequent work involved definition of a Loki ontology and the mapping of relational database entities into the resulting ontological concepts. Our approach of synthesizing the tag library layer of the architecture from an entity-relationship diagram proved substantially valuable in this work, as much of this mapping proved to be fairly formulaic through the use of the D2R relation to triple mapping tool. Furthermore, the clean partitioning of the logical components (e.g., demographics and publications) of the database layer allowed us to independently represent those components as discrete ontologies, and hence, discrete triple stores – supporting an overall LOD environment of interlinked triple stores that reflected the modularity of our initial tag library design.

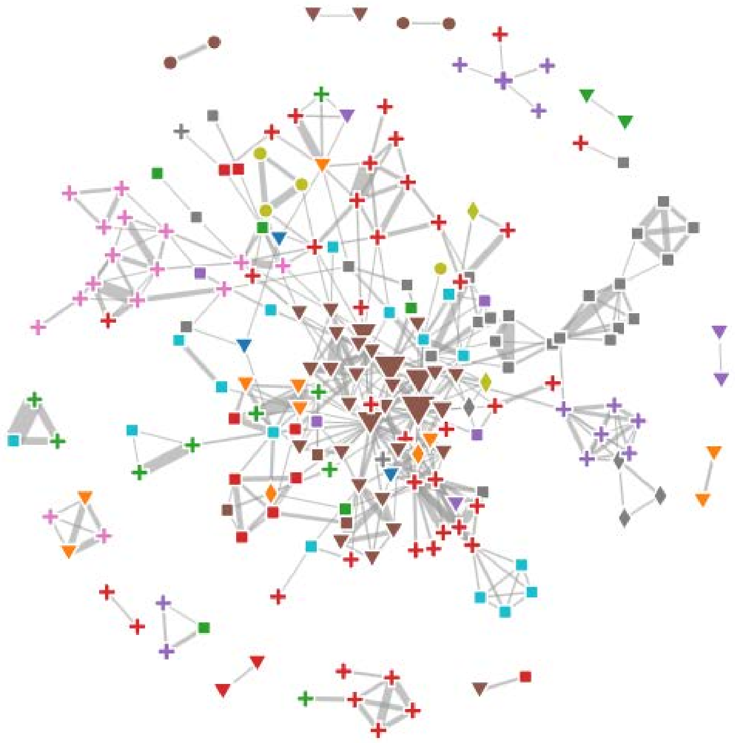
Ontological Mapping and Equivalences

The use of ontologies to model complex semantic relationships has become well-established, particularly in certain disciplines, such as biomedicine. Standardization on languages such as OWL have further demonstrated the utility and reusability of such formalisms. VIVO (the ontology) is an excellent example of community adoption of a shared semantic model, and projects such as CTSAsearch have demonstrated the potential for use of these models and the related data beyond that of the original context (i.e. VIVO the application).

As noted above, we opted initially to develop a Loki ontology that directly represents the semantics of our local environment. This was a conscious design decision, as we wished to demonstrate in a practical fashion that the LOD goal of concept mapping and equivalence was possible, and indeed desirable in this domain. We subsequently mapped the Loki ontology to the VIVO ontology within the D2R specification file purely at the ontological level, demonstrating the value in maintaining separation between the representational and conceptual levels in our overall information architecture. At the level of SPARQL query, Loki now is indistinguishable from a native VIVO instance.

CTSAsearch

CTSAsearch(http://research.icts.uiowa.edu/polyglot) is a federated search engine using VIVO-compliant Linked Open Data published by 87 institutions using eight distinct platforms. Since its introduction in 2013, the query and visualization mechanisms in



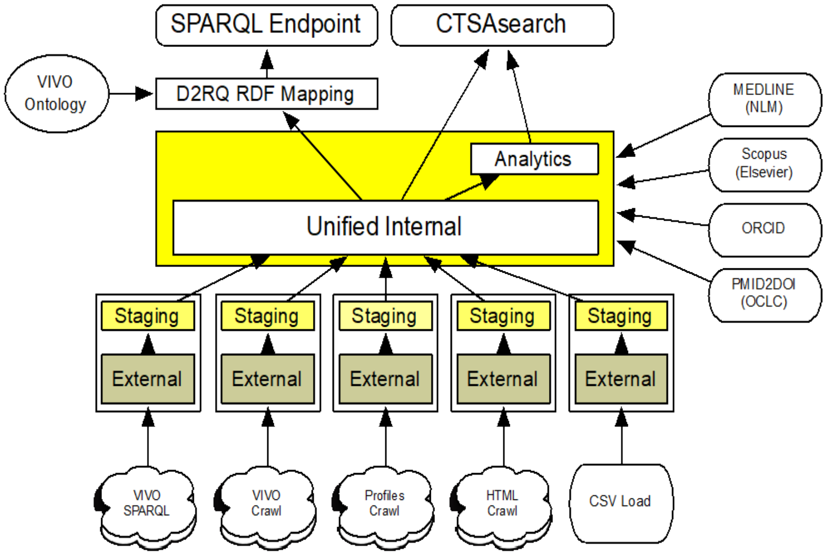
CTSAsearch have proven to be the primary elements of user interest. In particular,the coauthorship relationships between investigators at various institutions forms the principal visualization mechanism, as show in this figure. Each symbol indicates a particular investigator where the specific symbol indicates their home institution and the size of the symbol the relevance to the user's query. Edges between symbols indicate coauthorship with the thickness of the edge indicating the level of joint authorship.

Architecture

CTSAsearch draws both from research networking systems and from multiple data authorities. Given the diversity of information sources, modularity is critical to a robust, adaptable software architecture, as illustrated following.

CTSAsearch is currently comprised of the following:

- 1 VIVO-based SPARQL harvester
- 2 VIVO-based crawlers (due to differences in the respective ontologies
- 1 Profiles-based crawler
- 2 Platform-specific HTML crawlers
- 1 Proprietary API harvester (for Elsevier's Pure)
- 1 CSV-based loader



The resulting information space is comprised of 14.3 million VIVO v. 1.4 – derived triples, 129.3 million VIVO v. 1.6 – derived triples and 74.2 million Profiles-derived triples. The unified internal model aligns these representation variants into a single model which is used for indexing, retrieval and visualization.

Query Formulation using concept recognition

One of the early aspects of user feedback on CTSAsearch was a desire for more sophisticated search than that provided by a simple 'bag-of-words' relevance list. While this google-style search mode is available as an option, the default currently is one supporting full Boolean logic with a greedy concept recognizer processing the Boolean operands. For example, the query "stem cell & ferret" results in two operands, the first bound to UMLS concepts C0018956 and C0038250 (stem cell) and the second C0015859 (ferret). The recognizer aggregates the longest strings of tokens possible in each operand and the resulting concepts and any unrecognized strings are grouped as a Lucene query node. This approach has been very successful in pruning low relevance hits from results (e.g., matches on "cell" above). Finally, each Boolean operand is expanded with the subconcepts for each of its recognized concepts using the UMLS semantic network. This supports retrieval of profiles mentioning more specific descendant concepts by a more generic ancestor query concept.

Author-level co-authorship visualization

The co-authorship connections between the matched profiles are visualized using a force graph implemented in D3. Connections are pre-computed at profile harvesting time using multiple alternative identifiers (DOI, PMID, and PMCID) present in the profile data. OCLC pmid2doi crosswalk data is used to span the identifier spaces. As seen in the figure, useful force graph visualizations are possible for 'reasonable' result scales (n ~ 200). Challenges arise when results are larger – a query term such as "diabetes" returns thousands of results, leading to a network hairball.

Institution-level visualization

I have taken two different approaches to untangling the hairball. The first, and simplest, is aggregating results at the institution (i.e., VIVO instance) level. This clearly limits the number of nodes in the result to the number of VIVO instances for which I have data. However, for our diabetes query, there is little information discernable other than the degree of inter-institutional collaboration present for the topic. I am currently exploring the value of aggregation at smaller granularities (e.g., departments, institutes, etc.).

Inter-institutional community visualization

Focusing on community detection in the network structure is proving to be a far more robust approach to untangling large networks. I use a user-selectable set of community detection algorithms to aggregate community members into a single initial node, and then support zooming into an author-level visualization for a given community. I anticipate that this multi-scalar approach to visualization will accommodate scaling to entire research disciplines.

References

1. Obeid JS, Johnson LM, Stallings S, Eichmann D (2014) Research networking systems: the state of adoption at institutions aiming to augment translational research infrastructure. J Transl Med Epidemiol 2(2): 1026.
2. Weber GM, et al. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. J Am Med Inform Assoc. 2011 Dec; 18 Suppl 1:i157-60. doi: 10.1136/amiajnl-2011-000200. PubMed PMID: 220378.