Using OAI-PMH protocol to Data Ingest into VIVO Instances

Alexandre Rademaker* Violeta Ilik†

October 30, 2014

One main motivation for the adoption of VIVO by universities and research institutions is to create one central website that integrates information about researchers' scholarly outputs, such as publications, grants, activities, and researchers skills. This is one of the main reasons why VIVO is denoted as a research-focused discovery tool that enables collaboration among scientists across all disciplines. The discovery of trustworthy sources of data and the integration of these data sources is a challenge that all institutions and universities have to face in order to have a VIVO instance fully deployed. Digital libraries, or institutional repositories are among the sources of information that universities can use to populate their VIVO instances.

Getulio Vargas Foundation (FGV) started a project in 2010 with a goal to integrate information about researchers, publications, and academic activities. The current FGV VIVO instance [1] is in a pilot phase and contains 23K people and 72K research items. The data used to populate FGV's VIVO instance is collected from two main sources: (1) the Brazilian Lattes Platform [2]; and (2) The FGV's Digital Library.

The Lattes Platform is an online system used by almost all researchers in Brazil to maintain their curriculum vitae. Created by CNPq (National Council for Scientific and Technological Development) in the mid-80s, the platform is an instrument that guides investments in research in Brazil and evaluates the Brazilian research community. The Lattes data is available for institutions through a Web Service in a XML format. In the first part of the project we developed an XSLT stylesheet for transform a curriculum vitae file in XML to RDF using standard ontologies as FOAF, Bibo, SKOS etc. This transformation is released as an open source project called "Semantic Lattes". [3]

The FGV's Digital Library is comprised of two main systems: (1) The FGV's Digital Collections [4] which uses DSpace Platform [5] as digital repository management system; and (2) The FGV Journals Portal, which uses the Open Journal System. [6] Both systems implement the OAI-PMH protocol.

The Texas A&M University also has an ongoing project to implement VIVO. As FGV, Texas A&M also has an institutional repository based on DSpace architecture. Texas A&M institutional repository contains researchers' publications that are bound for integration with the local VIVO instance. [7] This give us a motivation to jointly investigate methods for data retrieval from OAI-PMH compatible sources for ingest into our VIVO instances.

Given that in general all digital libraries and institutional repositories are curated by librarians, publications data and authors names data is considered authoritative.

With this presentation, we aim to demonstrate the tools developed by FGV and Texas A&M for data retrieval from DSpace instances using OAI-PMH protocol, transforming the data into VIVO compatible RDF using XSLT and ingesting the data into VIVO using the VIVO SPARQL Update API. We plan to discuss the benefits of such tools, the current workflows at our two institutions, and the limitations.

*PhD, School of Applied Mathematics/FGV and IBM Research, RJ, Brazil.

†MLIS, Texas University, TX, USA.

[1] http://logics.emap.fgv.br:8080/vivo/

[2] http://lattes.cnpq.br

[3] github.com/arademaker/slattes/

[4] http://bibliotecadigital.fgv.br/dspace/

[5] http://dspace.org

[6] http://pkp.sfu.ca/ojs/

[7] http://repository.tamu.edu/