VIVO: A Research Networking Tool That Provides Contextual Information about Researchers, Resources, and Research Datasets

Layne M. Johnson, VIVO Project Director, DuraSpace
Dean B. Krafft, Chief Technology Strategist Cornell University Library

Abstract

VIVO is an open source semantic web application for integrating and sharing information and data about researchers and their activities and outputs at a single institution while supporting discovery of related work and expertise across a distributed network. VIVO is fundamentally interdisciplinary; it enables and promotes the discovery of research and scholarship across traditional boundaries of geography, organization structure and type, academic or clinical or applied domain, technology, language, and culture. This position paper will introduce VIVO and its underlying data standards and ontology, and describe the current status of VIVO and other research networking installations around the world. Examples where VIVO serves to create context for researchers and datasets are also presented.

Introduction

VIVO[1] enables the discovery of research and scholarship across disciplinary and administrative boundaries through interlinked profiles of people and other research-related information. VIVO is populated with information about researchers, including data about publications, grants, teaching, service, and more. Data can be programmatically imported into VIVO from authoritative data sources such as institutional records, bibliographic and grant databases as well as structured data sources generated during the research process. VIVO supports open development through simple, standard semantic web technologies. By storing and exposing data as RDF (Resource Description Framework) on the web and by using standard, well-defined ontologies (e.g. the VIVO Integrated Semantic Framework or VIVO-ISF), the information in VIVO is displayed both as a human-readable web page or delivered to other systems as RDF. Researcher metadata in VIVO can be harvested, aggregated, and integrated into the Linked Open Data Cloud.

VIVO is important because it is the only standard way to exchange information about research and researchers across diverse institutions. It provides authoritative data from institutional databases of record as Linked Open Data (LOD). Structured VIVO data support search, analysis and visualization across institutions and consortia. VIVO is highly flexible and extensible to cover research resources, facilities, datasets, and more. Critically important is that the VIVO Community includes an international group of committed open source code developers who are continually creating new tools and applications to extend the utility of the VIVO platform.

The Adoption of Linked Open Data Standards Ensures Interoperability

Because VIVO data are available as LOD and organized according to formal ontological principles, other systems based on RDF triples (including some digital repositories and many research networking systems) can easily communicate with VIVO. VIVO data standards were endorsed by the NIH Clinical and Translational Science Awards (CTSA) Consortium in October,

2011 when the CTSA Executive and Steering Committee recommended that all CTSA-funded institutions (61 nationwide) should encourage their institution(s) to implement research networking tool(s) institution-wide that utilize RDF triples and an ontology compatible with the VIVO ontology. Further recommendations endorsed the general principle that information in individual profiles at institutions should be publicly available as LOD. To ensure high quality information, the use of authoritative electronic data sources rather than manual entry was also emphasized[2].

The VIVO-ISF Ontology: Representing and Sharing Data

The VIVO-ISF ontology is an OWL open source ontology resulting from a merging of the eagle-i[3] and VIVO ontologies together with additional content to represent clinical activities and expertise. The ontology content provided by the eagle-i ontology represents research resources including instruments, techniques, services, specimens, organisms, reagents, etc. The eagle-i ontology permits the representation and sharing of data from various institutions across the network of institutions running the eagle-i application or compatible systems[4].

The content from the VIVO ontology complements eagle-i content by representing researchers' activities such as publications, grants, research projects, academic and organizational positions, etc. The VIVO ontology is also integrated into the VIVO application to support the capture and sharing of this information between various institutions. The goal of the VIVO-ISF representation is to relate basic science research (the eagle-i based content), academic professionals (the VIVO based content), and clinicians profiled through the outcome of their clinical encounters. Another primary goal is to enable the creation of multidisciplinary research teams as well as making this data available for third-party applications.

Although the initial goal of VIVO-ISF has been to integrate clinical activities, research resources, and researchers and their activities to build meaningful biomedical context, further extension of VIVO-ISF is currently focused on permitting the integration of data from any number of non-clinical disciplines, including the physical sciences, library catalogue data, and the humanities, for example.

The State of Adoption of VIVO and Research Networking Tools

In 2012, a survey of the adoption of VIVO and other research networking tools in the CTSA consortium was conducted. Johnson, et al.[5] and Obeid, et al.[6] reported that of the 61 institutions surveyed, 51 had implemented research networking systems (22% VIVO, 23% Harvard Profiles[7], 22% Elsevier SciVal Experts[8], and 25% other systems [including homegrown or commercial platforms]). Forty-seven institutions reported that they had either implemented or planned to implement LOD and VIVO-compliant ontology standards. Roughly 25 institutions had already included or were planning to include researchers from outside the biomedical domains. Approximately 168,000 individuals were represented in the systems that had been implemented and a little over half of these researcher profiles were available in LOD format and were compliant with the VIVO ontology.

The implementation of open source, VIVO-compliant, collaboration systems continues to increase. Harvard Profiles, an open source profiling system that is LOD and VIVO ontology compliant reports a robust worldwide community[9], and there are currently over 150 VIVO efforts across almost 50 countries[10]. Furthermore, the creation of multi-institution networks provides

researchers with the capacity to discovery expertise and resources outside of the walls of their own institutions. The Direct2Experts[11] cross-institutional federated search tool lists 76 participating organization, whereas VIVOSearch[12], a fully integrated, cross-institutional aggregated search platform built on semantic web, LOD and VIVO ontology standards currently includes 7 participating institutions representing >50,000 individual profiles which include publications, activities, events, and courses. A significant expansion of this network is currently envisioned.

VIVO: Providing the Context for Researchers and Research Datasets

Context is critical to finding, understanding, and reusing research data. Contexts include: narrative publications; the researcher, research resources, grants, etc.; dataset registries; structured Knowledge Environments; and the web of Linked Open Data. The capability to expose integrated information about researchers, resources, and research datasets provides users with significant context and a rich illustration of relevant facets related to specific research processes.

Deep Carbon Observatory

The Deep Carbon Observatory (DCO) is a 10-year global research program to transform our understanding of carbon in Earth. The DCO is a community of scientists, involving many different scientific domains from biologists to physicists, geoscientists to chemists, and many others whose work forges a new, integrative field of deep carbon science. The DCO's infrastructure includes public engagement and education, online and offline community support, innovative data management, and novel instrumentation[13]. DCO's focus is on the terrestrial biological carbon cycle, where observations for research encompasses the entire planet from crust to core, from a few meters to a few kilometers beneath the surface, as well as the lower mantle and core—regions where carbon may play chemical roles that are yet to be discovered.

The Data Science team of the DCO project is responsible for building an integrative, linked presence of all the research from the various science domains. It is responsible for developing the infrastructure to support the different DCO teams and to help them collaborate by providing a common, public interface[15]. From there scientists can deposit different pieces of information, from publications to datasets to photos and more. To create the DCO platform, the Data Science team combined three open-source tools: the CKAN repository (ckan.org), VIVO researcher profiling, and the Handle identifier system (www.handle.net). Researchers will be able to discover other researchers, research, and research data through the DCO portal and the underlying LOD platform.

Laboratory for Atmospheric and Space Physics (LASP) at CU Boulder

LASP[15] began in 1948, a decade before NASA, and is the world's only research institute to have sent instruments to all eight planets and Pluto. LASP combines all aspects of space exploration through expertise in science, engineering, mission operations, and scientific data analysis. As part of Colorado University Boulder, LASP also works to educate and train the next generation of space scientists, engineers and mission operators by integrating undergraduate and graduate students into working teams.

LASP has created a Semantic Metadata Repository for Scientific Data through the use of VIVO and other open source tools. LASP projects consist of a variety of disparate web sites to provide access to information about specific missions or education and outreach. Keeping data set information updated and in sync across these web sites is a problem when the information isn't interoperable. To address this and other issues, LASP created a semantically enabled metadata repository that holds information about data and that now offers semantic browse and search capabilities, such as search of data sets by type (currently spectral solar irradiance, total solar irradiance, and solar indices) or over a particular spectral range provided by the user. Through the use of VIVO to create a semantic database and by extending two existing ontologies to meet space physics domain needs, LASP is achieving significantly increased capabilities at a relatively low cost. They believe that their approach could help projects with limited resources achieve similar capabilities to manage and provide access to metadata.

Other VIVO-related Dataset Initiatives, Future Directions

The DataStar[16] initiative is focused on research data curation and metadata reuse by translating discipline-specific metadata into RDF. The project focuses on datasets and the creation and sharing of metadata. DataStar provides options to upload and share datasets and enables research data discovery through dataset registry, sharing metadata across domains and supporting the linkage of research data and VIVO to support discovery and reuse. The linkage of datasets to VIVO profiles thus allows users to perform searches like "Get all datasets authored by individuals funded under NSF CISE grants".

As part of the Australian National Data Service (ANDS), three Australian universities, the University of Melbourne, Griffiths University, and the Queensland University of Technology, have been using the VIVO platform to create a Research Activity Hub and a Research Data Registry. Both applications link researchers and research activity to research data, and serve as part of the Australia-wide ANDS national research data commons[17].

Various other VIVO instances have leveraged the power of accessing research, resource and data set information to create context. As the significance of data-sharing mandates from funding institutions increases and the focus on reproducible experimentation gathers attention, the future of open source, interoperable, formally-organized collaboration tools that provide access to a relevant milieu of research information and datasets looks quite promising.

Links and References

1. http://vivoweb.org
2. http://www.ctsacentral.org/best%20practices/research%20networking
3. http://code.google.com/p/eagle-i/
4. https://www.eagle-i.net/
5. Johnson, L.M., Stallings S., Eichmann D., and Obeid J.S. (2013). Adoption of Research Networking Systems in the Clinical and Translational Science Award (CTSA) Consortium. AMIA Summits and Translational Science Proceedings, (18):93.
6. Obeid, J.S., Johnson, L.M., Stallings S., Eichmann D. (2014). Research Networking Systems: The State of Adoption at Institutions Aiming to Augment Translational Research Infrastructure. Journal of Translational Medicine and Epidemiology. (Accepted for Publication June, 2014)
7. http://profiles.catalyst.harvard.edu/?pg=home
8. http://www.elsevier.com/online-tools/research-intelligence/products-and-services/scival
9. http://profiles.catalyst.harvard.edu/?pg=community
10. Personal Communication, K. Holmes, Northwestern University, VIVO Engagement Director. June 7, 2014.
11. http://direct2experts.org/
12. http://beta.vivosearch.org/
13. http://deepcarbon.net//about/about-dco
14. http://tw.rpi.edu/media/2014/01/23/c155/DCO-DS-Overview-Demonstration-VIVO20140123.pptx
15. http://lasp.colorado.edu/home/
16. http://datastar.mannlib.cornell.edu/
17. http://vivoweb.org/files/AustralianCommunity.pdf