Pushing, Pulling, Harvesting, Linking: Rethinking Bibliographic Workflows for the Semantic Web

Fabrizio Celli[1], Yves Jaques[1], Stefano Anibaldi[1], Johannes Keizer[1]

[1]Food and Agriculture Organization of the United Nations
A103, viale delle Terme di Caracalla, 00153, Rome, Italy. Yves.Jaques@fao.org

ABSTRACT

In this paper we describe the ongoing move of the AGRIS repository toward a decentralized approach based on Linked Open Data (LOD) (Bizer, *et al.*, 2008). This move has progressively required modifications and enhancements to data, models and workflows. The growing demand for freely accessible data has brought a rise in data distributed using LOD, which combines Resource Description Framework (RDF) (McBride, 2004a) and RDF Schema (McBride, 2004b) with vocabularies such as Dublin Core (DC) (Miles, *et al.*, 2009) and Simple Knowledge Organisation System, together with interfaces such as SPARQL query language for RDF (Prud'hommeaux, *et al.*, 2008). While LOD implementations are by now a well-established pattern, the impacts that such approaches have on underlying business processes is less well understood. The openness of the LOD paradigm can expose flaws in information management workflows. Poor metadata, lack of metrics, vague provenance; all can contribute to the inability of an LOD-enabled system to satisfy the demands of the Semantic Web.

Keywords: Linked open data, bibliographic data, agriculture, RDF, repositories

1. INTRODUCTION

AGRIS is an initiative set up by FAO in 1974 to make information on agriculture research and related sciences globally available. Its content, exposed using a qualified DC metadata format, is either manually created by experienced cataloguers or automatically harvested from a wide variety of OAI-PMH targets, a community that represents a global network of over 100 content providers.

The AGRIS service consumes metadata provided by the community and publishes it as open data. The metadata is captured either by *pulling* data through *harvesting* from clients, e.g. aggregators and institutional repositories using protocols such as OAI-PMH; or by data being *pushed* to AGRIS from clients, e.g. national libraries or journal publishers. The integration of this heterogeneous metadata, with issues ranging from multilingual content, text normalization, differing cataloguing rules and diverse metadata formats, has become a key issue as each of the sets or collections harvested is a unique case that must be handled and processed on its own.

F. Celli, Y. Jaques, S. Anibaldi, J. Keizer. "Pushing, Pulling, Harvesting, Linking: Rethinking Bibliographic Workflows for the Semantic Web". EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, 24-27 June 2013.

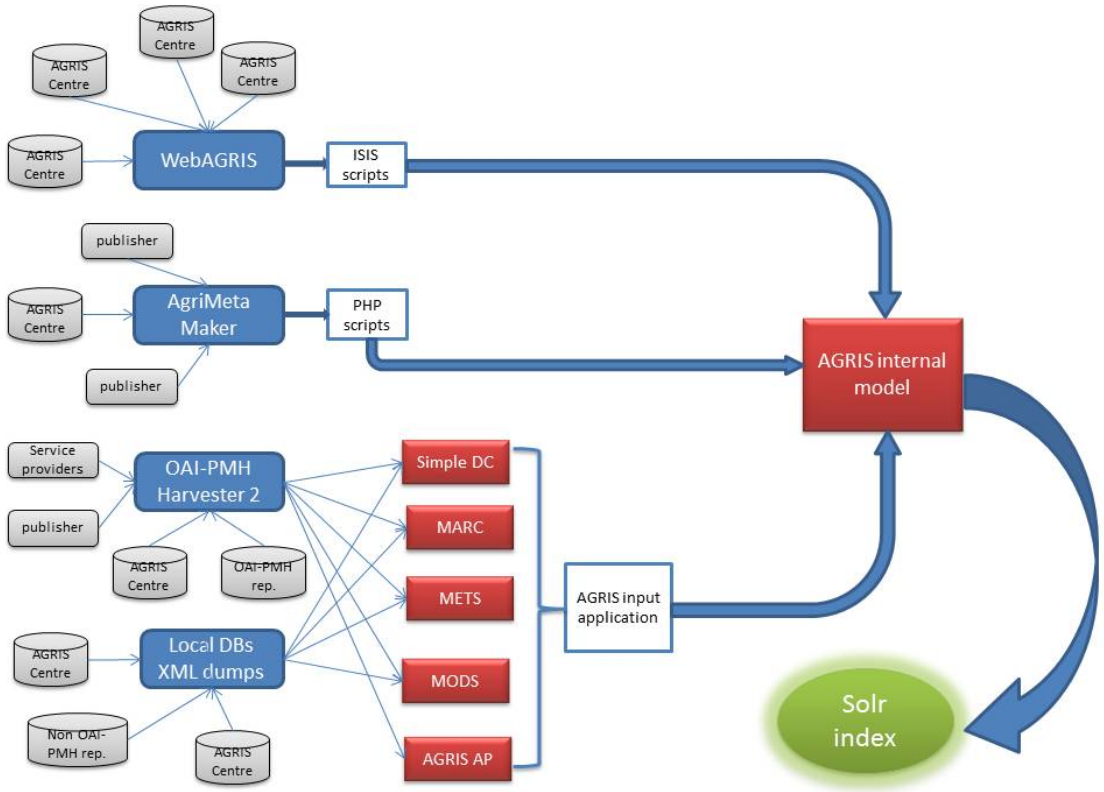

Figure 1 - AGRIS data consumption workflow

2. AGRIS DATA CONSUMPTION

The AGRIS database is a collection of more than 4.3 million bibliographic records many of which link to full-text documents. They are enhanced by the AGROVOC thesaurus which is extensively used by cataloguers world-wide to enrich data indexing in agricultural information systems. The AGRIS database can be queried via the AGRIS search engine (http://agris.fao.org/), a web application based on the Apache Solr search API.

In the paper era, data were catalogued and delivered to the central database by national libraries (AGRIS Centres) via paper worksheets and floppy disks. In recent years AGRIS dramatically improved its methods for harvesting and indexing metadata from content providers, also thanks to the growth of open access institutional repositories. In addition, not only traditional AGRIS Centres but also journal publishers now create metadata for publishing in the AGRIS database.

AGRIS uses various tools and technologies to consume metadata from content providers (Fig. 1), but stressing the importance of a low barrier for providers, AGRIS accepts any metadata record that meets the basic standards set out in the Meaningful

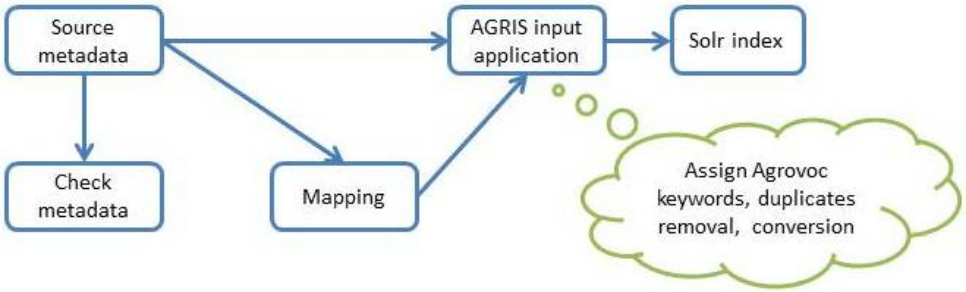

Figure 2 - AGRIS data normalization

Bibliographic Metadata (M2B) recommendations (Subirats, *et al.*, 2012), including standards such as simple DC, MODS, MARCXML, MARC21 and PubMed.

Looking at the *push* side of data collection, AGRIS content providers produce metadata using different methods and systems before sending them to AGRIS via an FTP upload page. They may also directly generate metadata records with the AgriMetaMaker; a web form developed using the Drupal CMS that allows users to create records with a predefined set of mandatory fields. One of the advantages of using the AgriMetamaker is that metadata can be directly indexed with AGROVOC keywords, querying and browsing the AGROVOC thesaurus to choose the desired ones.

On the *pull* side of data collection, AGRIS consumes metadata using an OAI-PMH *harvester*, a semi-automatic process which simplifies the overall workflow. Content providers publish their records compliant the OAI-PMH protocols and the AGRIS team uses a harvester tool to retrieve these data periodically. This step sometimes requires a preliminary selection of the most suitable and domain-specific list of target providers.

3. AGRIS DATA VALIDATION

Once the AGRIS Secretariat has received metadata in a specific metadata format it is necessary to transform and validate the record using a mixture of automatic, semi-automatic and manual methods. Metadata are randomly checked to look for inconsistencies or recurring semantic errors. This is a manual job, which can partially assure the high quality of the AGRIS content, even if the AGRIS team cannot be responsible for the content of its data providers.

After the metadata have been validated by the AGRIS team, they are mapped and translated to AGRIS' internal model and indexed using Apache Solr, a popular open-source search engine. AGROVOC keywords are also extracted and assigned to any record that does not already contain them. This task is performed using Maui, a popular open source keyword extraction framework that uses University of Waikato's Keyword Extraction Algorithm (KEA). Finally, duplicates are detected and removed, as the same record may be indexed in multiple collections or be duplicated in the same repository.

C0284
F. Celli, Y. Jaques, S. Anibaldi, J. Keizer. "Pushing, Pulling, Harvesting, Linking: Rethinking Bibliographic Workflows for the Semantic Web". EFITA-WCCA-CIGR Conference "Sustainable Agriculture through ICT Innovation", Turin, Italy, 24-27 June 2013.

Once metadata have been consumed, transformed and enhanced with AGROVOC (Fig. 2), they are ready to be indexed using Solr after which they can be searched and retrieved by the AGRIS search engine (http://agris.fao.org/).

4. THE ROAD TO LINKED OPEN DATA

Bibliographic records are often static and often don't contain sufficient information to answer a user's query, in particular when the record does not include the full-text link to the publication. In 2011, an AGRIS site analysis showed that many users reaching an AGRIS result page (a record) left the site immediately if they did not find a full-text link (Celli, *et al.*, 2011). The move to RDF and Linked Open Data has enabled AGRIS to better meet its customers' expectations by providing disambiguated, entity-based access to bibliographic and citation data and by mashing up this information with related data sources by taking advantage of the many formal alignments between AGROVOC and other knowledge organization systems. Following the innovation began by the porting of its indexing thesaurus AGROVOC to a Simple Knowledge Organization System (SKOS) concept-scheme published as LOD (Caracciolo, *et al.*, 2012), the decision was made to connect the AGRIS content to the LOD cloud as well, and to fully exploit the potentialities of the Semantic Web.

Becoming part of the LOD cloud meant translating the repository of more than 4.3 million XML bibliographic records to RDF, and publishing them on the Web. The conversion process required a design step to define vocabularies and properties needed in order to model the data. In the spirit of reusability the team almost completely avoided minting any new properties, reusing as much as possible available standard vocabularies such as BIBO, FOAF and Dublin Core. This process was facilitated by the M2B recommendations (Subirats, *et al.*, 2012), which support the selection of appropriate encoding strategies for producing meaningful Linked Open Data enabled bibliographical data (LODE-BD). But the definition of the vocabulary was not the only issue in this process; the AGRIS team had to review the entire business process, adding new steps in order to cope with the requirements of Linked Open Data while also modifying existing processes to facilitate the migration to RDF.

First of all, the move to RDF required the disambiguation of the AGRIS content and the cleaning of the data. Since 78% of AGRIS records are journal articles, the AGRIS team created a disambiguated RDF dataset of agricultural journals by combining journals from AGRIS, FAO, the Directory of Open Access journals (DOAJ), CABI and AGRICOLA, expanded with authoritative information from the ISSN Centre database (http://www.issn.org/). Unfortunately, AGRIS data were sometimes very dirty; in some cases the ISSN of the journal was present and assignment of a journal to an AGRIS record was easy. In other cases however, the ISSN was not correct or the record contained only the journal title, often with misspellings or abbreviations: in these cases

the AGRIS team compared journal titles by using string distance metric (Levenshtein algorithm) and assigning ISSN to titles. At the moment, there are more than 20,000 disambiguated agricultural journals stored as a set of almost 360,000 triples.

```
<bibo:Article rdf:about="http://agris.fao.org/aos/records/XS2010X00001">
  <dct:identifier>XS2010X00001</dct:identifier>
  <dct:title xml:lang="pt">Caracteristicas anatômicas ...</dct:title>
  <dct:title xml:lang="en">...</dct:title>
  <dct:creator>
    <foaf:Person><foaf:name>Mesquita, Alessandro Carlos</foaf:name></foaf:Person>
  </dct:creator>
  <dct:publisher>
    <foaf:Organization><foaf:name>Instituto Nacional de ...</foaf:name></foaf:Organization>
  </dct:publisher>
  <dct:issued>2010</dct:issued>
  <dct:subject rdf:resource="http://aims.fao.org/aos/agrovoc/c_6200"/>
  <bibo:abstract xml:lang="pt"><![CDATA[As estruturas envolvidas na produção ...]]></bibo:abstract>
  <bibo:abstract xml:lang="en"><![CDATA[The structures involved in latex production ...]]></bibo:abstract>
  <bibo:uri><![CDATA[http://www.scielo.br/scielo.php?pid=...]]></bibo:uri>
  <bibo:language>por</bibo:language>
  <dct:isPartOf rdf:resource="http://aims.fao.org/serials/c_e8d916a8"/>
</bibo:Article>
```

Figure 3 - The RDF/XML serialization of an AGRIS record

In order to represent AGRIS records in RDF it was necessary to select the set of RDF classes and properties that would best describe the records. It was also necessary to clean the content (for instance, to have dates in a common format), generate URIs for each record, associate each record to its unique RDF journal record and convert all AGROVOC terms to their RDF URI equivalents. Fig. 3 gives an example of the result.

The importance of having AGROVOC URIs associated to each AGRIS record lies in the fact that AGROVOC is used as a common schema to interlink to external datasets and sources of information. In fact, AGROVOC contains many alignments to other vocabularies (e.g. DBPedia, FAO Geopolitical Ontology, etc.) that allow querying triplestores to retrieve external resources. Moreover, AGROVOC keywords can also be used to query traditional Web Services to retrieve non-RDF data.

When the user comes to an AGRIS record using Linked Open Data technologies (http://agris.fao.org/openagris/), the system is able to display related information such as production statistics of aquatic species, species occurrence maps, World Bank indicators and more, all dynamically queried through a constellation of related keywords and vocabulary alignments. Currently, the AGRIS records dataset contains more than 115 million triples, stored in a triplestore that provides a public SPARQL endpoint.

Another important issue is related to the provenance of AGRIS resources (Jaques, *et al.*, 2012). Provenance is a broad term that may refer to various levels of granularity. With the shift to digital publishing and machine-readable resources tracing the provenance

chain has gained new importance. The AGRIS team follows W3C recommendations to cope with provenance and has begun implementing metadata provenance in AGRIS records. In AGRIS, each record has an identifier known as an ARN (AGRIS Record Number), which has a predefined structure and contains information on the data source together with the bibliographic record's year of creation. With this unique identifier, AGRIS provides precise and updated statements of the origins of the metadata. For instance, the ARN "IT2008000091" refers to a record created in 2008 from a specific AGRIS data provider in Italy, whose progressive number is 91. The team triplified information about the AGRIS data providers, providing unique URIs for each and adding triples (using the property *dct:source*) to identify this aspect of provenance, which previously was only implicit in the ARN.

5. NEW SCENARIOS

The road to RDF has had a big and continuing impact on the AGRIS workflow. Starting from the initial data consumption workflow (Fig.1), Fig.4 shows how the flow proceeds to the RDF conversion. In this yet to be implemented scenario, the AGRIS input application directly generates RDF as the AGRIS internal model and applies all needed filters: entity disambiguation (journals, publishers, and data providers), content cleaning, AgroTagger to generate AGROVOC URIs for the resource, provenance, etc. An RDF manager application is then responsible for adding triples to the AGRIS triplestore and indexing them as required by Solr.

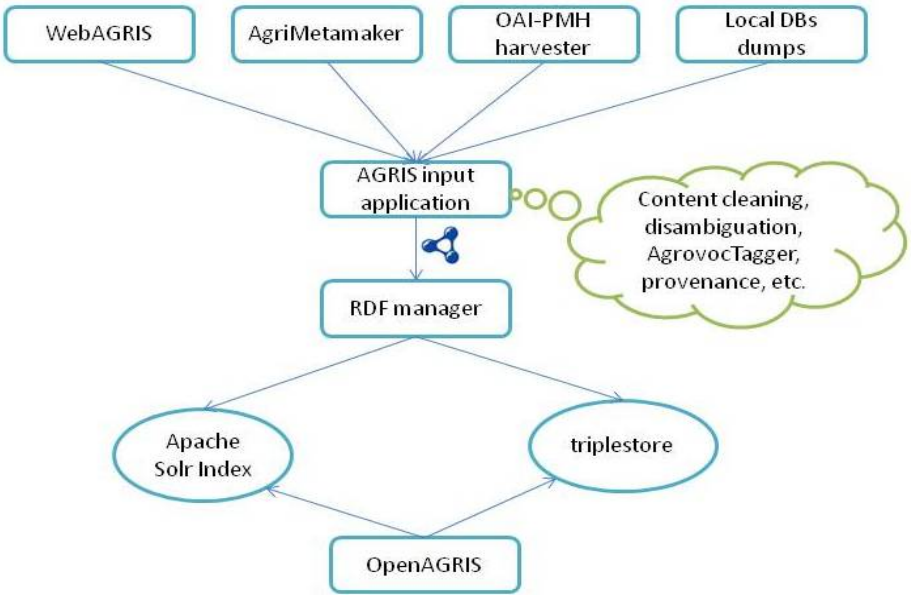

Figure 4 - The new AGRIS dataflow

At the end of this process there is the new RDF-aware system, currently known as OpenAGRIS (Fig.5), which is a mashup application that allows users to query the AGRIS content, interlinking all records to external sources of information and discovering the full text of a publication by using Google Custom APIs.

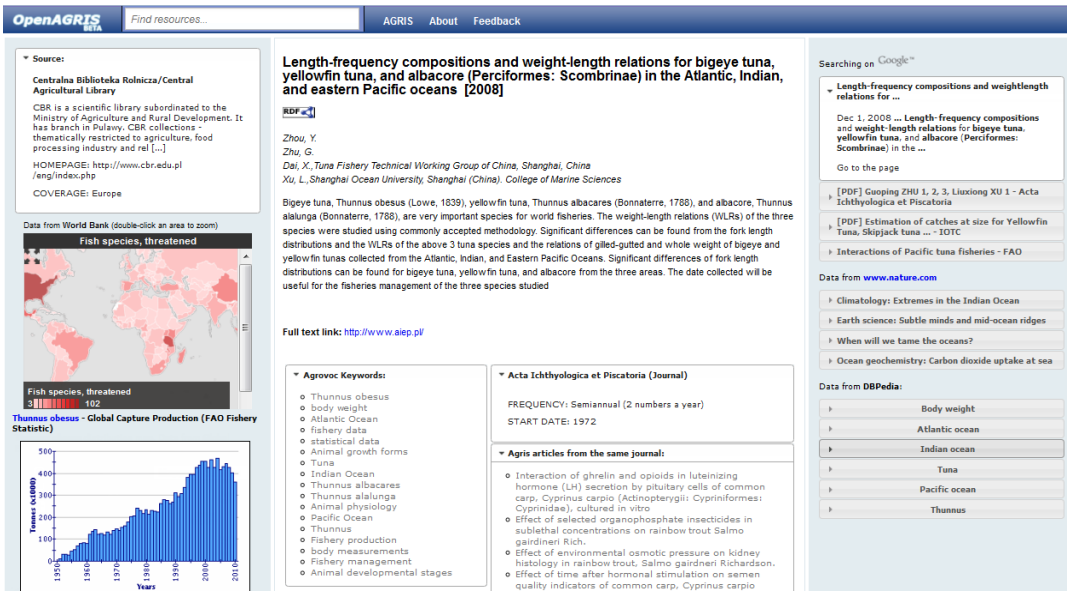

Figure 5 - The OpenAGRIS semantic mashup

In the future, the AGRIS team will have to cope with new issues and new functionalities that will again alter the workflow. To begin with there is the idea of a web bot able to crawl the web, extract triples and automatically index resources in the AGRIS triplestore. The automatic RDF consumption and the web crawler raise at least two important issues: the provenance of each triple added to the AGRIS triplestore and the selection of appropriate AGRIS content (which could be partially solved by manually selecting the initial subset of web sites to harvest and running the AgroTagger on the discovered resource to extract AGROVOC keywords) as well as the fact that the disambiguation process is at an early stage; the AGRIS team will soon begin to work on author disambiguation which will have a big impact on the entire system, with the possibility of the creation of a self-care system in which authors and/or institutions can modify their own profiles and data records.

6. RESULTS AND CONCLUSIONS

The conversion to RDF had a big impact on the AGRIS workflow. The team had to select RDF properties according to M2B and LODE-BD recommendations,

disambiguate AGRIS serials, check metadata content, and index AGRIS records with AGROVOC URIs to allow interlinking to other Web resources. Finally, the master repository became a triple-store with over 115 million triples, requiring changes to the consumption and indexing workflow which are still taking place. This move to RDF which vastly improved the dissemination and linking of the records nevertheless impacted the entire business process, including the model, the content and the validation. It also brought to light deficiencies in the collection of process metadata such as provenance. We continue to work to address issues such as author and institution disambiguation.

7. REFERENCES

Bizer, C., et al. 2008. Linked Data: Principles and State of the Art. [Online]. Available http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

Caracciolo, C., A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jacques e J. Keizer. 2012. Thesaurus Maintenance, Alignment and Publication as Linked Data - The AGROVOC Use Case. International Journal of Metadata, Semantics and Ontologies.

Celli, F., S. Anibaldi, M. Folch, Y. Jaques, J. Keizer. 2011. OpenAGRIS: using bibliographical data for linking into the agricultural knowledge web.

Jaques, Y., S. Anibaldi, F. Celli, I. Subirats, A. Stellato and J. Keizer. 2012. Proof and Trust in the OpenAGRIS Implementation. International Conference on Dublin Core and Metadata Applications (DC2012), North America.

McBride, B. 2004a. RDF Primer. [Online]. http://www.w3.org/TR/rdf-primer/

McBride, B. 2004b. RDF Vocabulary Description Language 1.0: RDF Schema. [Online]. http://www.w3.org/TR/rdf-schema/

Miles, A., S. Bechhofer. 2009. SKOS Simple Knowledge Organization System Reference. [Online]. http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Prud'hommeaux, E., A. Seaborne. 2008. SPARQL Query Language for RDF. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

Subirats, I., M. Zeng. 2012. Meaningful Bibliographic Metadata. [Online]. Available: http://aims.fao.org/metadata/m2b