OpenAGRIS: using bibliographical data for linking into the agricultural knowledge web

Fabrizio Celli, Stefano Anibaldi, Maria Folch, Yves Jaques, Johannes Keizer
FAO of the United Nations
Rome, Italy
Fabrizio.Celli@fao.org

***Abstract*— Spreading and exchanging agricultural information is a critical issue to allow researchers to access and use the knowledge in this sector. As a contribution to this goal, we propose a new approach that allows merging and integrating all information available on the Web about a specific agricultural topic by the usage of the most modern Linked Open Data technologies. We leverage on our previous work on AGRIS, a public domain database with nearly 3 million structured bibliographical records on agricultural science and technology. In this paper we illustrate a new Semantic Web platform, OpenAgris, which aggregates information stored in various sources available on the Web, providing much data as possible about a topic or a bibliographical resource.**

Keywords: linked open data, openagris, agricultural knowledge

I. BACKGROUND

AGRIS is an initiative that was set up by FAO of the United Nations in 1974 to make information on agriculture research globally available. The AGRIS portal [1] is one of the most important FAO web sites with an average 150,000 visits per month. Its core component is the AGRIS centralized database, a collection of nearly three millions structured bibliographic records, forming the corpus of one of the most important world-wide information systems in the area of the agricultural sciences.

The historical overall objective of AGRIS is to improve access and exchange of information serving "the information needs of developed and developing countries on a partnership basis" [2]. Over 150 participating data providers located in more than 100 countries are currently aggregating scientific and research publications, scholarly papers and grey literature which is not officially published in commercial channels.

Traditionally the AGRIS Centres sent periodically their data to the AGRIS Secretariat to have it published in the database. In the last years, not only traditional AGRIS Centres but also journal editors create their metadata to be published in the AGRIS database and with the growth of open access institutional repositories (IR), AGRIS has dramatically improved its methods for harvesting and indexing metadata from content providers. The OAI-PMH protocols currently allow the AGRIS service to facilitate the interoperability with service providers, by means of a more direct channel of communication between the two parties.

II. WHY OPENAGRIS

Bibliographic records are often static and do not contain sufficient information to the user, in particular if the annotation does not include the full text of the publication. A recent AGRIS site analysis shows that most of the times end users reach an AGRIS result page (a reference), when they see that the full text link does not exist, they exit and search again, using other search engines or databases, for other online resources. The "...digital environment has increased the range of user needs and expectations beyond the scope of the collections of most consortia. End-user services are now required to operate at web-scale and incorporate metadata from sources well beyond the traditional bibliographic record." [3]. By providing better access to bibliographic and citation data, scholars worldwide, specifically in the agricultural sector, will be the primary beneficiaries.

A. OpenAGRIS as Linked Open Data on the Semantic Web

With the exponential growth of open access repositories and web resources, it became clear that the AGRIS objectives required a thorough revision in the strategies and methods to use for a deployment of a system that would exploit the semantic richness of the AGRIS dataset.

"The ambition is not to collect comprehensively all bibliographic references in the subject area, but to use the latent knowledge in the AGRIS data to find, link and interpret relevant sources on the internet." [4].

AGRIS key scope in this project is to become the leading access point to scientific and technological knowledge in agriculture, in such a way that access to information will be facilitated for students and researchers, the end users who may be not satisfied when searching information either googling or ending up in multiple, sometimes federated databases.

AGRIS will be able both to interlink to other data sets in the relevant disciplines, in order to expand its intrinsic knowledge, and to become "the" Data Set to be consumed and queried by applications and systems in the agriculture sector.

The big challenge for the OpenAGRIS application is to produce a linked-data environment that will mashup the interlinked datasets, which will be combined to create meaningful results.

III. IMPLEMENTATION AND TECHNOLOGIES

Following the concepts discussed in previous sections we will illustrate OpenAgris, a web application entirely based on RDF that aggregates information from different sources to expand the AGRIS knowledge about a topic or a document. The idea is to develop an application that combines data in new ways and allows users to make connections, understand relationships that were previously hidden [5], and retrieve more information about something or hidden information (e.g. if a particular fish can reproduce in a special area with some environmental conditions). We want that OpenAgris will become the hub for agricultural information, a place where researchers and other stakeholders can find all the information they need about a topic, a bibliographical resource, an agricultural journal, and an author.

A. The process of RDF-ization

The main problem of AGRIS records is that, sometimes and in particular for old resources, there are only the title and the authors of a resource, while users look for the fulltext or for information related to the main topics of the resource. To solve this issue, we need to become part of the LOD cloud: in fact, Linked Data is the way to publish structured data and to interlink with other existing datasets. More, Linked Data "refers to data published on the Web in such a way that is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets" [6]. Becoming part of the LOD cloud means translating our repository of almost 3 million of XML bibliographical records to RDF [8], a language for expressing data models using statements expressed as triples (subject, predicate, and object), and publishing it on the Web. The translation process requires a design step to define vocabularies and properties we want to model. A vocabulary is a defined set of predicates that can be used to model semantic data: you can define a your own vocabulary for your application, but best practices suggest to use existing vocabularies (DC, FOAF, etc…) and to define new predicates only if there is nothing in the world that can express a desired relationship. We defined three different RDF datasets:

- *The AGRIS records dataset*, the direct translation of AGRIS XML records to RDF. Considering that AGRIS contains nearly 3 million of XML records, this new dataset will consist of nearly 50 million triples. This dataset is very relevant. In fact, even if sometimes there is a lack of information, AGRIS records are multilingual, so for a Chinese record we can have the English title and the Chinese one, and the same thing happens for the abstract. Moreover, we can interlink this dataset to the Agrovoc RDF dataset, and with the AGRIS journals one. As an example, the RDF/XML in Figure 1 explains a possible content of an AGRIS RDF record, even if authors URIs are at the moment fake.
- *The Agrovoc RDF dataset*: AGROVOC is the world's most comprehensive multilingual agricultural vocabulary that contains close to 40,000 concepts in over 20 languages covering subject fields in agriculture, forestry and fisheries together with cross-cutting themes such as land use, rural livelihoods and food security [7]. This thesaurus is used – among other things – to label AGRIS records and its RDF-ization allows interlinking to other thesauri (such as Eurovoc, NAL, DBPedia, etc…) and to extract more information about concepts.
- *The AGRIS journals dataset*: since the 82.11% of AGRIS records are journal's article, we created a dataset of agricultural journals with complete information about each journal (ISSN, start date, frequency, publisher…). The metadata information from the "AGRIS" scientific journals, together with other important data sources such as the journals of the FAO catalogue, the Directory of the Open Access journals (DOAJ [10]), CABI [11] and AGRICOLA [12] was expanded retrieving the authoritative information from the ISSN Centre database [13]. Although UDC and DDC authoritative classification systems codes are used to provided, it is with the AGROVOC thesaurus that semantic richness is provided to the dataset, enhancing the interlinking with other datasets. Its great power lays in its numerous term-to-term alignments to other important thesauri in the agricultural domain and to the direct and indirect mappings to other linked data sets, such as DBPedia. We interlinked this dataset with the AGRIS record one, assigning a journal's URI to an AGRIS record that belongs to that journal. This process is very complicated, since AGRIS data are sometimes very dirty: often we have the ISSN of the journal (i.e. a unique identifier for print or electronic periodical publications), and in this case the assignment of a journal to an AGRIS record is very easy. Other times the ISSN is not correct or we have only the title of a journal, that can contain misspellings or abbreviations: in this situation we need to compare journals' titles by using some string distance metric (since a simple "equals" between strings is not sufficient) and to assign an ISSN to a title (or an ID if some journals/series don't have an ISSN). At the moment, we have disambiguated the 90.3% of AGRIS journals, obtaining for them complete information. We have nearly 20,000 agricultural journals stored in a triple store, for a total of almost 320,000 triples. Each journal has also been classified according to some thesauri, such as Agrovoc and DDC [9]. Figure 2 shows an example of the AGRIS journals RDF/XML.

B. The implementation

OpenAgris is a Semantic Web application that aggregates information stored in various triple stores available on the Web. The core of this application is the AGRIS records RDF dataset, stored in an Allegrograph RDF store [15]: starting from the resource requested by the user - and identified by a special number called ARN - OpenAgris queries the Agrovoc RDF repository to extract keywords for the specific resource and relationships to other datasets, such as DBPedia. If the resource is a journal's article, the engine also queries the AGRIS journals dataset, obtaining complete information about the journals and related articles from the same journal and about the same topics. This process can be extended to all areas of interest of the linked open data cloud, obtaining all possible

information about the specific resource and its main topics. Figure 3 explains this data flow.

From an architectural point of view, since filtered Sparql queries on a triple store can be not so very efficient, OpenAgris uses an Apache Solr index to show immediately information about the record. Then, various threads are responsible to query the other triple stores which we are linking to by using their Sparql endpoints when available, or other available APIs. A beta release of OpenAgris is available from the AGRIS website (http://agris.fao.org/). As an example, you can exploit the resource:

http://agris.fao.org/openagris/search.do?recordID=JP2010001379

The most significant process is the RDF-ization of AGRIS records. First of all, we need a unique URI for each resource (AGRIS record): each of the published URIs represents a unique means to identify a specific resource and provides researchers with an exhaustive map of the global research community, linking formal outputs (papers, conference proceedings, etc.) with other information available in other datasets. Then there is the process of translation of the XMLs to RDFs. As we explained in previous sections, AGRIS is a repository of nearly 3 million of XML bibliographical records and every month we receive new records, so the translation to RDF cannot be performed only the first time but it requires continuous automatic updates. Moreover, the whole process is based on operations of transformation and enrichment of the data, so it requires well defined and structured steps with a precise flow of information between each step. We can summarize these steps as shown in the Figure 4. The process takes in input a set of AGRIS XML records (but it can be expanded for any data source, like data crawled from the Web) and loads them into a DBMS, in an object-relational data model. This step is necessary to ensure that no duplicate records will be added to the repository and to modify data using a common query language, SQL. Then, these data are consumed by a filter that transforms them (data type conversion, data cleaning, data validation) and enriches them with information taken from different controlled and authoritative sources, like the ISSN Centre database [13], the FAO Open Archive, etc… At the end, data are ready to be converted to RDF, loaded into a triple store and disseminated on the Web.

This process of RDF-ization is completely automatic: data come directly from AGRIS records, which are structured data so we only needed to define a mapping between the AGRIS XML and the AGRIS RDF, with some filters to check the correctness of some information (i.e. the date's format, ISSNs, etc…) and to add more data by using predefined data sources that have been correctly mapped to AGRIS. Then, when new records are added to the AGRIS repository, they are automatically translated to RDF. Therefore, accuracy is 100% from our side, since no uncertainty is added to the dataset: critical operations, such as the mapping from Agrovoc to other datasets and requests for more information to ISSN Centre database [13], require a human intervention, while the rest of the translation process is controlled by filters that check data correctness and trust in the data sent by AGRIS centers as bibliographical data.

Moreover, maintaining a wealth of knowledge such as that OpenAGRIS is publishing implies the implementation of a mechanism that will keep track of the provenance of such information, which is coming from diverse data sources – it will be required essentially to give credit to the AGRIS partners when other datasets will reuse their information. The W3C Provenance Incubator Group [14] is currently working towards the development of a roadmap in the area of provenance for Semantic Web technologies. It is not clear yet what tools and how this will be effectively done.

```
<bibo:Article rdf:about="http://agris.fao.org/resource/CN2009001544">
  <dct:identifier>CN2009001544</dct:identifier>
  <dct:title xml:lang="en"><![CDATA[Cultural supernatants of
erythrocytes from adult and infant dogs promoting Lymphocytes
proliferation]]></dct:title>
  <dct:title xml:lang="Zh"><![CDATA[不同年龄犬红细胞培养上清对
淋巴细胞增殖的作用]]></dct:title>
  <dct:creator rdf:resource="http://agris.fao.org/author/zhangwenli"/>
  <dct:issued>jun2008</dct:issued>
  <dct:subject rdf:resource="http://aims.fao.org/aos/agrovoc/c_2352"/>
  <dct:subject rdf:resource="http://aims.fao.org/aos/agrovoc/c_27517"/>
  <bibo:language>Zh</bibo:language>
  <dc:isPartOf rdf:resource="http://aims.fao.org/serials/c_67c3411a"/>
</bibo:Article>
```

Figure 1. RDF/XML of an AGRIS record.

```
<bibo:Journal rdf:about="http://aims.fao.org/serials/c_249cf355">
 <bibo:ISSN>0028-4793</bibo:ISSN>
 <dct:title>New England journal of medicine</dct:title>
 <dct:alternative>The New England journal of medicine
 </dct:alternative>
 <ags:publisherPlace rdf:resource=
   http://aims.fao.org/aos/geopolitical.owl#United_States_of_America
   />
 <ags:publisherName>Massachusetts Medical Society.
   </ags:publisherName>
 <dct:language>eng</dct:language>
 <dct:issued>1928</dct:issued>
 <foaf:homepage>http://www.nejm.org/content/index.asp
   </foaf:homepage>
 <dct:accrualPeriodicity>weekly</dct:accrualPeriodicity>
 <dct:subject
   rdf:resource="http://aims.fao.org/aos/agrovoc/c_10394"/>
 <dct:subject rdf:resource="http://dewey.info/class/610/about"/>
 <bibo:coden>NEJMAG</bibo:coden>
</bibo:Journal>
```

Figure 2. RDF/XML of an AGRIS journal record.

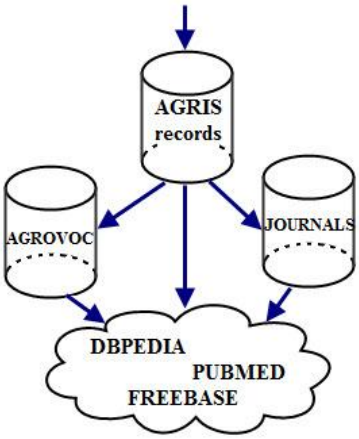

Figure 3. The OPENAGRIS data flow.

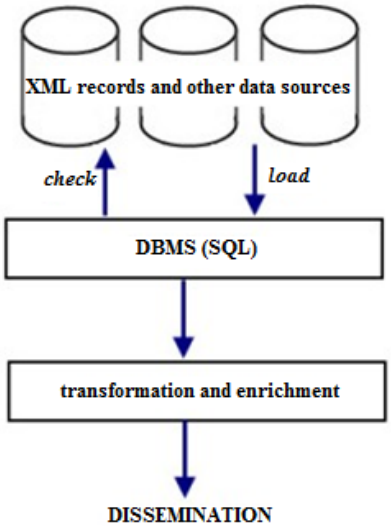

Figure 4. The process of RDF-ization.

IV. CONCLUSIONS AND FUTURE WORK

The ways in which users search and discover information has changed enormously with the rise of the Internet. It is becoming less and less likely that users are browsing bibliographic data in order to find a print publication that they will then order from a publisher or research library. Users are more and more expecting to simply search subject categories in a generic search engine and click on the results to open full text journal articles, conference proceedings and book excerpts. This implies that traditional bibliographic databases need to find new ways in which to offer value to end users. They must somehow leverage their often rich set of authoritative, disambiguated and interrelated metadata to take advantage of the new sea of external data that is increasingly being exposed via web services and exploited using technologies such as Asynchronous JavaScript and XML (AJAX).

The OpenAGRIS pilot does just that. By combining disambiguated, authoritative records for journals together with a rich set of multi-lingual thesaurus concepts that are in turn broadly linked to other subject thesauri, OpenAGRIS is able to maximize a powerful set of related metadata by using it first against an ever increasing set of semantic web services and then pragmatically against other, useful web services whether semantic or not.

In its original conception OpenAGRIS was to be a purely RDF-based application. Experience showed that only a few domains intersecting with the agricultural domain offered RDF web services. Pragmatic considerations of basic data availability thus necessitated the use of non-RDF APIs.

Despite the relative lack of RDF-based APIs for agricultural use, the field is growing. As new, relevant RDF APIs become available OpenAGRIS will continue to exploit them, while at the same time doing its part by producing AgriBase, a Freebase-like aggregation API for agricultural data. This challenging endeavor will necessitate wrapping and re-interpreting non-RDF APIs. It will be available via SPARQL endpoint as well as via RDF and non-RDF web services.

Along with a recent rise in the number of RDF-based services, data providers are increasingly providing web-service widgets returning complex functionality such as static and dynamic tables, maps, charts and graphs. OpenAGRIS will also begin to exploit such rich services, relying on its unique set of citation-based metadata relationships to deliver relevant information in graphical form to end users.

Finally, behind the scenes the group responsible for the publication of AGRIS has been quietly publishing and linking an ever increasing set of authoritative RDF vocabularies, from thesauri, to journals, to corporate authors. This growing web of formally linked metadata will continue to improve the relevance of AGRIS records as they reach ever further into the linked data cloud.

REFERENCES

[1] The AGRIS Portal, http://agris.fao.org

[2] "AGRIS introduction", FAO 1981, printed.

[3] Bibliographic Data Cluster, W3C, 2010 http://www.w3.org/2005/Incubator/lld/wiki/Cluster_BibData

[4] "AGRIS 2008-2010: A portal to access global knowledge in agricultural research and technology", FAO, 2008

[5] T. Segaran, C. Evans, and J. Taylor. "Programming the semantic web", O'Reilly 2009.

[6] C. Bizer, T. Health, and T. Berners-Lee. "Linked Data – The Story So Far". IJSWIS.

[7] The AIMS website, http://aims.fao.org

[8] Resource Description Framework (RDF), http://www.w3.org/RDF/

[9] Dewey Decimal Classification, http://dewey.info/

[10] Directory of Open Access Journals, http://www.doaj.org/

[11] Centre for Agricultural Bioscience International, http://www.cabi.org/

[12] National Agricultural Library Catalog, http://agricola.nal.usda.gov/

[13] ISSN portal, http://www.issn.org/

[14] W3C Provenance Incubator Group Wiki, http://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki

[15] Allegrograph RDF store, http://www.franz.com/agraph/allegrograph/