Pre-print
Paper presented at the 2nd Computer Science Online Conference, 2012

AgriVIVO: an Ontology-based Store of URIs and Relations between Entities in Agricultural Research

1. John Fereira, Cornell University Ithaca NY 14853-2801 US, jaf30@cornell.edu.

2. Valeria Pesce, Global Forum on Agricultural Research (GFAR) 00153 Rome Italy, valeria.pesce@fao.org.

3. Jon Corson-Rikert, Cornell University Ithaka NY 14853-2801 US, jc55@cornell.edu.

4. Ajit Maru, Global Forum on Agricultural Research (GFAR) 00153 Rome Italy, ajit.maru@fao.org.

5. Johannes Keizer, Food and Agricultural Organization of the United Nations (FAO) 00153 Rome Italy, johannes.keizer@fao.org.

Abstract

In order to facilitate better collaboration between agricultural research actors and ensure more effective management of research projects and more rational funding, it is desirable to have access to comprehensive information on people's expertise, areas of activities of Institutions, existing projects in specific areas and countries, related events and publications.

An information system that aimed at giving access to such information should: a) go beyond closed communities and directories (search several communities and directories, allow to share people profiles, affiliations, competencies, projects, publications across communities); b) go beyond serendipity, gathering information systematically, organizing data by discipline, affiliation, topic, geographic scope and providing context, in order to discover what is happening and who does what through meaningful relationships.

An existing project that aims at something similar is VIVO, started at Cornell University in 2003. VIVO is a research-focused discovery tool that enables collaboration among scientists across all disciplines at Cornell University. It allows to browse information on people, departments, courses, grants, and publications following an ontology-based navigation.

Cornell University, the Global Forum on Agricultural Research and the Food and Agriculture Organization of the United Nations are working on an adaptation of the VIVO model for agricultural research, called AgriVIVO.

AgriVIVO integrates data from several large agricultural research management communities into a VIVO RDF store, customizing the ontology model to the organization of agricultural research, focusing on the relationships between people, institutions, projects, topics, events, and geographic locations.

AgriVIVO works as a common store of URIs and relations to interlink the data managed in the existing communities and databases. Data management can remain decentralized as well as data browsing, as VIVO's search functionalities can be integrated in other websites.

Keywords

distributed system, entity disambiguation, information systems, ontologies, rdf, semantic web, solr indexes, user interfaces, web services

Introduction

The need for more collaboration between agricultural researchers, practitioners, and information managers across Institutional and geographical boundaries has been highlighted in many conferences, publications and projects in recent years[1].

To ensure that collaboration is as effective as it can be, it is essential to identify the right collaborators across Institutional and geographical boundaries. Additionally, in emergencies, it is essential to be able to quickly identify the right people or the key Institution or the relevant project that can make the difference in providing the necessary knowledge.

Most of the contacts and common connections researchers seek exist implicitly within the many information sources available today, but the lack of coordination among sources makes it difficult to identify and leverage shared connections or make them visible and persistent (figure 1).

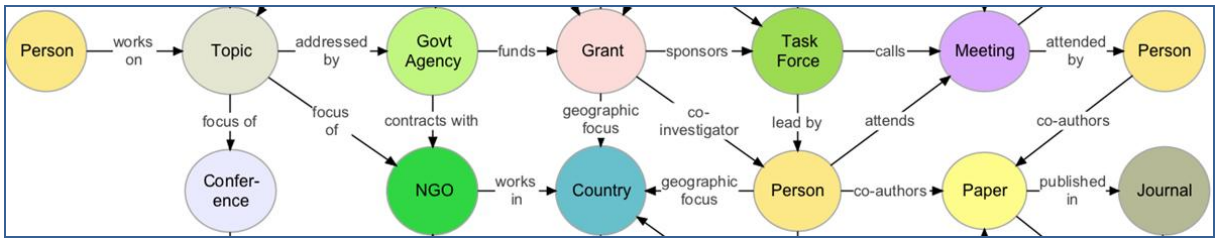

Fig. 1. Beyond serendipity: Building a network by representing the nature of relationships and using them to connect people, organizations, topics, activities, and outcomes

An information system aimed at facilitating the discovery of such connections should: a) go beyond closed communities and directories (search several communities and directories, allow to share people profiles, affiliations, competencies, projects, publications across communities); b) go beyond serendipity, gathering information systematically, organizing data by discipline, affiliation, topic, geographic scope and providing context, in order to discover what is happening and who does what through meaningful relationships.

An existing project that aims meets these criteria is VIVO, started at Cornell University in 2003. VIVO is a research-focused discovery tool that enables collaboration among scientists

across all disciplines at Cornell University. VIVO is based on an entity-relationship ontology model to organize and present information on people, research, and education activities[2].

Cornell University, the Global Forum on Agricultural Research and the Food and Agriculture Organization of the United Nations are working on a customization of the VIVO model for agricultural research, called AgriVIVO.

The AgriVIVO pilot will demonstrate a strategy for bridging across separately hosted directories, websites, and online communities without requiring centralized control. What AgriVIVO will endeavour to provide is the ability to search multiple databases and directories and to share people profiles, affiliations, competencies, and publications across disparate agriculture related communities.

AgriVIVO will integrate key existing sources of information about agricultural researchers by discipline, organization, or topic, using a consistent ontology framework. It will build upon the essential features of VIVO: a) the ontology approach to model the organization of agricultural research and the relations between all its entities; b) the RDF framework to build and update the ontology and to store URIs for all the entities identified in the ontology and the relations between them, expressed through properties from the most appropriate RDF vocabularies and from specialized AgriVIVO vocabularies; c) data harvesting tools for parsing, mapping and importing data from systems that use different machine-readable mechanisms to expose their data; d) special algorithms for disambiguating entities imported from different sources, also exploiting what has already been done by other initiatives in this area; e) internal and external authority data, so that AgriVIVO can offer a reference environment for any other information system: alongside existing authority data for geographic information and for agricultural concepts, AgriVIVO will integrate and align internal and external authority data for institutions and for people.

Methods

The ontological model

The original VIVO project developed at Cornell was based on an ontology that represented the organization of research at a University. The VIVO ontology tries to model types and relationships as they are in the real world (“ontology realism”: see figure 2).

VIVO uses "an entity-relationship ontology model to organize and present information on people, research, and education activities.”[2]. The VIVO core ontology is based on the following main classes: Person, using properties from the FOAF[1] namespace and from the

[1] http://www.foaf-project.org/

VIVO core namespace; Information Resource, using VIVO and BIBO[2]; Organization, using FOAF and VIVO; Area, using the FAO Geopolitical Ontology[3]; Event (Event Ontology[4]).

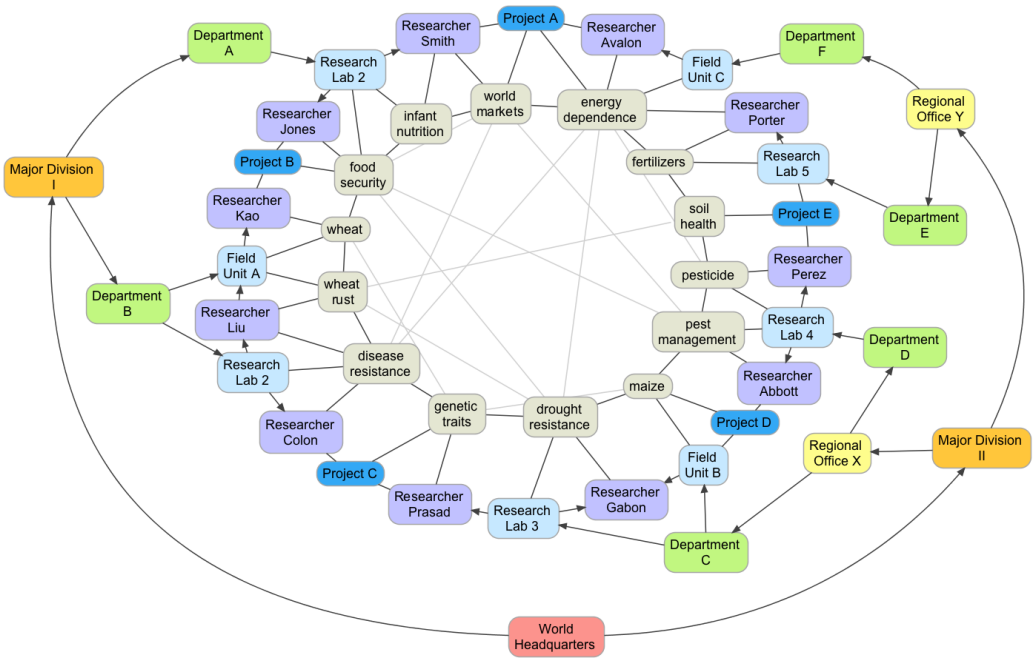

Fig. 2. An overview of the VIVO ontology.

Ontology extensions can be implemented in any VIVO installation to add greater specificity, but the core VIVO reasoner will always be able to infer more general type statements and thanks to the sharing of the core ontology consistent queries can be run across Institutions, even if some Institutions use local ontology extensions.

The VIVO concept and RDF model can be adapted to fit the way agricultural research works and implement a "research discovery tool" in agriculture by extending the core ontology and “hiding” classes and properties that relate to entities that don’t belong to the “real world” of agricultural research.

The AgriVIVO model shares with the VIVO model the basic entities of expert profiles, Organizations, projects, publications, events, geographic locations and all reciprocal relations. Most of the work on the ontology extension consists in the adjustment of sub-classes for these entities: in the VIVO model, specific “types” of the above entities are sub-classes, so for

[2] http://bibliontology.com/

[3] http://www.fao.org/countryprofiles/geoinfo/geopolitical/

[4] http://motools.sourceforge.net/event/event.html

instance sub-classes have to be created under the Organization class to include non-academic organization types.

Another essential aspect of the extension of the model is the integration of the AGROVOC[5] thesaurus to use agricultural concepts to "annotate" any entity in the model: this is done through the use of the "terminology annotation" object available in VIVO, which has a "referenced term" property that can point to any external concept if available at a URI and expressed in SKOS[6]. In addition to referring to AGROVOC terms through their URI, the VIVO extensible external concept services has been used to call the AGROVOC web services and integrate a free-text search on the AGROVOC thesaurus to facilitate indexing within the system.

In the future, in line with the objective of facilitating the identification of available competences all over the world, the AgriVIVO model could broaden to include (part of) the eCOTOOL[7] model that defines competence concepts and structures, or frameworks, within particular occupational domains.

Components and technologies

AgriVIVO uses the same technology as VIVO, which is implemented as a Java web application, currently running in a Tomcat servlet container. VIVO supports OWL ontologies through a Jena SDB triple store. An Apache Solr instance facilitates searches on the triple store. Additional APIs and custom harvesting tools are being developed for the AgriVIVO customization in order to implement a distributed architecture where data can be harvested by other data providers, made available as triples, exposed using a SPARQL endpoint, and made easily accessible through use of a Linked Data API[8] implementation to allow re-use by other data consumers.

VIVO also offers an ontology and data management environment: an ontology editing environment, an ontology import functionality, a data curation environment (updating information through a simple user interface automatically modifies the underlying triples) and data ingest tools for importing data.

URIs, disambiguation, authority data

Multiple URIs are a fact of life. In VIVO, each entity has a URI, but whenever other identifiers can be found for the same entity they are stored as well: owl:sameAs statements are asserted while keeping source data distinct. Additionally, sameAs statements can be removed,

[5] AGROVOC is a thesaurus of concepts in the agricultural domain, now published also as Linked Data: see http://aims.fao.org/standards/agrovoc/

[6] Simple Knowledge Organization System vocabulary: see http://www.w3.org/2004/02/skos/

[7] http://www.competencetools.eu/

[8] http://code.google.com/p/linked-data-api/wiki/Specification

both through manual data curation or automatic procedures, if better information becomes available.

AgriVIVO aims at also becoming a "hub" of authority data in agriculture, referencing existing authority data whenever possible (AGROVOC concepts for topics, the FAO Geopolitical Ontology for geographic locations, existing authoritative bibliographic databases and catalogues for publications, existing directories for institutions) or creating new authority data, always mapping alternative authority data for the same entity through owl:sameAs statements.

Results

The current prototype is an AgriVIVO server instance that integrates information on experts, related competencies and belonging Institutions from a few key sources: the e-agriculture community[9], the CIARD[10] and RING[11] user base, the AIMS[12] community and the EGFAR[13] contact list. It also includes events in agricultural harvested from the AgriFeeds[14] portal.

The final AgriVIVO server instance will allow to submit/harvest structured information on expert profiles, belonging Institutions, related projects, related publications and events in one RDF store that can be used by any information system as reference data from which information can be pulled and to which information can be pushed. This will allow different websites and information systems to share the same data and also to map their local data to a unique "authority" integrated research management information system.

AgriVIVO will not replace any existing community or database; it will work as a common registry of URIs to interlink the data managed in the existing communities and databases. Communities and databases will share data through AgriVIVO.

Beside the server instance, several applications are planned that will re-use the data from the server instance in order to provided added-value services: a search engine over this integrated AgriVIVO and other designated sources (e.g., USDA VIVO or VIVO at land grant universities in the U.S.) patterned on http://vivosearch.org; search functionalities in any website: VIVO's search functionalities can be integrated in other websites through remote

[9] http://www.e-agriculture.org

[10] Coherence for Information for Agricultural Research for Development (CIARD): http://www.ciard.net

[11] Routemap to Information Nodes and Gateways (RING): http://ring.ciard.net

[12] Agricultural Information Management Standards community: http://aims.fao.org

[13] The web space of the Global Forum on Agricultural Research (GFAR): http://www.egfar.org

[14] The agricultural news and events aggregator: http://www.agrifeeds.org

calls. In this way, specialized and targeted search engines can give access to and offer highly customized "views" of the data coming from AgriVIVO.

Discussions

Authority data, disambiguation and interactive data curation

Current research activities in the VIVO project include the management of overlapping or duplicate data from multiple sources through "sameAs" assertions and the ability to assert that an author is not the author of a publication when ambiguous author names may risk repeat assertions of incorrect data. These VIVO functionalities will contribute to the international availability of authoritative data for publication authors, in collaboration with the Open Researcher and Contributor ID (ORCID) initiative and other efforts to uniquely identify organizations, people, and projects.

AgriVIVO can also be used as a community platform for interactive data curation. Users can add/remove "relations" in which they are part of the relation (person A "is author of" publication B, person A "participates in" project C): in this way they can "claim / disclaim" publications or associate / remove themselves with / from a project.

Users may also assert that they are attending an Event, allowing other User to see who, within their area of expertice may also be planning to attend the same event.

Conclusion

We feel that AgriVIVO can be a powerful tool for collaboration between agricultural researchers, practitioners, and information managers across institutional and geographical boundaries and a discovery mechanism to find others outside existing closed communities with expertise in a specific agricultural domain. The use of an ontological model and semantic technologies suits this scenario perfectly, while the main challenge remain deduplication and the disambiguation of entities for which no authoritative source exists.

References

1. Global Forum on Agricultural Research. *The GCARD Road Map. Transforming Agricultural Research for Development (AR4D) Systems for Global Impact*. FAO, 2011. ISBN 978-92-5-106908-0

2. Devare, Medha et al. Connecting People, Creating a Virtual Life Sciences Community. D-Lib Magazine, July/August 2007, vol. 13 no. 7/8,