DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2012

Proof and Trust in the OpenAGRIS Implementation

Yves Jaques
FAO of the UN, Rome
yves.jaques@fao.org

Stefano Anibaldi
FAO of the UN, Rome
stefano.anibaldi@fao.org

Fabrizio Celli
FAO of the UN, Rome
fabrizio.celli@fao.org

Imma Subirats
FAO of the UN, Rome
imma.subirats@fao.org

Armando Stellato
Univ. of Tor Vergata, Rome
stellato@info.uniroma2.it

Johannes Keizer
FAO of the UN, Rome
johannes.keizer@fao.org

Abstract

The AGRIS repository is a bibliographic database covering almost forty years of agricultural research. Following the conversion of its indexing thesaurus AGROVOC into a concept-based vocabulary, the decision was made to express the entire AGRIS repository in RDF as Linked Open Data. As part of this exercise, a semantic mashup named OpenAGRIS was developed in order to access the records and use them to dynamically display related data from external systems through both SPARQL queries and traditional web services. The overall process raised numerous issues regarding the relative lack of administrative metadata required to compellingly address the top proof and trust layers of the semantic web stack, both within the AGRIS repository and in external data dynamically pulled into OpenAGRIS. The team began by disambiguating the journals in which the articles were published and converting them into RDF but quickly realized this was only the beginning of a series of necessary steps in moving from a closed to an open world paradigm. Further disambiguation of institutions, authors and AGRIS Centres as well as the use of the VoiD vocabulary and of quality indicator models are discussed and evaluated.

Keywords: Semantic Web; proof; trust

1. Introduction

The Proof and Trust layers (Fig.1) of the Semantic Web stack are well researched although functioning examples that implement these layers in robust end-user production systems are few. Although "Linked Data should be published alongside several types of metadata, in order to increase its utility for data consumers" (Bizer et al., 2009), Linked Data has finally gotten off the ground by focusing on the lower layers, figuring the rest will sort itself out. This may be a mistaken assumption given that data consumers who need proof and trust typically have no relationship to data producers in a position to provide it.

This paper discusses the OpenAGRIS (Celli et al., 2011) semantic mashup implementation whose requirements necessitated a move on the part of the AGRIS (http://agris.fao.org) bibliographic repository from closed to open world assumptions. This migration brought to light deficiencies in data production, in particular the handling of proof and trust in a world of machine-readable linked data. The paper covers a number of initial issues that were resolved and finishes with an overview of proposals aimed at partially or wholly remedying the remaining proof and trust deficiencies in the AGRIS repository.

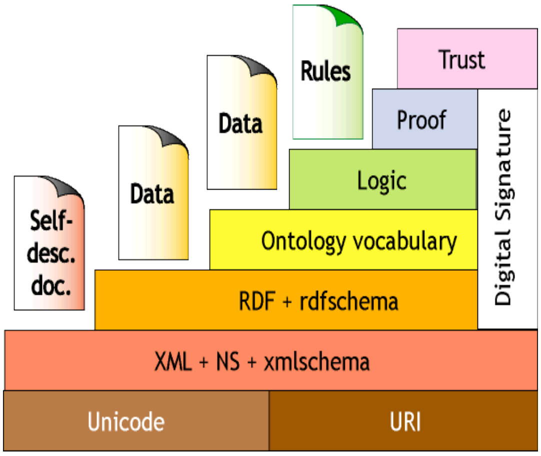


FIG. 1: The Semantic Web layers (Berners-Lee, 2000)

OPEN ACCESS

This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952136043

Since 1975, following an FAO (Food and Agriculture Organization of the United Nations) initiative, the International Information System for the Agricultural Sciences and Technology (AGRIS) has been collecting and disseminating bibliographic information on scientific and socio-economic publications issued on a wide variety of food and agricultural matters from over 150 heterogeneous Institutional Repositories worldwide.

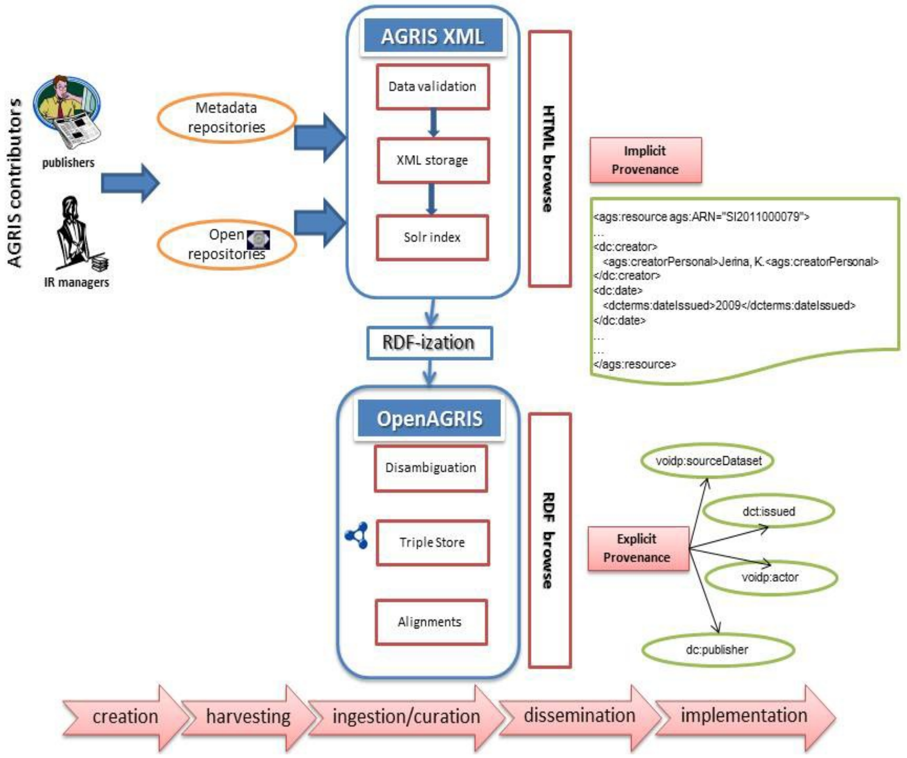
AGRIS is an international cooperative system that serves developed and developing countries in order to give scientists and students free access to agricultural knowledge. The AGRIS repository, a collection of nearly 2.9 million bibliographic references is encoded in an XML qualified Dublin Core metadata format that eases sharing of information across dispersed bibliographic systems. The AGROVOC thesaurus, extensively used by cataloguers to enrich data indexing in agricultural information systems, enhances its high quality content description.

2. The Road to Linked Data

2.1. The AGRIS artifact and its administrative data

In recent years the life cycle of an AGRIS record has changed enormously. In the past, data were catalogued and delivered to a central database by national libraries (traditional AGRIS Centres) via floppy disks and email. However, with the advent of the Open Access movement and the proliferation of OAI-PMH repositories, AGRIS modified its approach and began to also index data harvested from service providers such as DOAJ (Directory of Open Access Journals), whose content comes from external publishers.

When AGRIS decided to publish its records as linked data in RDF, it quickly became clear that crucial metadata necessary in addressing issues of proof and trust were missing. It goes without saying that "as the number of repositories and aggregators increases, so too does the number of potential formal or informal metadata sources" (Tonkin et al., 2006). Fig. 2 shows the long flow of an AGRIS artifact, from genesis to dissemination. Every phase generates administrative metadata and for each record, AGRIS has always registered authors, titles, dates and the cataloguing institute.



```
<ags:resource ags:ARN="SI2011000079">
...
<dc:creator>
  <ags:creatorPersonal>Jerina, K.<ags:creatorPersonal>
</dc:creator>
<dc:date>
  <dcterms:dateIssued>2009</dcterms:dateIssued>
</dc:date>
...
...
</ags:resource>
```

FIG. 2: Derivation history of an AGRIS artifact

OPEN ACCESS
This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

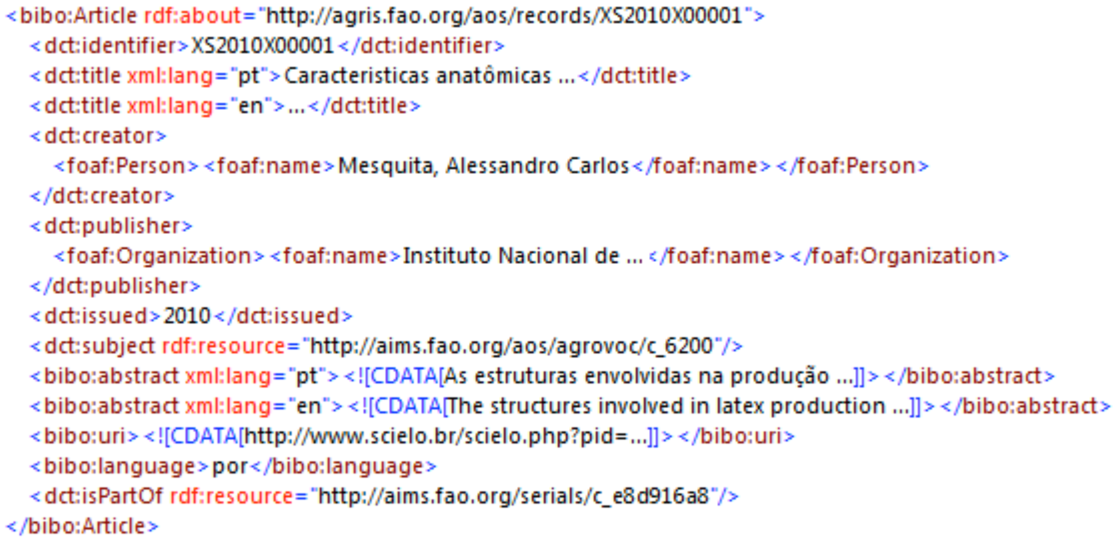
https://doi.org/10.23106/dcmi.952136043

The work of converting AGRIS to RDF quickly brought out the need to not only produce administrative data, but to associate wherever possible to disambiguated data in order to enhance traceability. While the eventual goal is to completely disambiguate journals, authors and institutions, AGRIS has thus far only disambiguated its journal references, an arduous nine month process that resulted in 20,000 unique journals, in themselves a precious resource for the agricultural research for development community.

2.2. The OpenAGRIS Semantic Mashup

In the last two years, AGRIS has focused attention on the metadata in which end users are interested and on the ways it is possible to enrich these metadata. AGRIS references often suffer from a lack of complete information and in particular of full text links. Only 4% of the entire collection has a working full text link. Accordingly, if a user wants to get more information on a specific topic they must use Google or other search engines to retrieve the publication.

The team decided to treat AGRIS records abstractly as metadata sets that could be leveraged to automatically access and display related data. It developed OpenAGRIS, a semantic mashup that aggregates information from different Web sources using AGRIS records exposed as sets of triples in a Linked Open Data environment. An AGRIS record represented in RDF (Fig. 3) thus becomes the entry point for a mechanism that discovers related web resources primarily via AGROVOC keywords. AGROVOC, organized using Simple Knowledge Organization System (SKOS), contains many alignments to other vocabularies (e.g. DBPedia, FAO Geopolitical Ontology, etc.) that allow querying triple stores to retrieve external resources. Moreover, AGROVOC keywords can also be used to query traditional Web Services (e.g. World Bank, FAO fisheries dataset, etc.) to retrieve non-RDF data. The system currently displays aquatic species production statistics, species occurrence maps, World Bank indicators and more, all dynamically queried through a constellation of related keywords and vocabulary alignments.



```
<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>
```

FIG. 3: The RDF/XML serialization of an AGRIS record

OpenAGRIS is an environment that allows the team to test and raise issues related to the implementation of end user systems based on the Semantic Web, and helps reveal problems in the proof and trust layers of the AGRIS system as well as in most of the systems from which it retrieves data. The first major issue to arise was that of provenance. AGRIS records contained only partial provenance metadata, and most external sources that were accessed contained even less. DBPedia is a typical example in that by its very nature, WIKI data is highly collaborative and almost immune to strong provenance tracking.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952136043

The AGRIS triple store currently contains almost 60 million triples, and ideally each should be linked to administrative metadata. In fact, information about a specific AGRIS record often comes from different entities, each of whom have provided part of the metadata during the record lifecycle. It should be possible to know from where the information was extracted, who submitted it and the primary data source. Although this primary source -- trustworthy or not -- creates the information, it is often only possible to determine the last provider in the chain (e.g. DOAJ).

Licensing is another aspect which is often underestimated and/or kept off the explicit data level. Quality indicators are also typically non-existent. Finally, there are semantic issues related to vocabulary alignment that can impact the correctness of dynamically retrieved data. A common situation is that two vocabularies representing the same concept with the same name are attached to data that is contextually very different and even wrong from the user's point-of-view. For example, one vocabulary may have a commercial view of a concept while another has a scientific view. SKOS' deliberately fuzzy definition of exactMatch and closeMatch properties, while avoiding ontological over-commitment does little to assist in this regard.

3. Improving AGRIS Proof and Trust

The issues discussed in the previous section refer to the world of Linked Open Data (LOD) and are typically ignored either partially or wholly partially or wholly ignored in closed world systems. In open world systems, data are designed to be accessed by potentially any person or machine, and it is then that problems of proof and trust move to the forefront. OpenAGRIS as a system designed to access such data has to cope with these issues and has been an ideal vehicle to get to the top proof/trust layers of the LOD stack to experience the issues firsthand.

Provenance in particular is a broad term that may refer to various levels of granularity. Data provenance was not a historical concern as AGRIS always took for granted that the agreement with national governments (and the national libraries) provided AGRIS with a license allowing its secretariat to ingest and disseminate data without specific rights or provenance statements. With the shift to digital publishing and machine-readable records however, tracing the provenance chain gained new importance. Dublin Core defines some properties to describe provenance but they are not sufficient in and of themselves to cover all provenance levels, while the W3C Provenance Working Group (W3C, 2011) is working to define an updated and comprehensive data model. In AGRIS, we can define at least the following provenance levels:

- main metadata, such as the title, authors, institutions and publication date;
- metadata that relates to the information of publishers and license, if any;
- the entire record submitted in AGRIS;
- the entity that submitted the record and who is not necessary the entity that created the record, so provenance may refer also to the main source;
- provenance of each triple, especially when there is an enrichment of metadata by accessing other data sources.

In AGRIS, each record has an identifier called an ARN (AGRIS record number), which has a predefined structure and contains implicit information about the AGRIS centre that submitted the record together with the submission year (that is not the year of publication of the resource described by the record). In the ARN, AGRIS is not providing the entire provenance information, but it is able to, at least, determine precise and updated statements of the origins of the artifact itself, i.e. the entity that submitted the record in the AGRIS database. For instance, the ARN "ES2011001090" represents a record submitted in 2011 from the AGRIS center in Spain, whose progressive number is 1090. Especially for legacy data, ARNs are very important pieces of meta-information since it is very difficult to retrieve provenance information for decades-old records with poor metadata. Thus, as immediate work, the team will triplify information about AGRIS centres and other sources of AGRIS records, providing unique URIs for each and adding triples to identify this aspect of provenance which is today implicit in the ARN. More ambitiously the team plans to move on to the institutions and even the authors of each record.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952136043

In fact, provenance must be considered for each piece of metadata and for each triple since AGRIS knowledge has increased during the RDF conversion of the database, e.g. data such as the unambiguous link to the journal of a publication has been added. Thus, some triples may have a different provenance than the record, thus the previous solution is no longer sufficient as it causes a loss of granularity.

Nevertheless, questions remain on the appropriate set of properties with which to encode such metadata. Vocabulary of Interlinked Datasets (VoiD) is a likely candidate. Intended as a bridge between publishers and consumers of RDF data (Cyganiak et al., 2011), it is organized around four metadata areas: descriptive metadata, access metadata, structural metadata and *linking metadata* that is "helpful for understanding how multiple datasets are related and can be used together" (Alexander et al., 2011). For descriptive metadata, VoiD recommends Dublin Core (DC) and Friend of a Friend properties which together cover basic provenance issues. However to more fully cover provenance, extensions such as the EnAKTing Group's voidp Vocabulary for Data and Dataset Provenance (Omitola et al., 2011) are desirable while earlier initiatives such as the Open Provenance Model (Moreau et al., 2010) suffer from complex serializations and no reuse of existing properties.

Looking at metadata domains other than provenance, VoiD also covers license issues by using DC extensions and contains some recommendations regarding common license types. Where VoiD and its extensions are silent is in the trust layer area of quality which will become an important issue as more competing Linked Data comes online and machines need to make dynamic value judgments on which sources to prefer. Though beyond the scope of proposed work in AGRIS, an interesting initiative which bears future examination is Olaf Hartig's tRDF, which "proposes a set of criteria to assess the quality of Linked Data sources" (Hartig et al., 2010).

References

Alexander, Keith. Richard Cyganiak, Michael Hausenblas and Jun Zhao. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved March 20, 2012, from http://www.w3.org/TR/void/ .

Berners-Lee, Tim. (2000). Semantic Web on XML. XML 2000, Washington DC. Retrieved March 09, 2012, from http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html .

Bizer, Christian. Tom Heath, and Tim Berners-Lee. (2009). Linked Data - The Story So Far. In Tom Heath, Martin Hepp, Christian Bizer (Eds.), Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 5(3).

Celli, F. Stefano Anibaldi, Maria Folch, Yves Jaques, and Johannes Keizer, (2011). OpenAGRIS: using bibliographical data for linking into the agricultural knowledge web. AOS 2011.

Cyganiak, Richard. Jun Zhao, Keith Alexander and Michael Hausenblas. (2011). Vocabulary of Interlinked Datasets (VoID). Retrieved March 22, 2012, from http://vocab.deri.ie/void .

Hartig, Olaf. (2010). Quality Criteria for Linked Data sources. Retrieved March 22, 2012, from http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources .

Moreau, Luc. Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, and Jan Van den Bussche. (2010). The Open Provenance Model core specification (v1.1). In Future Generation Computer Systems, July 2010.

Omitola, Tope. Christopher Gutteridge. (2011). voidp: A Vocabulary for Data and Dataset Provenance. Retrieved March 22, 2012, from http://www.enakting.org/provenance/voidp/ .

RDF Working Group. (2004). Resource Description Framework. Retrieved March 23, 2012, from http://www.w3.org/RDF/ .

Tonkin, Emma. Julie Allinson. (2006). Signed metadata: method and application. In Proceedings of the 2006 international conference on Dublin Core and Metadata Applications: metadata for knowledge and learning

W3C Provenance Working Group. (2011). Retrieved March 23, 2012, from http://www.w3.org/2011/prov/wiki/Main_Page.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952136043