A Framework for Knowledge sharing and Interoperability in Agricultural Research for Development

Caterina Caracciolo and Johannes Keizer

Food and Agriculture Organization of the United Nations (FAO)
v.le Terme di Caracalla 1
00154 Rome
Italy
{Caterina.Caracciolo,Johannes.Keizer}@fao.org

Abstract. In an ideal world all data would be produced using open formats and would be linked directly to other related data on the web. This would give the possibility for service providers to set up information systems by mixing and matching data from different distributed repositories. A scenario like this is no science fiction. Nevertheless most data (of all kinds) resides in database and repository silos, and efforts to create one stop access to distributed data lack functionalities, robustness or sustainability. The CIARD initiative[1] is working to make agricultural research information publicly available and accessible to all, by acting on both those issues. Among its actions are advocating and promoting open access, improving applicability and enabling effective use of data and information in agricultural research and innovation. In this paper we present the CIARD initiative and concentrate on FAO's contribution to it. We present the Linked Data approach, the vocabulary editor VocBench, the domain specific tagger AgroTagger and the RING registry of services and tools.

Keywords: Interoperability, Information sharing, CIARD, FAO, Linked Data.

1 Information in agriculture: where are we?

Scholarly communications through articles and conferences with collateral exchange of limited datasets have been the means of data sharing in the past. However, scientific publication and data production is growing at a much faster rate than ever before. Figure 1 charts the number of articles indexed by MEDLINE from 1950 to 2010. The steep increase that has begun in 1995, the internet era, is clearly visible.

---

[1] http://www.ciard.net

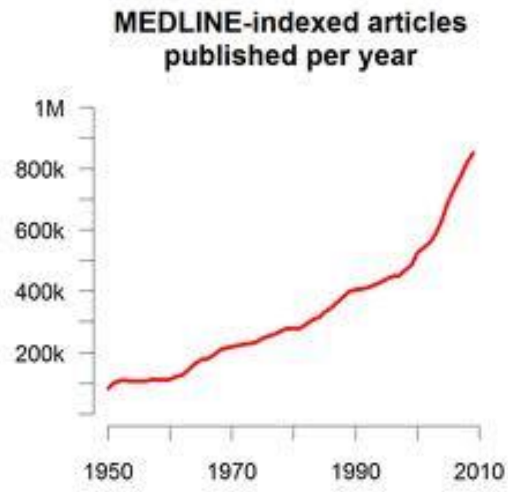

Fig. 1. MEDLINE-indexed articles published per year. Source: http://altmetrics.org/manifesto/

But the production of peer-reviewed scholarly journals is only the tip of the iceberg. Behind growing numbers of scholarly publications there is a growing amount of scientific data, such as experimental data, published as data sets with their associated metadata and quality indicators. Furthermore, scholarly papers are no longer the only way in which scientific information is exchanged. Document repositories of white papers or technical reports are widely used, and so are other web-based repositories, that may be considered as Knowledge 'derivatives' such as collections of descriptions of agricultural technologies, learning object repositories, expertise databases, etc. Researchers are also using more social platforms such as blogs[2] to discuss results before they are published in scholarly journals or after they have been published. Personal communications are now often made more general and turned into community communications, thanks to blogs and professional platforms that may have access to directories of peoples and institutions. This growth of scientific output calls for a growth of the instruments at disposal of scientists to orientate themselves into this wealth of information.

[2] See as an example ScienceBlogs http://scienceblogs.com/

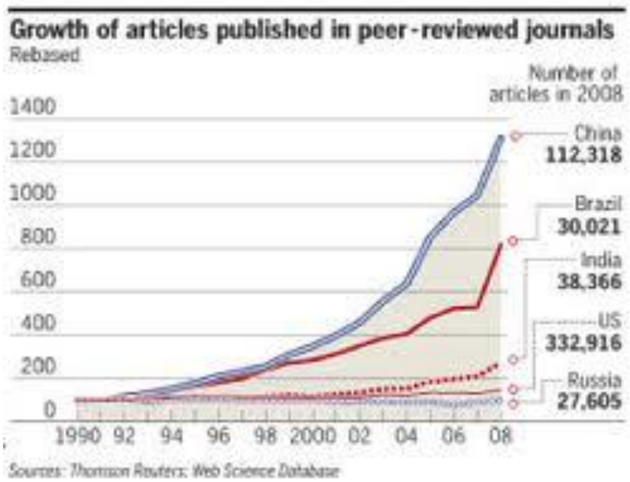

Fig. 2. Growth of articles pubished since 1990. See the outstanding growth of China, Brazil, India. Source Thomson Reuters. Web of Science Database.

Another factor to take into account is the expansion of the scientific world to include a few larger countries with so-called "emerging economies", especially China, Brazil and India, into the scientific mainstream. This phenomenon is especially clear since the year 2000, as it is shown in Figure 2. For example, taking 1990 as a base, Brazil has increased its scientific production by 800% and China by 1200%. The growth of these key new players in the scholarly communication arena has made interoperability a more important global issue, with special regard to the handling of languages other than English.

Data on publication rates in agriculture are not easily available, but it is quite clear that the trend will be similar, although perhaps less dramatic, to the other life sciences which are monitored by MEDLINE[3]. Agricultural scientists should be responsible for actively ensuring that the new knowledge generated from their research is easily accessible and so easily taken up by their colleagues and the agricultural community. This issue is also vital in the face of rapidly shifting challenges such as climate change, trans-boundary pests and diseases, and agricultural trade tariffs.

Some of the principal partners in the CIARD initiative have organized an e-consultation with the goal of identifying how to enhance the sharing and interoperability of information for global agricultural research for development. This event was followed by an expert consultation in Beijing in June 2011 with the same title. These two events are exercises to describe the current status and analyze the needs for tools, standards and infrastructures, leading on to defining future actions.

[3] http://www.nlm.nih.gov/

2 Data sharing and interoperability: where are we?

The IEEE definition for interoperability refers to the ability of two systems or components to exchange information and use the information that has been exchanged. In practice, the actions that require some degrees of interoperability include the transferring of data from one repository to another, the harmonization and harmonization of different data and metadata sets, the creation of a virtual research environments, the creation of documents from distributed data sets, the reasoning over distributed datasets, and the creation of new information services using distributed data sets. This list, although not inclusive, shows that the interoperability of systems implies “interoperability” of data. Nowadays, this implication is widely acknowledged and we can claim that we are in a data-centered world.

Data sets (consider for example those mentioned in the previous section) are often connected in a way or the other, as in the case of two scientific papers about the same experiment, or the scientific data and the published paper based on that data. However, much of that connection is often implicit. One of the reasons of such a limited availability of connections between data sets is that each data set usually has in independent life cycle, with big variety in the way data is structured and accessed. This is a well-known problem among researchers and practitioners in the area of relational databases.

There are several interesting examples of successful data exchange between distributed datasets, and some of them in the area of agricultural research and innovation. A common characteristic of most examples is that they are based on specific ad-hoc solutions more than on a general principal or architecture. Usually, they rely on pre-existing agreement between the people or organizations involved in the exchange, which lead to the sharing of data through web services, or accessible APIs.

Some well known initiatives try to achieve data sharing for scholarly publications, mostly organized by the publishers[4]. Other initiatives concentrate on document repositories and open access. Certainly the best known of the initiatives dealing with document repositories, is the Open Archives Initiative[5] (OAI), which promotes interoperability standards to facilitate the efficient dissemination and sharing of content.

Initiatives related to data sharing are also happening in specific domains. The Open Geospatial Consortium[6] (OGC) represents a strong community in the area of geospatial information, which promotes geospatial and location standards. It has spurred important open source projects such as GeoServer[7]. Considerable cooperation among international organizations has focused around the problem of sharing statistical data, such as census data and time-series, leading to initiatives such as the Statistical Data and Metadata Exchange[8] (SDMX) and the Data Documentation Initiative[9] (DDI), an

[4] For example, Science Direct by Elsevier http://www.sciencedirect.com/
[5] http://www.openarchives.org/
[6] http://www.opengeospatial.org/
[7] http://geoserver.org/display/GEOS/Welcome
[8] http://sdmx.org/
[9] http://www.ddialliance.org

XML based standard for statistical data exchange, and a specification for capturing metadata about social science data, respectively. These standards are embraced by various international organizations, including the World Bank[10], the International Monetary Fund[11], United Nations Statistical Division[12], and the Organization for Economic Co-operation and Development[13]. Similar experience is the Gene Ontology Consortium[14], which promotes the standardization of the representation of gene and gene product attributes across species and databases. Specific initiatives are Singer System[15] and GeoNetwork[16] in the area of genetic resources and georeferenced maps respectively.

Apart from the virtuous examples represented by the initiatives just mentioned, the general problem regarding the actual way information systems developers access, collect and mash up data from distributed sources is still to be solved. Mainly, data sources are as disconnected and unorganized as before. The problem then is how to connect diverse yet intrinsically connected data sets, and organize efficient collaboration processes.

3 Interoperable is not centralized. Or, the Linked Data approach

One approach to data sharing consists in making all interested parties use the same platform. In other words, through centralization. Quite naturally, in this way interesting degrees of interoperability. *Facebook* and *Google* are the largest examples of centralized systems that allow easy sharing of data among users, and some interoperability of services. Uniform working environments (software & database schemas) help create interoperability despite physically distributed information repositories. But there are social, political and practical reasons why centralization of repositories or unification of software and working tools will not happen. This does not mean that a certain concentration of data in specific servers on the web or a common set of software tools is useless. It just means that this cannot be the unique road to allow sharing and interoperability.

The alternative to centralization of data or unification of working environments is the development of **standard ways to encode, transmit and process data**, to make distributed data sets interoperable and shareable. This is exactly the idea of the "semantic web", that of a web of global interoperability, launched more than 10 years ago. Since then, a number of standards promoted by the W3C have contributed to the achievement of that goal. The underlying idea of these standards is that interoperabil-

[10] http://www.worldbank.org/
[11] http://www.imf.org/external/index.htm
[12] http://unstats.un.org/unsd/default.htm
[13] http://www.oecd.org/
[14] GeneOntology Consortium http://www.geneontology.org/
[15] Singer System http://singer.cgiar.org/
[16] GeoNetwork http://www.fao.org/geonetwork/srv/en/main.home

ity is achieved when machines understand the meaning of data and are able to programmatically process them.

The latest paradigm for interoperability on the web is the data publication style called **Linked Data**[17] [1] [2]. Instead of pursuing ad hoc solutions for the exchange of specific data sets, the concept of linked data establishes the possibility to express structured data in a way that it can be linked to other data sets that are following the same principle. This style of publishing data is gaining much attention and new services such as the New York Times and the BBC are already publishing data that way. Some governments too, are pressing heavily to publish administrative information as Linked Data.

Linked Data focuses on data and its links to other pieces of data, independently of their physical location around the world. Linked Data is then about using the web to link together what was not previously linked, and exploit the potentiality of web publication. It is the equivalent for data of what an HTML document (the hypertext of the early days of the internet) is for text. But, unlike HTML documents, linked data may be used programmatically by machines, and so exploited within web applications.

More than a specific technology, Linked Data is a web publication style based on the use of a few web standards and protocols. Primarily, Linked Data is based on the Resource Description Framework[18] (RDF), which facilitates the exchange of structured information regardless of the specific structure in which they are expressed at source. Any database can be expressed using the RDF, but also structured textual information from content management systems can be expressed in RDF. Once data is expressed as RDF and exposed through a web server with a SPARQL[19] end-point, the web client may be presented with either a human-readable (HTML) [3], or a machine readable version of the data, depending on the request. The RDF version of the data makes it understandable and processable by machines, which are able to mash-up data from different sites. The advantage of data mesh-up based on Linked Data is that data is exposed with uniform interfaces. On the contrary, in common data mesh-up techniques, ad-hoc procedures are required for each data source.

There are now mainstream open source data management tools like Drupal[20] or Fedora commons[21] which includes already RDF as the way to present data. The other important standards for Linked Data is the URIs to name and locate entities in the web, and the HTTP protocol to exchange web content through web browsers.

Once these three mechanisms (URIs to locate and name data objects, RDF to encode them, and HTTP to move them around in the web) are implemented, data can be linked to any other relevant data set published using the same technologies and protocols. And the whole thing, the resulting linked data, may be exploited in applications.

[17] Linked Data - Connect Distributed Data across the Web http://linkeddata.org/

[18] Resource Description Framework http://www.w3.org/RDF/

[19] http://www.w3.org/TR/rdf-sparql-query/

[20] http://drupal.org/

[21] http://fedora-commons.org/

Let us suppose that a data owner publishes a data set about sardine captures as linked data. Then it could link its data to the a third party data set, say a reliable source of information about biological profiles of fish species. The resulting linked data may be of interest to humans, but also programmatically used in applications exploring the connection between fish capture and the climate.

When Linked Data is published with an explicit open license, or when it is at least openly accessible from a network point of view (not behind an authorization check or paywall), then it is Linked Open Data. The Linked Open Data project[22] collects information about data sets published that way.

FAO, as well as many other large data producers (BBC, the Library of Congress, the US government, to name just a few), has started publication of its data as Linked Open Data. The very first data set published by FAO is its multilingual thesaurus AGROVOC[23] [3], available in over 20 languages and covering all subject areas related to FAO interest. To date, AGROVOC is linked to ten relevant thesauri and vocabularies in specialized domains, and a few more are in the pipeline.[24] **Fig. 3** shows an intuitive view of the fact that AGROVOC is linked to other vocabularies, by representing three concepts belonging to three vocabularies (NALT[25], Eurovoc[26], and GEMET[27]). That figure also captures the fact that once various vocabularies are linked, one also gains the access to the document repositories associated to them. With the publication of AGROVOC as linked data, not only has FAO exposed its first data set in the linked data world, but the largest data set about agriculture is now out there for public use.

[22] http://richard.cyganiak.de/2007/10/lod/
[23] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[24] For the entire list of vocabularies connected to AGROVOC; and a human-readable version of the connected data sets see: http://aims.fao.org/standards/agrovoc/linked-open-data
[25] http://agclass.nal.usda.gov/
[26] http://eurovoc.europa.eu/
[27] http://www.eionet.europa.eu/gemet

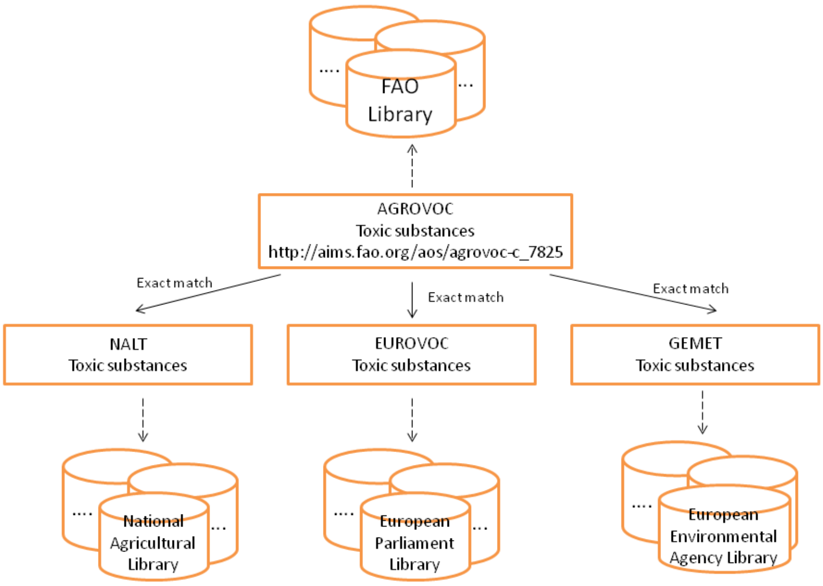

Fig. 3. AGROVOC concepts are linked to other vocabularies (solid lines). Vocabularies are then connected to one or more document repositories (dashed lines) because they index the documents contained in them.

We envisage that this is the way to go to achieve interoperability of data in the agricultural domain. In our view, the four key elements to make this happen are:

1. common standard vocabularies to allow for the use of RDF in specific domains and facilitate automatic data linking (thesauri, authority files, value vocabularies)
2. appropriate tools to store, manage and exploit data in RDF, including content management systems, RDF wrappers for legacy systems, and RDF based information extractors
3. a common way for users to know what vocabularies, data sets and tools are available
4. people able to understand these mechanisms, to produce linked data and exploit it in relevant applications. This implies that training and capacity development activities are needed.

This is, in a nutshell, the perspective endorsed by CIARD, which we explain in detail in the next section.

4 The CIARD perspective

The Coherence in Information for Agricultural Research for Development (CIARD) [28] was established in 2008 with the aim of making research information and knowledge accessible to all [4]. Through consultations, virtual and face-to-face, with a wide range of organizations it was agreed that data and information sharing needs to be enhanced and improved starting from within organizations at national, regional and global level. In 2011, two such international consultations took place, to analyze and understand the state of the art in outputs of agricultural research. The result was a plan of actions, to implement the idea that data should be available and shareable avoiding central repositories or standardized software, but rather enabling horizontal communication. Interoperability is then a notion interpreted at the data level, as interoperability of distributed data sets. In this sense, the CIARD perspectives couples very well with the Linked Data perspective introduced previously.

Data should be made available using formats suitable for the web and for common consumption, such as RDF, in combination with standard vocabularies. In case of data born in other formats, it will have to be made available in the web format by means of suitable software, then stored. Tools, software and services will have to be available to institutions and users in a common and accessible place. Also, data storage will have to be possible. Finally, it is important that awareness and capacities are well developed in the community of users.

In this section we point out at these components individually, and try to provide the reader with a concise, high level summary of the CIARD view. For a more detailed account of that, the interested reader should consult the CIARD website[29]. Note that there is no a priority ranking in our listing, but rather an attempt to provide a narrative line to our view.

1. **Tools to convert existing data into formats suitable to sharing.** Most data repositories will keep their current format and structure for a long time. In principle, this is not a problem, also because many of these systems have been created to optimize data processing. However, this implies that data should be converted in order to be sharable over the web and accessed by other parties in a programmatic way. Various technologies to convert data born in other formats into RDF are available. Also, Linked Data is a publication style that can already be easily adopted. However, such processes of conversions are usually performed according to ad-hoc procedures. Our goal is to streamline these processes, and make available to the community of information producers in agriculture suitable methodologies to address these problems. In this way we can produce "triple stores" of RDF data that interlink distributed information sources and make them accessible from aggregations sides or specialized services.

[28] http://ciard.net

[29] Core documents can be accessed at: http://www.ciard.net/ciard-documents, while the entire repository is available at: http://www.ciard.net/repository.

2. **Vocabularies.** The notion of standard vocabulary is needed in all information tagging activities, including the traditional activity of document indexing. In fact, that is the reason why vocabularies like AGROVOC were originally developed. Beyond AGROVOC, a number of vocabularies, thesauri and specialized glossaries are available in the agricultural domain. They need not only to be mapped and linked, but also to be made available for the purpose of data markup. Tools such as Open Calais[30] and AgroTagger (see next section) are examples of software for data markup: based on specific vocabularies, they use entries in the vocabularies to mark up documents. We would also mention another important application of standard vocabularies in information sharing, since data cannot be interoperable without an explicit qualification of its intended meaning. Specifically, being RDF essentially a data model based on the triple structure of "subject-predicate-complement", we need to be clear about what the intended meaning of the predicates used. Lists of predicates are also called vocabularies, and are essential to information sharing.
3. **Data Storage on the cloud.** Data for the web, be that pure RDF, Linked Data or other formats, will have to be stored in a way that is accessible to the world. This requires some infrastructure that may be costly for many institutions involved into data production. We see here a need that could be addressed by CIARD, by providing an infrastructure available to the community for the storage of their data. This will be a sort of "cloud storage" of data.
4. **Tools for data processing and storing.** A selection of the tools currently available for the publishing of interoperable data should be made, according to the requirements and possibilities of the actors represented by CIARD. What we need is then a suite of tools and services to process and store data, that can be installed and used at institutional levels.
5. **Interfaces to distributed triple stores.** Once data is made available for consumption, it is important that interfaces are also available to make it exploited by applications. The CIARD partners should produce a library of such interfaces with inbuilt data selectors.
6. **Registry for data and software.** We need a one stop access to existing services from the community. The registry of tools and services developed within CIARD, called RING[31], should become a hub of data streams, from which the community can channel their data.
7. **Awareness and Capacity.** Last but not least, a clear commitment to open access for data and publications from all partners is needed. Managers, researchers and information managers must understand that sharing will enhance our capacity to create knowledge for agricultural development. They also must understand the instruments that have been created by international standard bodies to achieve this.

[30] http://www.opencalais.com/
[31] http://ring.ciard.net/

The **Fig. 4** provides an intuitive view of the data flow from individual, local repositories into web based applications, through web services that consume linked data.

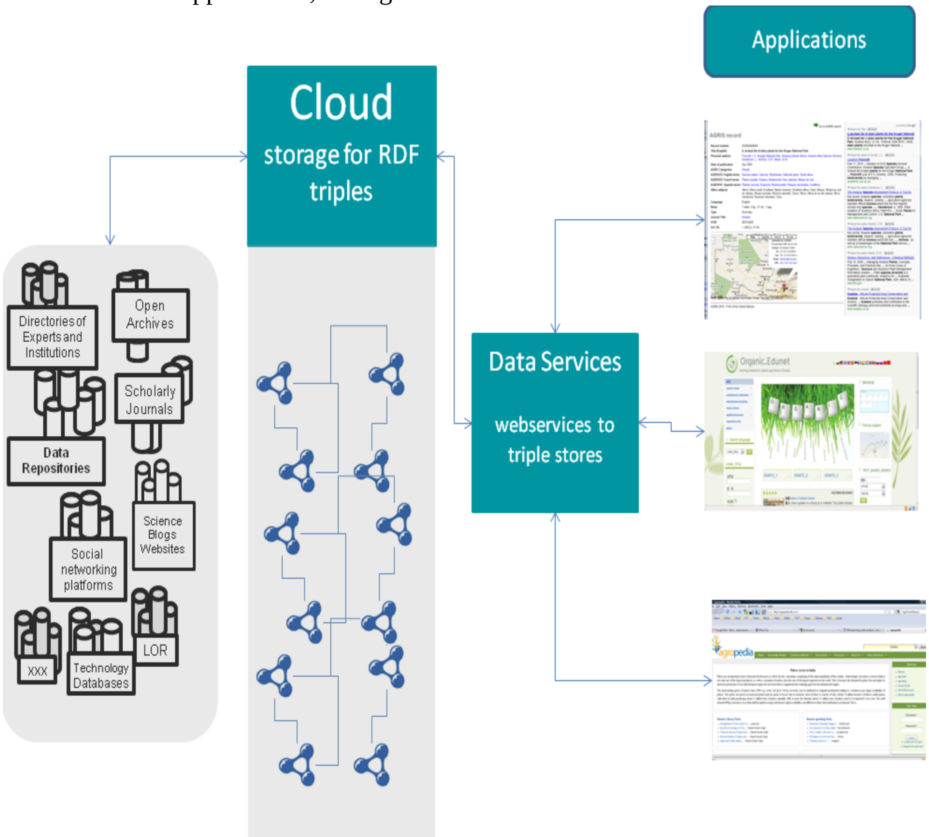

Fig. 4. Data flow from local repositories to data accessible through the web.

This is a program that can be implemented gradually. FAO is already working towards the implementation of some of these points, with activities supported by both its own budget and also external funding. For example, recently the EU has funded the agINFRA project[32], that deals with the implementation of an infrastructure for the agricultural domain. This fact proves that the CIARD view may be presented in a convincing way to donors and interested stakeholders. In the following section we go on presenting in details FAO´s contribution to the CIARD view.

[32] http://aginfra.eu/

5 How FAO is contributing to the CIARD view

FAO is contributing to two of the main components of the CIARD view. FAO is active in the area of “Vocabularies” (point 2 above), because it maintains the AGROVOC thesaurus. We have introduced AGROVOC in Section 3 above, together with the Linked Data project. AGROVOC is central to FAO’s effort within CIARD, because it triggered the development of VocBench, a tool for vocabulary editing, and because it is used in other applications such as AgroTagger, an agriculture-specific tagger, that we introduce below. Moreover, FAO is planning on turning AGRIS[33], originally a bibliographic repository, into a pilot service based on an infrastructure dedicated to interoperability.

VocBench[34] is a web-based tool originally developed to manage the AGROVOC thesaurus. Nowadays, many ontology editors have been developed over the years, including Protégé[35], the NeOn Toolkit[36], Altova Semantic Works[37], and TopBraid Composer[38]. But a few years back, very few editors were available and usually they did not support collaborative work, a formalized workflow with user roles and editing rights also by languages, or UTF-8 – and none all these features together. This is the reason why FAO started the development of its own web-based, fully multilingual vocabulary editor supporting collaboration structured into an explicit workflow. Moreover, in order to allow individuals and organizations to contribute to AGROVOC while maintaining the information about the provenance of their authorship, VocBench applies a fine-grained mechanism of track changes. These features have made the interest around VocBench grow, so that its community of users has grown beyond the one originally envisaged. VocBench is now used to maintain the FAO Biotechnology Glossary[39] and much of the bibliographic metadata used by FAO. Currently in version 1.3, VocBench supports the export of data into RDF-based format, and it will soon support RDF format natively.

Much of the available data was born before the Linked Data era, so it will need to be converted to formats suitable to that style of publication. This is certainly the case for both relational data and for unstructured textual documents. In particular, in the case of unstructured textual documents, information will have to be extracted from documents and then turned into sets of RDF triples. The information extraction process may involve the extraction of the author and title of the document at hand, or the events discussed in it, or specific topics such as “salmon” or “crops” and the like. For the case of agricultural related topics, IIT Kanpur[40], FAO and MIMOS Berhad[41] have developed Agrotagger, a web-based tool to tag documents according to keywords

33 http://agris.fao.org/
34 http://aims.fao.org/tools/vocbench-2
35 http://protege.stanford.edu/
36 http://neon-toolkit.org/wiki/Main_Page
37 http://www.altova.com/semanticworks.html
38 http://www.topquadrant.com/products/TB_Composer.html
39 http://www.fao.org/biotech/biotech-glossary/en/
40 http://www.iitk.ac.in/
41 http://www.mimos.my/

taken from AGROVOC. Currently in beta version, AgroTagger is available for trial through a web interface[42], where one may select a document and get in return the AGROVOC keywords suggested by AgroTagger, completed with their URIs. The set of suggested keywords can also be downloaded as an XML document. AgroTagger may also be accessed via web services so that one may be submit document in bulk, for use within other applications.

Finally, we want to mention the Routemap to Information Nodes and Gateways (RING)[43], a registry of tools, services and data sources developed by the Global Forum on Agricultural Research (GFAR)[44]. The idea behind the RING is to make available from a single access point the data and services relevant to agricultural information professionals. Currently implemented as a Drupal[45] website, it aims at offering extra services on top of the role of registry of data and tools. For example, it could not only provide links to RDF repositories, but also provide view/navigation functionalities to the data. Or it could include harvesters for all data and metadata providers registered in it. It could also become a repository of code to the benefit of programmers wishing to implement services embodying the philosophy shared by the CIARD.

6 Conclusions and future work

The AIMS team at FAO has concentrated its efforts for years on providing semantic standards to facilitate cooperation and data exchange. Now, thanks to the community grown around the CIARD initiative, a more comprehensive view about information sharing in agriculture is taking shape. All aspects involved are being analyzed, both the technical level, and the policy and capacity level. In this paper we concentrated on the technical level, and placed the main products of the AIMS team within the infrastructure envisaged by the CIARD view. In the coming years, we will continue working towards the completion of the CIARD view, and publishing in the form of Linked Data as much of the FAO data as possible. Our first step will be to open up the Agris bibliographic repository to the web. The connection will be provided by AGROVOC, which is used to index all documents in Agris: a search in Agris will then provide access to data residing in all data repositories linked through AGROVOC. It is for applications like that the Linked Data Given seems to show all its potential.

We are convinced that the landscape of data production will remain heterogeneous for many years, which makes important that tools to convert existing data into RDF and Linked Data formats be widely available to the community. Also, we encourage the development and sharing of applications that exploit the data produced. In this sense, we expect that interesting results will come from the recently started, EU funded project agINFRA.

---

[42] http://kt.mimos.my/AgroTagger/

[43] http://ring.ciard.net/

[44] http://www.egfar.org/egfar/

[45] An open source content management system. See: http://drupal.org/

Acknowledgments. Several people and groups in FAO, GFAR, CIARD, IIT Kanpur, and MIMOS Berhad have contributed to the ideas, tools and data sets mentioned in this paper. Among them, special our team in FAO Rome. Special thanks to MIMOS Berhad which founded the participation of Johannes Keizer in the Knowledge Technology Week in 2011, which took place in Malaysia.

References

[1] C. Bizer, T. Heath and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue on Linked Data*, vol. 5, no. 3, pp. 1-22, 2009.

[2] C. Bizer, R. Cyganiak and T. Heath, "How to Publish Linked Data on the Web," 27 July 2007. [Online]. Available: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

[3] A.-S. Dadzie and M. Rowe, "Approaches to Visaulising Linked Data: A Survey," *Semantic Web,* no. 1, 2011.

[4] C. Caraccciolo, A. Morshed, A. Stellato, G. Johannsen, Y. Jaques and J. Keizer, "Thesaurus Maintenance, Alignment and Publication as Linked Data: the AGROVOC use case," in *Metadata and Semantic Research 5th International Conference Proceedings*, Izmir, Turkey, 2011.

[5] Coherence in Information for Agricultural Research and Development, "Manifesto," 2010. [Online]. Available: http://www.ciard.net/sites/default/files/CIARD%20Manifesto%20EN_0.pdf.