Semantic-based Researcher Profile Management System: A Case Study on VIVO

Mr. P Kannan, Scientist C (LS)

Abstract

The Internet has provided ample opportunity to this generation of scholars to communicate, share and discover information. Development of profile management system is crucial to organise research activities of an organisation and showcase them to the peer group. The semantic web is the extension web standard to organise the information through common data format and exchange protocol. This article explains about the semantic technologies such as Resource Description Framework, Web Ontology Language, Open Linked Data. The purpose of this article is to discuss about what is profile management system, how it is used in the scholarly communication, etc. The article emphasis on semantic-based profile management system called: VIVO and its architecture, data integration tools and major features and functionalities, etc.

1. Introduction

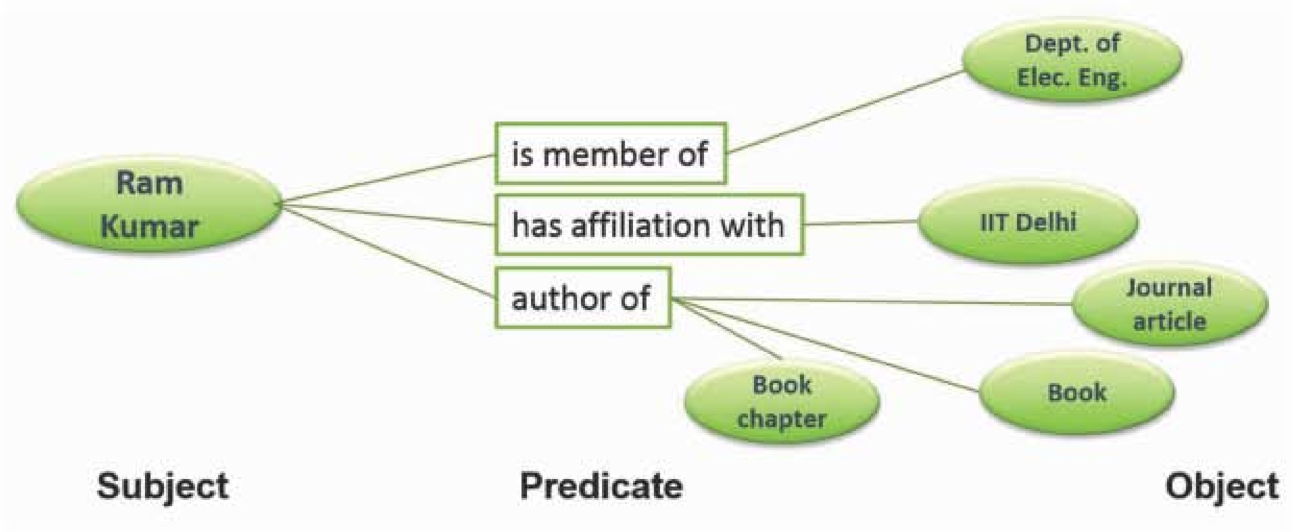
The rapidly growing scholarly community are connected electronically to their peers, colleagues, academic administrators, funding agencies and policy makers. The Internet has provided ample opportunity to this generation of scholars to communicate, share and discover information at push of a button. Evaluation, assessment and its impact play a vital role in academic institutions, R&D organisations and researchers has paramount importance. Sophisticated software system and database are needed to collect, process and analyse the huge amount of complex data produced by the scholars all over the world. The Profile Management system is a web-based tool to organise research activities carried out by a researcher such as interests, skills, experience, expertise, mentor, projects, publications, etc. The profile management system act as knowledge management as well as discovery tool for the researchers, faculty, scientists, etc. within the institutions as well as across the institutions. Development of profile management system compatible with international standard is vital for any organisation to showcase the research activities to the peer group or the funding agencies. There are number of proprietary and open source profile management system available with different features and functionalities. The article emphasis about important components of semantic technologies and case study of semantic-based profile management called VIVO.

2. Semantic Technologies

The content available on internet is very scattered in nature, to consume this information from various sources, semantic technologies can be used, Semantic technologies are to promote standards and exchange protocol to represent information on internet in meaningful way. Semantic technologies are extension of web standard.

2.1 RDF and OWL

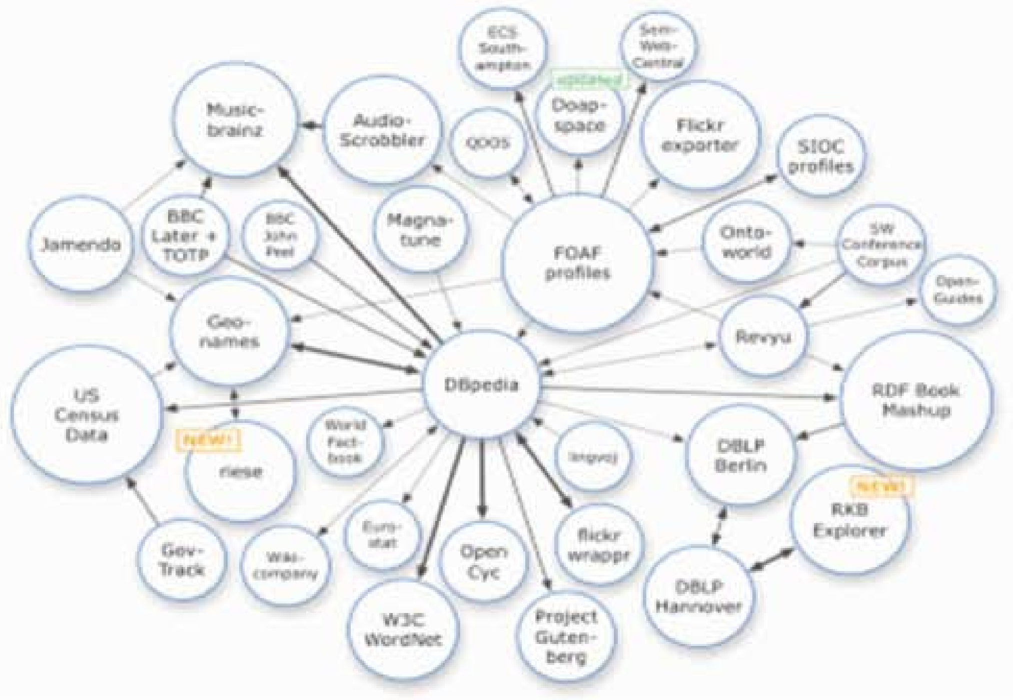
Resource Description Framework / Resource Description Framework Schema is a data module used to store data on the semantic web. The components of RDF are subject, property and object. Subject or resource are the things, which talks about author, publishers, person, organisaiton, etc, and can be represented by Uniform Resource Identifier (URI). The properties describe the relationship between subjects and objects i.e. written by, colour, name, worked in, etc. The object may be the literal value or subject of another triple store e.g. Kumar, red. For self-description of the data in RDF the Web Ontology Language (OWL) is being used. It has been designed in such a way so that data could be processed by computer and human in an efficient way.



Data Representation in the RDF Graph

2.2 Linked Data

World Wide Web relies on HTML files linked by the hyperlink and HTTP protocol to process the data, whereas Semantic Web relies on the Linked Data or Open Linked Data, which consist of RDF data model called triple statement with self-described URI. Linked Data refers to data published on the Web in such a way that it is machine-readable as well as understandable by human, it is a self-described data linked to other URI in the same domain as well as outside the domain, it lead to open linked data and enable us to generate the linked data graph based on the semantic inter relationship (Bizer, Heath, & Berners-Lee, 2009). The semantic-based application store the datasets in RDF/XML format. Also the semantic-based application are capable to generate the linked data or open linked data, which could be interlinked with the same kind of resources inside or outside the organisation and reused beyond the World Wide Web.



Linking Open Data cloud Diagram

2.3 SPARQL Query Language

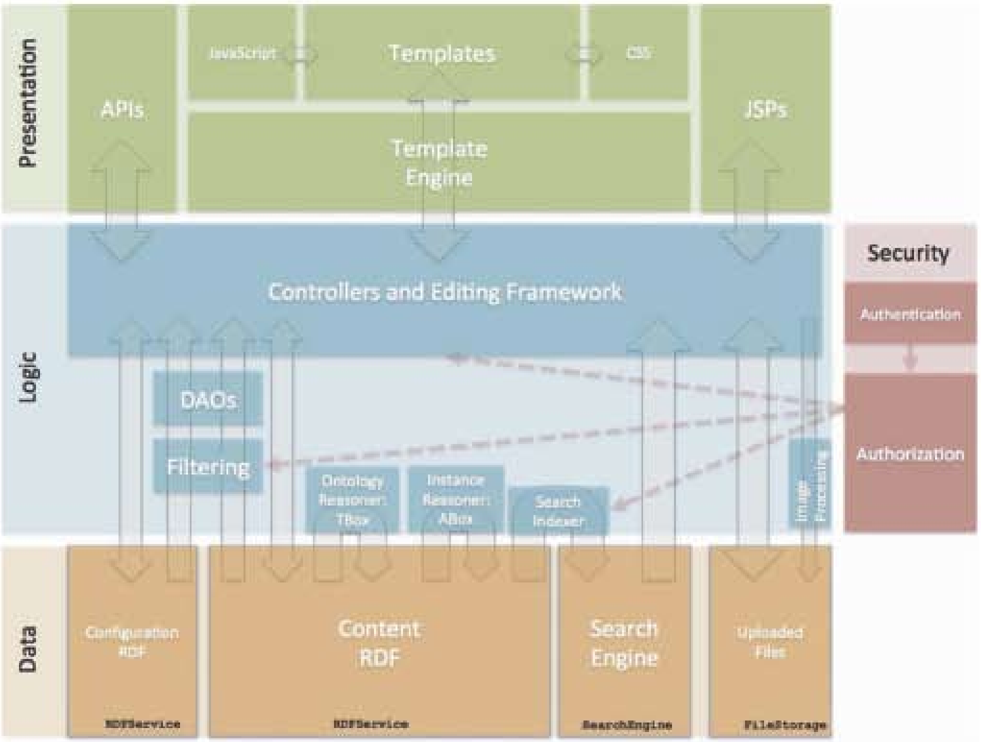
The Simple Protocol and RDF Query Language (SPARQL) is a W3 recommendation to retrieve the data from the RDF triples. The semantic-based application represent the data in RDF format such as N3, Turtle and deploy SPARQL endpoint to process the query from the client. SPARQL uses four form of query i.e. SELECT, ASK, CONSTRUCT and DESCRIBE and most of the semantic application support for the SELECT clause. A SELECT query allows you to identify which subset of the selected data is returned. The WHERE clause allows you to define graph pattern find the match in the data set. The semantic-based application has the capability to consume the RDF data from various source and produce the report, analysis, visualisations as desired by the client.

3. VIVO Overview and Architecture

Developed in 2004 by Cornell University Library, the VIVO is a semantic-based, open source, the community-maintained software tool for research discovery and networking of scholars. The VIVO software is a part of multi-million dollar project to promote inter-disciplinary collaboration in the field of Life Science (1997) and Social Science (2004). In 2009, the University of Florida along with Cornell University received a grant of about $12.2m from the National Center for Research Resources of the National Institutes of Health (NIH) to use VIVO as a framework to develop national level network of science community in US (Krafft etal., 2010). The VIVO enable the administrators to store, retrieve, analyse, visualise and showcase the research activities of the scholars in the effortless way using Linked Open Data.

The VIVO application consisting of ontology editor, data ingest tools and lightweight content management system. The ontology editor defines the type of information to be modelled, define the relationship between types and develop data property statement to describe individual, object property statement to connect individual. The VIVO data ingest interface enable import of information about the scholars such as personal, affiliation, education, accomplishment, projects, publication, etc. from different sources. The content management system provides the visual interface to display the scholars' information in a sophisticated way.

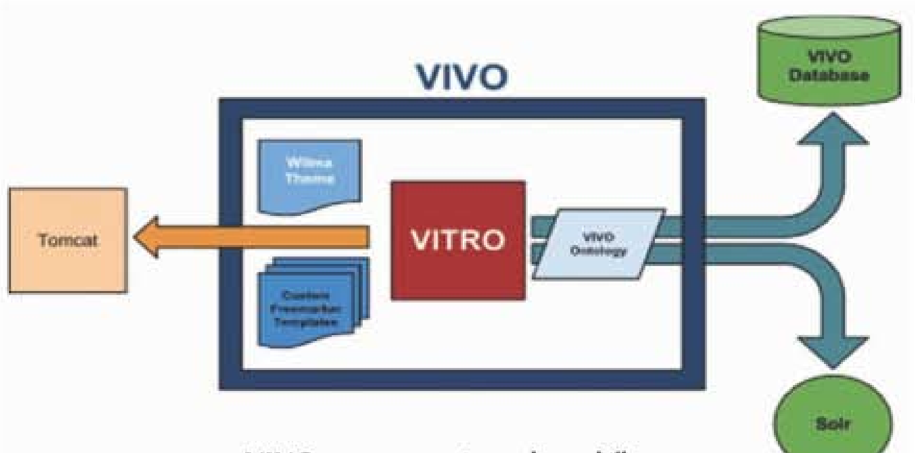
The software architecture consist of three layers including data, business logic and presentation. The data layer represent the information about individual person or institution or resources and their relationship. The logic layer consist of database connection, user privileges and authorisation. The presentation layer is the front end, consisting of templates, style sheets and JSP pages to display information to the user.



Software Architecture Overview

4. Open Source Components used for building VIVO Application

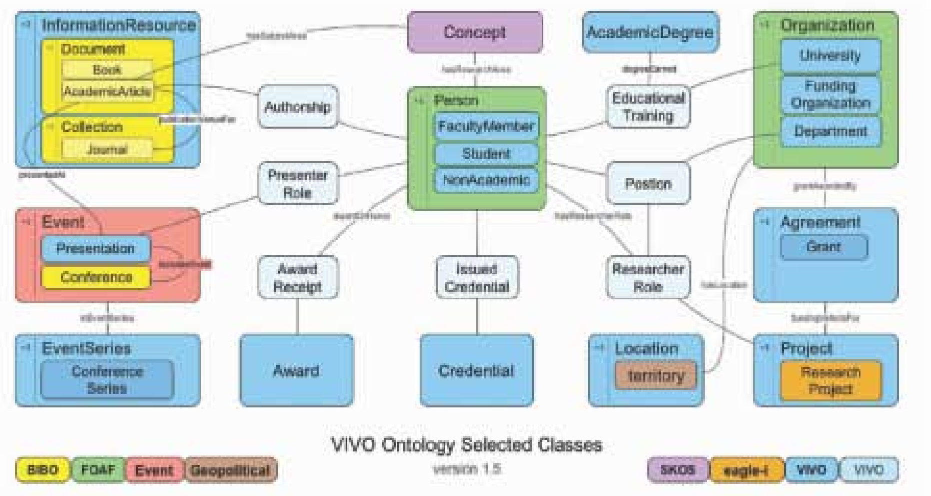
The VIVO application is constructed based on open source tools and libraries. The Jena Semantic Web framework is the container to build Java-based VIVO web application and deployed in Servlet Engine to execute the JSP files. The entire VIVO application, placed in the HTTP Server, in front of the Servlet container receives a request from the clients and process the data in triple store. The VIVO uses Jena Simple Relational Database or Trivial Database storage model to save the triple store in a relational database such as MySQL, PostgreSQL and also supports commercial relational database. VIVO uses Apache Solr search server to index the data and speed up the query result with the facet. The Free Marker Java Template Engine Library used to display the data in the desired manner. Apart from the above, JQuery and other JavaScript Libraries are used for asynchronous response and interaction.



VIVO components and workflow

5. VIVO Core Ontology

The Integrated Semantic Framework Ontology for VIVO called VIVO-ISF ontology is explicit and formal specification of information about organisations, researchers, research activities, events and their relationship (Krafft et al., 2010). It provides a set of types (classes) and relationships (properties) to represent the researcher's and their research activities. The VIVO-ISF or VIVO Core ontology is the extension of existing ontologies for various entity model such as Friend-of-Friend (FIFO), ontology provides the specification for people and organisation. Bibliographic (BIBO) Ontology provides the specification for research publications such as books, journal articles. Event ontology is the model to describe the conference, workshop, and Geopolitical (Geo) ontology is used to define the geographical location, and Simple Knowledge Organisation System (SKOS) used to describe the research interest and area of expertise. Apart from the above, VIVO uses various ontologies as per the organisational need and allow the developer to extend the existing ontologies for specific localisation such as expert ID, Google Scholar ID, Microsoft Academic ID, etc. Every entity such as person, article, department, event in VIVO is represented by the URI. Human readable web pages will be generated, when you access the URI from the browser and machine readable triple store or JSON file will be generated, when you access the same URI from the semantic application.



VIVO Ontology with Selected Classes

6. Tools Used to Process the Data

The Institute or organisation gather a large amount of data about the researchers, organisations, courses, events, grants, research articles, etc. from various sources in different formats such as MySQL, XML, CSV, etc. Feeding this data into VIVO by manual means is a difficult task and huge wastage of money and manpower. The VIVO team from participant universities have developed various tools to add, edit, update and ingest data from different sources and different formats. The VIVO provides built-in ingest tool to transform the CSV and XML file into RDF triple store. The SPARQL Constructer used to map the RDF triple store with the ontology, generate new triple store with proper meaning while importing in the VIVO system.

6.1 VIVO Harvester

The Clinical and Translational Research Informatics Program, University of Florida developed the java based library tool called VIVO Harvester (Barnes et al., 2012). The VIVO Harvester fetch data from external sources such as PubMed, SCOPUS, Web of Science, CSV file, OAI repositories, relational databases, translate into self-described RDF with the VIVO ontology, map with the VIVO data, and generate triple statement for VIVO store. The data fetching process uses XML as the intermediate and XSLT as translator to generate XML/RDF triples compatible with VIVO ontology.

6.2 GoogleRefine + VIVO

OpenRefine, formerly known as GoogleRefine, is an open source tool for cleaning the unstructured data. It enable the user to clean the unstructured data such as excel worksheet, comma delimited txt files, CSV files and convert into desired format such as XML, RDF,etc. (OpenRefine, 2015). The GoogleRefine has the capability to convert grid data into graph data and export it into the triple-store format that is building block for Semantic Web. The Weill Cornell Medical College developed the Reconciliation API interface for VIVO and the extension of GoogleRefine, called GoogleRefine+VIVO (Cole, Dickinson & Lee, 2013). The Reconciliation API, ingest data from VIVO, extract data from other reliable sources against the data from VIVO, reconcile datasets with additional data, align data to the VIVO Schema and export standardise RDF triples for VIVO.

6.3 Karma

The Karma is an open source information integration tool developed by Information Sciences Institute, University of Southern California, that enables users to quickly and easily integrate data from a variety of data sources including databases, spreadsheets, delimited text files, XML, JSON, and Web APIs (Karma, 2015). The user-friendly graphical interface has the capability to import the structured datasets from external sources, map data with the vocabulary in VIVO ontology and generate standard RDF. These RDF triples are directly imported into the VIVO application in an efficient way.

7. VIVO Features and Characteristics

The VIVO is a ready-to-use researcher profile management system, which provides extensive range of features, functionalities and also accustomed to international standard for reusability and interoperability. Some of the significant features of VIVO are integration with various bibliographical information providers (e.g. ORCID, SCOPUS, Web of Science, Google Scholar) for research information import, co-author network, co-investigator network, map of science network, QR code, OpenSocial gadget, etc.

7.1 Research Information Standards and Interoperability

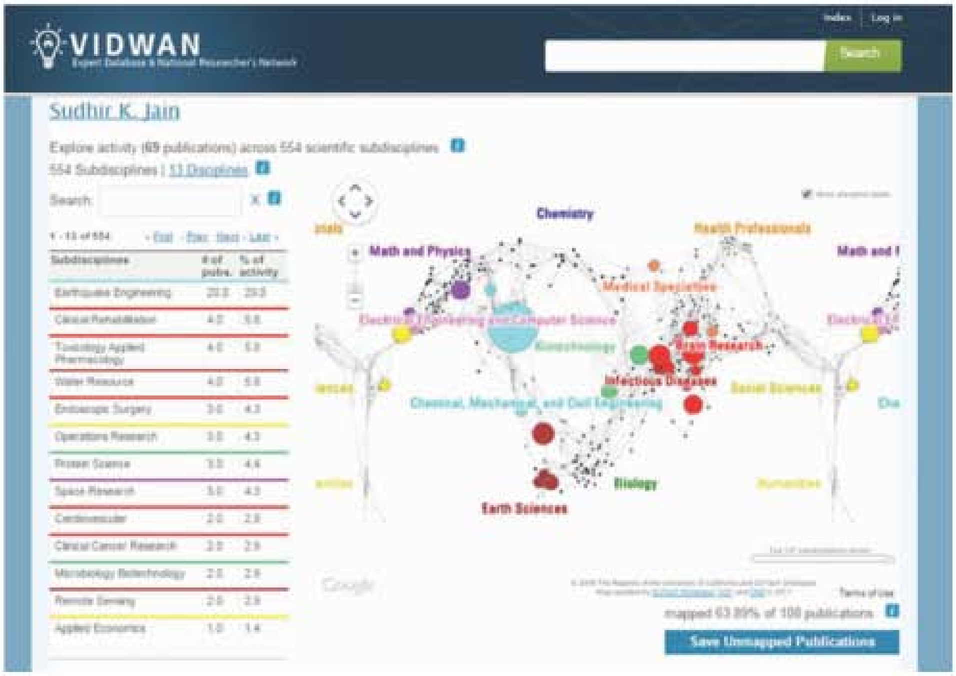
There is an increased need to develop structured CV for project proposals and to make effective communication between applicants and funders. The HR system maintaining the personal informational of the faculty or researcher may not have proper information about the research activities. There are hundreds of funding agencies requesting CV for funding opportunity in different form. These CVs should be standardised for effective communication. The Consortia Advancing Standards in Research Administration Information (CASRAI) is a not-for-profit standard development organisation and an international community of leading research funders and institutions collaborating to develop standard vocabulary for research information interoperability. European Current Research Information Systems (euroCRIS) is another not-for-profit organisation established to develop and promote Common European Research Information Format (CERIF) for European member state institutes. The VIVO collaborates with CASRAI and euroCRIS to advance the common approach to research interoperability and also enhance the implementation of VIVO around the World.

7.2 Visualisation and Reporting

The principle use case of the VIVO profile management is the discoverability of researchers based on their expertise, finding collaborators for research project based on the research output of organisation and create the national level network of faculty and scientists. Organisations and funding agencies are very keen and give emphasis on Return on Investment related to research activity. Academic administrator need the research report or research impact analysis in the graphical representation, which lead to greater understanding of the research progress and decision making for research funding. The VIVO provides temporal, geospatial, topical, and network analysis and visualization of data at the individual (micro), local (meso), and global (macro) levels (Tank etal., 2012) and also provide support for outside application to access the RDF data for various network analysis. A Sparkline is a small line chart that is typically drawn without axes or coordinates (Engelhardt, 2007). Total no of publications with year and co-authors of the faculty member or expert represents by Sparkline graph. Temporal visualisation allow the users to compare the publications and grants information of faculty or department in an organisation through visual graph as well as numerical value.

7.3 Map of Science

VIVO uses UCSD map of science to visualise the expertise or specialisation of the people, department, organisation, group of organisation. The science map allow the user to compare specialisation of people in the department, compare the department within the organisation, etc. The UCSD map of science and classification system developed by University of California San Diego (UCSD) comprising of article level data of the journals from Elsevier's Scopus and Thomson Reuters' Web of Science (WoS) for the years 2001-2010 (Börner et al., 2012). The 25,000 journals grouped into 13 broad discipline and regrouped into 544 sub discipline and these discipline represented by specific colour. Article not published in the 25,000 journals are represented by grey colour. The graphical representation of research activities helpful to the student to overview the specialisation of the organisation or department or faculty, which lead to find mentor or guide for the research project. Faculty and researcher could find potential collaborator and analyse the research output of respective project in efficient way. Funding agency could use the graphical representation as a tool to visualise the project progress, funding pattern and growth of the inter-disciplinary research also help them in decision making for project funding.



Sudhir K. Jain

Explore activity (69 publications) across 554 scientific subdisciplines

554 Subdisciplines | 13 Disciplines

Search:

1 - 13 of 554 « First ‹ Prev Next › Last »

Subdisciplines	# of pubs.	% of activity
Earthquake Engineering	20.0	29.0
Clinical Rehabilitation	4.0	5.8
Toxicology Applied Pharmacology	4.0	5.8
Water Resource	4.0	5.8
Endoscopic Surgery	3.0	4.3
Operations Research	3.0	4.3
Protein Science	3.0	4.4
Space Research	3.0	4.3
Cardiovascular	2.0	2.9
Clinical Cancer Research	2.0	2.9
Microbiology Biotechnology	2.0	2.9
Remote Sensing	2.0	2.9
Applied Economics	1.0	1.4

mapped 63.89% of 108 publications

Save Unmapped Publications

Area of Expertise Represented through Map of Science

8. VIVO Implementation at Institute and National Level

The data about researcher and research activity such as personal information, projects, publications, expertise, accomplishment, etc. scattered across the institute / organisation website or locked inside the institute database. These data are less utilised or underutilised due to lack of standard format for interoperability. The ultimate aim of VIVO is to explore the data into the outside world through open linked data. It enable the researchers and funding agencies to reuse for discovery and research data analysis. The W3C SWEO Linking Open Data community project assert that the aim of Linked Open Data is "to extend the Web with a data commons by publishing various open data sets as RDF on the Web and by setting RDF links between data items from different data sources" (The World Wide Web Consortium, 2015). The library play a vital role in any organisation to collect the research information, processing and disseminate to the academic community and it is a trusted entity between resources and faculty or researchers. Since the VIVO application itself first developed at Cornel University Library, the librarian or information scientist should play the lead role in the process of setting up infrastructure, content development and outreach activities to implement the VIVO at institute level. The VIVO is also used to interconnect the scholars working across the institutions through national level research network. The VIVO community project has developed various tools, applications and API for multi-institute harvest and search, which facilitates the end user to search the researcher profile of various institutions from a single interface.

9. Summary

There is an increased demand for ROI, impact factor analysis, ranking of institution to evaluation and assessment of research done by an organisation or institution. Semantic web and linked open data has been growing enormously, academic institution and R&D organisation have shown interest in joining linked open data and Govt. also promotes such activities that leads to national level open data. The VIVO application enable the organisation to showcase the research activity of faculty and researcher and will be the helpful tool for the administrator to analyse the research focus area, progress of research and funding opportunity. Since 2013, DuraSpace become the incubator for developing the VIVO software through open source community and promote the VIVO to the world. National level institutions such as Information and Library Network Centre, National Institute of Science Communication and Information Resources, Documentation Research and Training Centre should take the lead role to promote VIVO among the academic and R&D organisation in India.

10. References

1. Barnes, C., Williams, S., Sposato, V., Skaggs, N., Raum, N., Corson-Rikert, J., ... Blake, J. (2012). Extending VIVO. *Synthesis Lectures on Semantic Web: Theory and Technology*, 91.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. http://doi.org/10.4018/jswis.2009081901
3. Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., ... Boyack, K. W. (2012). Design and update of a classification system: the UCSD map of science. *PloS One*, *7*(7), e39464. http://doi.org/10.1371/journal.pone.0039464
4. Dr. Curtis L. Cole, Dan Dickinson, Kenneth Lee, E. C. (2013). Extending Google Refine for VIVO. Retrieved April 20, 2015, from https://wiki.duraspace.org/display/VIVO/Extending+Google+Refine+for+VIVO
5. Engelhardt, Y. (2007). Edward R. Tufte. Beautiful Evidence. Graphics Press LLC, Cheshire, Connecticut, 2006. *Information Design Journal*, *15*(2), 190–193. http://doi.org/http://dx.doi.org/10.1075/idj.15.2.13eng
6. Karma. (2015). Karma: A Data Integration Tool. Retrieved December 12, 2015, from http://usc-isi-i2.github.io/karma/
7. Krafft, D. B., Cappadona, N. A., Devare, B. M., Lowe, B. J., & Corson-rikert, J. (2010). VIVO/ : Enabling National Networking of Scientists. *Technology*, 24 months. Retrieved from http://journal.webscience.org/316/
8. OpenRefine. (2015). OpenRefine. Retrieved May 3, 2015, from http://openrefine.org/
9. Tank, C., Linnemeier, M., Kong, C. H., & Börner, K. (2012). Analyzing and Visualizing VIVO Data. *Synthesis Lectures on Semantic Web: Theory and Technology*, 141.
10. The World Wide Web Consortium. (2015). SweoIG/TaskForces/CommunityProjects/LinkingOpenData-W3C Wiki. Retrieved May 5, 2015, from http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/ LinkingOpenData