An Automated System for Tracking the Growth of Expert Profiling Systems

Robert P. Light
Cyberinfrastructure for Network Science Center, SLIS, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
1-812-856-3465
lightr@indiana.edu

Chin Hua Kong
Cyberinfrastructure for Network Science Center, SLIS, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
1-812-856-5417
kongch@indiana.edu

Katy Börner
Cyberinfrastructure for Network Science Center, SLIS, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
1-812-855-3256
katy@indiana.edu

ABSTRACT

Over the last decade, a number of expert profiling systems have been developed. Some systems are proprietary, e.g., Elsevier's SciVal experts, others are open source, e.g., the NIH funded VIVO system. Analogous to Facebook and LinkedIn, these systems make it possible to create personal profiles interlinked by professional and collaborative ties. They differ from social networking sites in that much of their data comes from high quality institutional repositories and they offer novel functionality for researchers, e.g., printout of resumes, and science policy makers, e.g., visualizations of expertise profiles or collaboration networks. Given the value of high quality expertise profiles, there is a great interest in tracking the spread and growth of these systems over time and geographic space. This paper describes the International Researcher Network site that was developed to track and visualize basic characteristics of major expert profiling instances to communicate the quality and coverage of their data holdings. The site's data is automatically gathered by the Expert Profile Resource Counter. A Google map interface supports the interactive exploration of the data.

Categories and Subject Descriptors

Semantic Networks

General Terms

Measurement, Documentation

Keywords

VIVO, Semantic Web, Expert Profiling

1. INTRODUCTION

The last ten years have seen the rapid emergence of expert profiling systems, web-based services used to describe researchers and their work. There are a multitude of systems available today, with three of largest being VIVO, Harvard Catalyst Profiles, and Stanford CAP. VIVO is an open-source semantic web application designed to facilitate the discovery of research and researchers and encourage collaborative efforts.[1] VIVO was founded in 2003, with local deployment at Cornell in 2004. With NIH funding for a collaborative development effort, VIVO has been deployed at dozens of institutions around the world. Harvard Catalyst Profiles was originally developed under Dr. Griffin Weber, also with support from NIH. Today it serves universities and research organizations worldwide. Stanford's Community Academic Profiles was announced to the world in October 2011 and serves the faculty of Stanford University exclusively. While there are efforts to build federated searching tools, both within and between these systems, questions of growth have to date focused on the number of instances a particular system has implemented rather than the quality of data within those systems and their growth over time. The International Researcher Network site[2] was created in an effort to compile and visualize this information. Initially, all relevant data was manually collected and recorded. Recently, an Expert Profile Resource Counter was implemented to remedy this situation by leveraging standardized templates within the various expert profiling systems to automatically acquire and record the number of basic types of entities in a set of documented instances.

The remainder of the paper is organized as follows: Section 2 provides background information on different expert profiling systems. Section 3 introduces the International Researcher Network site together with the Expert Profile Resource Counter. Section 4 reports first results of a meta-analysis of the data holdings by different expert profiling systems. Section 5 concludes the paper with a discussion of future work.

2. BACKGROUND INFORMATION

The International Researcher Network site supports four expert profiling system types that are detailed here. All these systems aim to support researchers to discover others who may have similar interests or to find researchers with completely disparate skillsets in order to form collaborations. Institutional administrators can use these systems to gain an understanding of the volume and type of work being done across the various departments and organizations of the institution, e.g., they can track the changes in the level of funding intake or research output over time. They can also see which departments are successfully forming collaborations as opposed to those in which the researchers are largely working independently. Institutional and national science leaders can use the systems to determine how effective efforts to encourage collaboration beyond the institutional level have been. They can see whether scientists at a given institution keep to themselves, work only with a few preferred other groups, or form many linkages across the country or around the world.

2.1 VIVO

The VIVO System and VIVO Ontology have been in development since 2003 as a way to semantically describe the work of academic researchers [1]. With funding from the National Center for Research Resources at the NIH, VIVO development has been led by Cornell, the University of Florida, and Indiana

[1] http://vivoweb.org/about

[2] http://nrn.cns.iu.edu

University[3]. VIVO is primarily implemented in Java, using MySQL as the database platform for the triple store. VIVO is an open source project that encourages outside developers to participate in creating data integration tools, applications and widgets. The ontology includes classes that describe many types of academic output in an effort to thoroughly describe the work and expertise of a given researcher, as well as to enable users to explore the many types of connections that a researcher can form during his or her career. Data described in the VIVO Ontology is of value to a number of members of the research community.

VIVO offers a number of visualizations of the data in the instance. Co-authorship and co-investigator networks show a given individual's links to other researchers [2]. Temporal graphs show the evolution of a person or groups output over time. VIVO also, uniquely, offers science mapping of publications based on the UCSD Base Map of Science [3]. This process maps the publications based on the publication venue (journal or conference) to one of 554 subdisciplines and then displays this on a set basemap. With this, users can view and compare the expertise profiles of a person, department, or university.

2.2 Harvard Catalyst Profiles (HCP)

Harvard Catalyst Profiles is a system implemented using the Profiles RNS software[4]. This open source platform utilizes the VIVO Ontology for researcher description and offers algorithms that weight the importance of works to a researcher's career based on attributes like age, position in the author list, and citation information. It also includes a 'Disambiguation Engine' designed to help reconstruct a publication history. HCP systems allow for cross-institutional searching and are currently piloting a system to allow for searching across other systems that utilize the VIVO ontology.

2.3 SciVal Experts

SciVal Experts is Elsevier's foray into the field of expert profiling[5]. This closed-source, commercial system is designed to be fully compatible with the VIVO ontology in support of data interoperability. This system offers visualizations showing a researcher's topical interests over time, as reflected in their publications, as well as offering a listing of "Similar Experts" that share research interests.

2.4 Other Systems

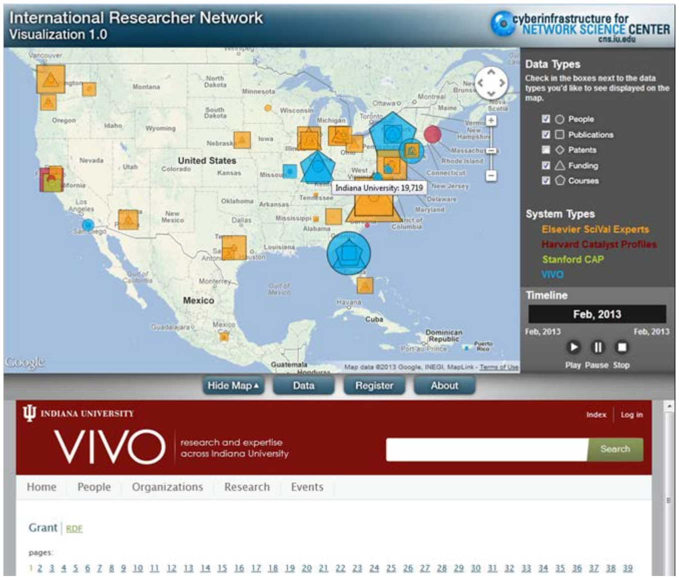
There are currently at least 40 other profiling systems in existence[6]. Several of them have exactly one instance, e.g., the Stanford Community Academic Profiles (CAP) system developed and deployed at Stanford University, the LOKI system at University of Iowa, or LatticeGrid by Northwestern University. Others are under development. To our knowledge, the three above mentioned systems are the only ones that are VIVO ontology compliant. Great care has to be taken in drawing data from non-compliant instances to ensure that the data gathered measures the same semantic concepts.

NIH has also funded DIRECT2Experts[7] through a Clinical & Translational Science Award. This cross-instance search system can perform a federated search across registered instances using all of the above systems (VIVO, SciVal, HCP and several others). DIRECT2Experts lists 47 participating organizations as of December 2012, though the site notes that not all have research tools connected to the system yet.

3. INTERNATIONAL RESEARCHER NETWORK PORTAL

3.1 Goals

The International Researcher Network site (http://nrn.cns.iu.edu) was created in 2010 to track the growth of expert profiling systems. It visualizes expert profile instances using a proportional symbol map of the world. Changes over time can be animated over time from 2010 to date, see Figure 1. Different data types and system types can be selected. Hovering over a symbol brings up a count. Clicking on an individual profiling instance brings up the respective profiling site, e.g., the Funding site at Indiana University in Figure 1. Managers of new sites can register their instances so that they show on the map.



Check in the boxes next to the data types you'd like to see displayed on the map.

Figure 1. International Researcher Network Homepage[8]

3.2 Data Coverage and Cleanliness

While VIVO was the first semantic web application developed for expert profiling, a number of similar projects have taken root in the last several years, including efforts by Harvard, Stanford and Elsevier. In order to determine whether one type of system is more prone to seeing continued use and growth than others, multiple types of systems need to be tracked over time.

While the three primary expert profiling systems make use of the VIVO Ontology, not all instances have implemented it in a consistent manner. For example, the University of Melbourne instance only includes Faculty Members, while the American Psychological Association instance has only one entity classified as a Faculty Member. Another example is how publications are

[3] http://mannlib.cornell.edu/news/nih-grant-funds-networking-scientists

[4] http://profiles.catalyst.harvard.edu/

[5] http://info.scival.com/experts

[6] http://en.wikipedia.org/wiki/Comparison_of_Research_Networking_Tools_and_Research_Profiling_Systems

[7] http://direct2experts.org/?pg=about

[8] http://nrn.cns.iu.edu/

counted: In most cases the VIVO Ontology is followed, with the Article as the main publication class, with Academic Articles as a subset of these, but in others, most publications are listed only as Academic Articles, with the Article as a completely separate and much smaller class. The Expert Profile Resource Counter must be able to accept settings for each instance in order to gather meaningful data from instances that may interpret the ontology in different ways.

3.3 Expert Profile Resource Counter

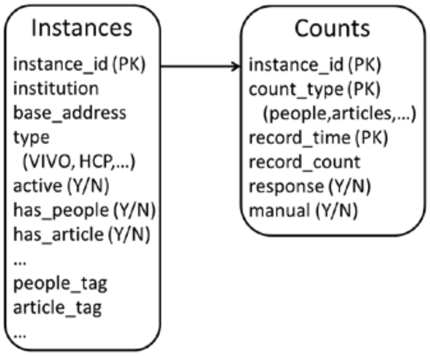
This section presents the Expert Profile Resource Counter that automates the update of the data that powers the International Researcher Network. The primary purpose of the system is to count basic entity types (researchers, articles, grants, and courses) across a variety of semantic expert profiling instances.

3.3.1 Software Used

The Expert Profile Resource Counter was implemented in Python 2.4.3 using a PostgreSQL 8.4.5 database as the primary data store. In both cases, the choice of platform was driven primarily by previous software deployment decisions on the servers to be used. Standard Python packages were utilized whenever possible, with the psycopg2[9] bridge to PostgreSQL being the lone exception.

3.3.2 Database Schema

The database for the Expert Profile Resource Counter has to serve two functions. First, it must keep a record of the instances to be tracked. Secondly, it must store the collected counts from those instances. This is achieved with a simple two table structure, as shown in Figure 2.



Instances
instance_id (PK)
institution
base_address
type
(VIVO, HCP,...)
active (Y/N)
has_people (Y/N)
has_article (Y/N)
...
people_tag
article_tag
...

Counts
instance_id (PK)
count_type (PK)
(people,articles,...)
record_time (PK)
record_count
response (Y/N)
manual (Y/N)

Figure 2. Database Schema

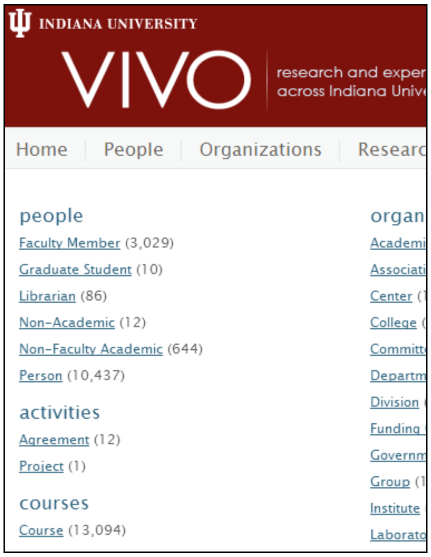
3.3.3 HTML Scraping and Parsing

The Expert Profile Resource Counter relies on the consistencies in the structure of the overview pages for the various types of expert profiling systems to parse out the relevant data. Different parsers are detailed here.

VIVO

In VIVO, the critical data is found on a page that breaks down all entities by type. While these pages can look very different, based on the CSS template, the underlying HTML is identical. See Figure 3 for an example.

VIVO HTML data is acquired using the python urllib package and its urlopen method. This data is then parsed with the HTMLParser package. By focusing on the "siteMap" class, the parser can scan for the appropriate label and then extract the associated number.



INDIANA UNIVERSITY
VIVO
research and exper
across Indiana Univ

Home | People | Organizations | Researc

people
Faculty Member (3,029)
Graduate Student (10)
Librarian (86)
Non-Academic (12)
Non-Faculty Academic (644)
Person (10,437)

activities
Agreement (12)
Project (1)

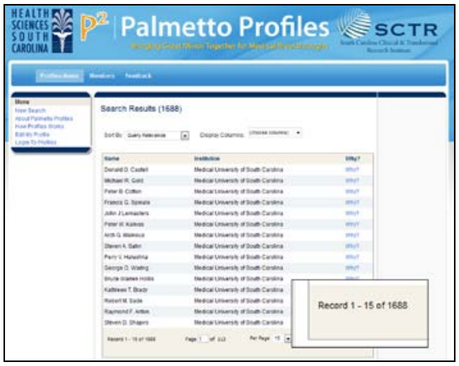
courses
Course (13,094)

organ
Academi
Associati
Center (
College (
Committ
Departm
Division (
Funding
Governm
Group (1
Institute
Laborato

Figure 3. Indiana University VIVO Instance Screenshot

Harvard Catalyst Profiles (HCP)

In HCP, the number of people is available by passing a blank search into the search engine. This provides a table with a standardized format. By parsing out the number of records from the bottom of the table, the number of person entities can be retrieved. In Figure 4, the South Carolina Clinical & Translational Research Institute site is shown as an example, but all HCP instances share a similar table with similar code.



HEALTH SCIENCES SOUTH CAROLINA
P2 Palmetto Profiles
SCTR
South Carolina Clinical & Translational Research Institute

Profiles Home | Members | Feedback

Menu
New Search
About Palmetto Profiles
How Profiles Works
Edit My Profile
Login To Profiles

Search Results (1688)

Sort By: Query Relevance
Display Columns: (choose columns)

Name	Institution	Why?
Donald O. Castell	Medical University of South Carolina	Why?
Michael R. Gold	Medical University of South Carolina	Why?
Peter B. Cotton	Medical University of South Carolina	Why?
Francis G. Spinale	Medical University of South Carolina	Why?
John J Lemasters	Medical University of South Carolina	Why?
Peter W. Kalivas	Medical University of South Carolina	Why?
Arch G. Mainous	Medical University of South Carolina	Why?
Steven A. Sahn	Medical University of South Carolina	Why?
Perry V. Halushka	Medical University of South Carolina	Why?
George O. Waring	Medical University of South Carolina	Why?
Bruce Warren Hollis	Medical University of South Carolina	Why?
Kathleen T. Brady	Medical University of South Carolina	
Robert M. Sade	Medical University of South Carolina	
Raymond F. Anton	Medical University of South Carolina	
Steven D. Shapiro	Medical University of South Carolina	

Record 1 - 15 of 1688 Page 1 of 113 Per Page 15

Record 1 - 15 of 1688

Figure 4. South Carolina Clinical & Translational Research Institute HCP Search Results

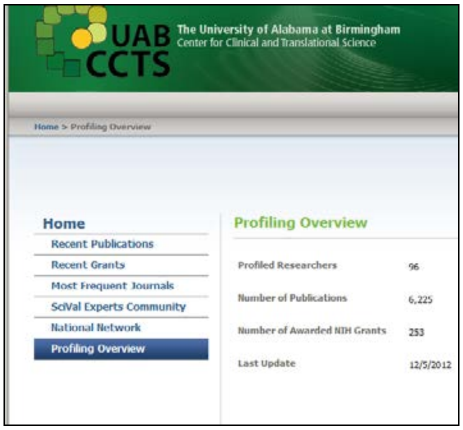
When parsing the HTML from HCP instances, the critical segment is at the bottom of the search results table, and is identified by the span ID "ctl00_ctl00_middle_MiddleContentPlaceHolder_grdSearchResults_ctl18_lblPageCounter". This number is also displayed at the top of the table in HCP, but this particular usage was more likely to be missing as implementing institutions modified the template to their tastes.

SciVal Experts

SciVal Experts instances provide the most straightforward summary of the available data from the "Profiling Overview" page. The information available here varies from instance to instance and can include counts of people, publications and grants. The example used in Figure 5 is from the University of

[9] http://initd.org/psycopg/docs/

Alabama-Birmingham, but almost all pages are identical in HTML structure. It is worth noting two things with this particular system, though. The Overview page seems to be optional, as one instance had removed it altogether, and it is very common to see the wording on the individual lines changed from instance to instance. As this wording is used to determine which numbers represent what type of data, it is critical to set these phrases correctly prior to scraping.



Profiling Overview

Profiled Researchers	96
Number of Publications	6,225
Number of Awarded NIH Grants	253
Last Update	12/5/2012

Figure 5. University of Alabama-Birmingham SciVal Experts Profiling Overview

In this case, the critical table is denoted by the div id “innerContentOneArea” in the HTML. Within that div, each line on the table has two data values, with the first indicating the type of data and the second giving the number of entities.

4. RESULTS

The International Researcher Network site gets about 6000 hits each month. It supports the study of the evolution of researcher networking sites over the last 3 years.

The Expert Profile Resource Counter is currently reading data from 44 different expert profiling instances, out of the 54 known to the authors at the time of its creation, collecting 117 different measurements per run. The run time for the entire set is approximately 20 seconds. Data from the remaining 10 instances is collected by hand on a monthly basis. The most recent totals are shown in Table 1 below and growth over time is shown in Figure 6, right. Note that an instance’s totals are only added to the graph from the point where the NRN team became aware of its existence.

Table 1. Entities tabulated during Expert Profile Resource Counter session 12-10-2012*

	VIVO	SciVal Experts	HCP	Total
People	262,596	44,817	11,346	**318,759**
Articles	269,522	1,604,519	N/A	**1,874,041**
Funding	54,432	89,417	N/A	**143,849**
Courses	38,963	N/A	N/A	**38,963**
Total	**625,513**	**1,738,753**	**11,346**	**2,375,612**

*Only those entities counted via the automatic parser are included in the table.

Of those that are not currently being collected, 3 are inactive, 5 use systems that are not covered by the parser, and two have made template changes such that the parser cannot retrieve the needed data.

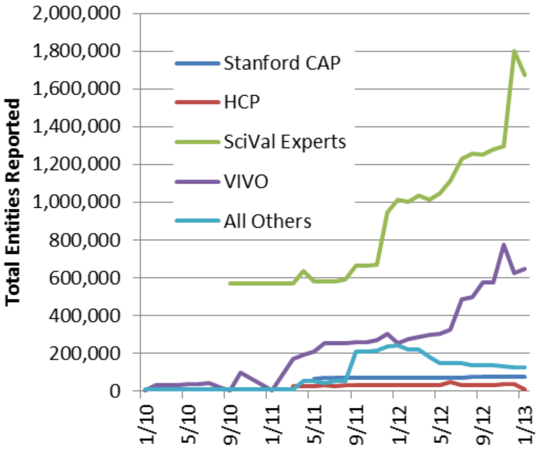


Figure 6. Total number of entities reported over time by system type

5. FUTURE WORK

As the number of researcher profiling sites grows, the Expert Profile Resource Counter becomes an essential and effective tool for keeping the International Researcher Network site up to date. Planned improvements comprise: Revised error logging and handling allowing for the system to handle malformed HTML rather than being forced to skip over such instances. An error may suggest that an instance has been closed or moved. A drop in the number of entities may suggest that the data in an instance has been restructured and that its tag entries need to be updated.

The system can also be expanded to allow for more robust record collection. While the average VIVO instance collects dozens of different types of data, aggregate data on only a handful are actively collected at this time. As SPARQL endpoints become available, those could be used to harvest data.

The field of expert profiling is still evolving and the coming years will certainly see new systems emerge into the market, such as LabRoots. As new instances are registered, system-specific parser functions will need to be implemented to handle these nascent technologies.

6. ACKNOWLEDGMENTS

VIVO has been funded by the National Institutes of Health, grant number U24-RR029822. This project was carried out as part of a graduate project under the auspices of Indiana University and with the support of the Cyberinfrastructure for Network Science Center.

7. REFERENCES

[1] Mitchell, S. et al. (2011). The VIVO Ontology: Enabling Networking of Scientists . *Proceedings of the ACM WebSci'11* (pp. 1-2). Koblenz, Germany: ACM.

[2] Börner K, Conlon M, Corson-Rikert J, Ding Y. (Eds.) (2012) *VIVO: A semantic approach to scholarly networking and discovery.* San Rafael, Calif.: Morgan & Claypool

[3] Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, et al. (2012) Design and Update of a Classification System: The UCSD Map of Science. PLoS ONE 7(7): e39464. doi:10.1371/journal.pone.00394