Research Graph VIVO Cloud Pilot

Abbreviated Proposal

Introduction

Many VIVO implementers find collecting, mapping, and loading data into VIVO to be quite difficult. For example, data on publications, grants, and datasets produced by an institution's faculty can be difficult to find and disambiguate. Understanding the ontologies used to describe data in VIVO and mapping faculty data to those ontologies involves a steep learning curve. Also, transforming the data to a linked data format, such as VIVO RDF, has proven difficult for most implementers due to gaps in skills and knowledge. These barriers have prevented organizations from joining the VIVO community and adopting the technology that enables access, discovery, and analysis of scholarship data.

Research Graph[1] is an integrated network of information about researchers, their publications, grants, and datasets, across global research infrastructures such as ORCID, DataCite, CERN, CrossRef, and funders such as National Institutes of Health (NIH). The Research Graph network currently connects more than 20 million research object across Australia, Europe, United States and Japan.

For example, when provided "seed data," such as a simple list of researchers, Research Graph will identify publications, grants, and/or datasets related to those researchers and represent the information in a graph. These are referred to as "first order" connections. Research Graph is also capable of identifying and linking collaborators of the people in the "first order" data and linking their publications, grants and datasets. These collaborator links are referred to as "second order" connections. We are not aware of any other technology or product, open source or proprietary, that can offer "second order" connections.

A recent collaboration between VIVO and Research Graph[2] developed and demonstrated a repeatable process for using seed data to build first and second order graphs, and to export, transform, and load those graphs in VIVO RDF format to a hosted VIVO instance.
This process has been accomplished by using Research Graph. The outcome has the potential to resolve common difficulties experienced while finding, disambiguating, transforming, and mapping data for ingest into VIVO. Figure one shows the outcome of the augmentation process for the data from the National Computational Infrastructure in

[1] Research Graph Home Page. http://researchgraph.org. Accessed December 2, 2017.
[2] Conlon, Michael, and Amir Aryani. "Creating an Open Linked Data Model for Research Graph Using VIVO Ontology," July 24, 2017. https://doi.org/10.4225/03/58ca600d726bd.

Australia.

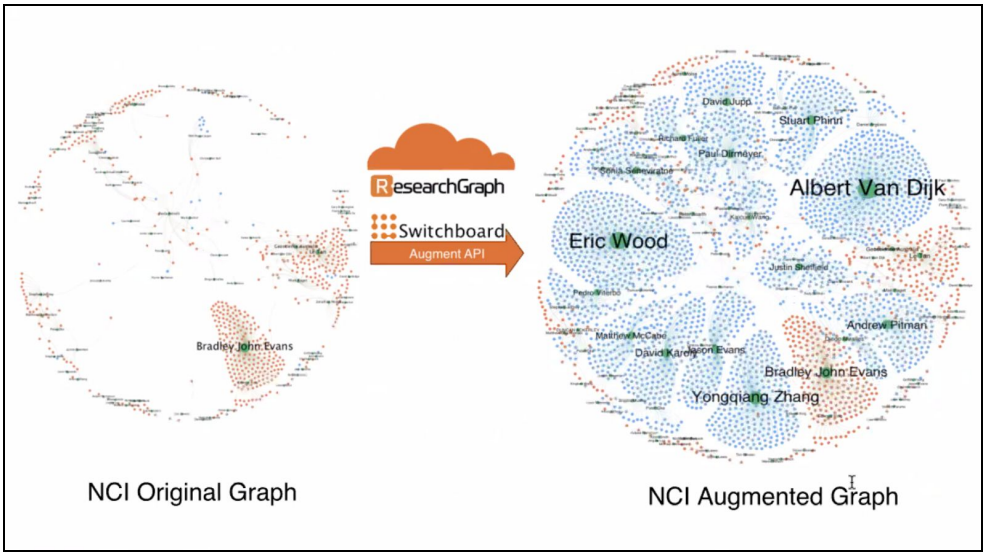

Figure One: Visualization of seed data from the National Computational Infrastructure before and after augmentation via Research Graph.

There are three types of organisations that can benefit from the collaboration between VIVO and Research Graph.

- Type A (Repositories, Research Institutes): These institutions can benefit from transforming their seed data (people, and/or publications) to an augmented graph of people, publications, grants, and datasets. Furthermore, these institutions will benefit from the smooth transformation of their augmented graph to a VIVO instance with no requirement of high technical expertise.

- Type B (Semantic Web Sites): Government and research organisations such as GeoScience Australia who already have their research data in RDF format can use VIVO as a data exchange platform between Research Graph and their semantic web system. Furthermore, these sites will be able to take advantage of VIVO visualisations of their collaboration network -- the capability that is often missing from their local RDF-based research system.

- Type C (Current VIVO Sites): Using the connection between VIVO and Research Graph, these sites can enrich their data and ingest new information into their systems.

VIVO's Project Director, Dr. Mike Conlon, Research Graph's Director, Dr. Amir Aryani, and more recent collaborator, DuraSpace's Business Development Manager, Erin Tripp propose a joint Research Graph VIVO Cloud Pilot project. It will investigate how a pilot organization can provide seed data and have the joint partners of the Cloud Pilot produce a fully populated, turn-key, hosted VIVO with linked researchers, publications, grants, and datasets that can be access and searched. A successful Cloud Pilot project will determine

the value and potential of a long term collaboration between VIVO and Research Graph in the form of new services that could reduce barriers for organizations that want to implement VIVO.

Potential benefits of the Cloud Pilot include:

- Development of a process for producing a turn-key hosted VIVO based on Research Graph data, significantly reducing the effort of new implementers
- Production of linked data that could be used by the pilot organization for any purpose, including augmenting an existing VIVO, other research information, or data analysis systems
- Opportunity for analysis by network researchers at pilot organizations to study the outputs related to specific research areas/efforts
- Introducing Research Graph's "second order" connections to the VIVO community. This provides not only the entities related to the seed data, but also entities related to the entities in the seed data. Second order data analysis can provide answers to questions such as "who do my collaborators collaborate with?" This is a tremendous advantage in the competitive landscape
- Potential to greatly improve the adoption of VIVO, and grow the VIVO community
- Development of technical automation to form the basis of services that can be provided to the VIVO community that supports the production of VIVO data and/or provides VIVO hosted software

An Expression of Interest (EOI) notice soliciting participation in a Research Graph VIVO Cloud Pilot was distributed at the Open Repositories conference in June 2017 as well as at the eResearch conference on October 2017, both located in Australia. The notice was also distributed to peers in Germany and Canada. Four organizations formally expressed interest in participation. Another five organizations informally expressed interest in participating..

Pilot Phases and Structure

The organizations that expressed interest in the Research Graph VIVO Cloud Pilot have varying levels of knowledge of the projects and technologies. Two of the organizations are extremely knowledgeable and involved in the VIVO and Research Graph communities. They are located in Australia and Germany.

The collaborators on this project recommend taking a phased approach to the pilot, starting with the two organizations that are most knowledgeable and skilled in phase one and working with less knowledgeable and skilled organizations in phase two. We recommend the formation of a Working Group to implement the Cloud Pilot including the

Pilot Team and representatives from each pilot organization.

This first phase will enable the Cloud Pilot Team to confirm what they consider to be the most important and unknown technical variable of the project, e.g. how large the seed data will grow after first and second level connections are made and what impact that will have on hosted server resources, cost, and performance.

The Cloud Pilot will also provide excellent information on scaling the service. Using subsets of Pilot Organization data, we can identify small to medium graph sizes. Using their entire faculty as seed data, we can load and scale test to large sizes. The organizations identified for the first phase of the Cloud Pilot are thought leaders in linked data, repositories, and the open science community. Their involvement lowers project risks and helps build understanding of a potential international service offering.

Pilot Deliverables

A successful, multi-phased Cloud Pilot will produce information suitable for a go-no go decision regarding the creation of a joint Research Graph VIVO Cloud Service, including:

- A formal service definition
- Market analysis, including what we consider acceptable annual pricing. Anecdotally, we see evidence of demand for such a service, particularly among US and European research institutes with smaller staffs, and universities and institutes interested in VIVO concepts, but unwilling to master VIVO technologies (ontologies, triple stores, and linked data), and sites seeking outsourced services
- A go-to-market pitch describing the key value positions for go-to market
- Technical pipeline and deployment model for the production of VIVO data and hosted VIVO sites. A deployment model under consideration is a full pipeline at Research Graph, including VIVO hosting. Duraspace would handle promotion, sales, customer relations, and billing
- Cost model for one-time and annual pricing for small, medium, and large size seed data sets. This includes analysis of the resulting graph size from various seed data sets
- Staffing model, describing the roles which would be required to set up, operate, and support the service, and the division of labor between Research Graph and DuraSpace
- A recommendation regarding go-to market decision

Pilot Assumptions and Risks

- Market analysis[3] -- We assume that there is a market for the production of VIVO data and turn-key hosted VIVO sites at a reasonable price. However, the market analysis may indicate there is no market for the service

- Technical effort -- Duraspace is currently gapped with respect to VIVO technical knowledge due to staff turnover. We assume this gap can be quickly filled by existing Duraspace staff, assisted by Dr. Conlon

- Customization -- we assume that the customer can be satisfied with simple theming (colors, logo, site name) of the turn-key hosted site, and that the theming can be delivered at reasonable cost

- Graph size -- the number of entities and triples -- resulting from particular seed datasets is not well understood. Based on the experience of large VIVO sites (Duke, Vidwan, Florida), we do not expect this to be an issue, even for large seed datasets (10,000 researchers). We propose a limit on the size of the final graph to be 500,000 entities for the purposes of this Cloud Pilot

- Data value -- the data produced by Research Graph must be of high value to the customer, including significant coverage and accuracy of first and second order entities

- Team member commitment -- we assume that the Pilot Team members (see below) and pilot organizations can provide the required effort in the required timeframe to participate in a Cloud Pilot Working Group

Proposed Pilot Timeline and Effort

We propose a three month term for phases one and two, running from February 2018 to April 2018. This will provide suitable results for a presentation at the 11th RDA Plenary Meeting in Berlin, Germany, March 21-23, 2018.

The following effort is estimated for the Pilot Team during the term: Market analysis and service definition lead (10%), Project manager (10%), VIVO subject matter expert (10%), Duraspace technical resource (20%), Research Graph subject matter expert (5%), and a Research Graph technical resource (20%). We recommend The Pilot Team meet with pilot organizations weekly during the term, forming the Cloud Pilot Working Group[4].

[3] The pilot can be conducted with its pilot sites, a market analysis conducted, and deliverables successfully produced without risk to the pilot itself. The risk is to the resources used in the pilot, which may be to no avail if the market analysis indicates that there is no market for the service.
[4] Pilot sites may wish to commit additional effort in order to achieve local goals deliverables.