

# AGRIS—From a Bibliographic Database to a Semantic Data Service on Agricultural Research Information

Angela Fogarolli, Johannes Keizer, Stefano Anibaldi, and Dan Brickley

*NOTE:* Originally presented at the 13th IAALD World Congress held in Montpellier, 26–29 April 2010.

**ABSTRACT:** AGRIS has for many years provided a huge collection of bibliographic references, such as research papers, studies and thesis, each including metadata such as conferences, researchers, publishers, institutions, and keywords from different thesauri as AGROVOC.

With the rise of full text search and online availability of more research material, the role for bibliographic metadata can appear redundant. When considered instead as a form of modeling that emphasizes relationships, connections and links, bibliographic metadata grows in value as the Web grows in connectivity, and can provide researchers with a map of the global research community, linking formal outputs (papers, data) with a wider grey literature (preprints, drafts) and with communication platforms (blogs, forums) that help researchers put formal findings into a wider context.

This paper describes the evolving role of the AGRIS bibliographic database as it becomes a hub of agricultural research literature. The huge silo of 3 million agricultural resources, collected by more than 150 institutions over the last 35 years, becomes the starting point to access the diverse knowledge in agricultural science and technology available globally on the Web.

**RESUMÉ:** Depuis des années AGRIS offre une énorme collection de références bibliographiques, notamment des documents de recherche, des études et des thèses, toutes accompagnées de métadonnées, telles que conférences, chercheurs, éditeurs, institutions, et de mots-clés provenant de différents thésaurus comme AGROVOC.

Avec l'accroissement de la recherche en texte intégral et la disponibilité en ligne de documents de recherche de plus en plus nombreux, le rôle des métadonnées bibliographiques peut paraître superflu. En revanche si on les considère comme une forme de modélisation qui met en évidence les relations, les connexions et les liens, leur valeur augmente en même temps que la connectabilité sur le Web, et elles peuvent offrir une carte de la communauté mondiale des chercheurs, établissant le lien entre

## Introduction

Through exploration of the evolving role of databases such as AGRIS, it has become clear that the connectivity patterns among the things described in such a database (researchers, topics, institutes, places) can be better reflected online through a more explicit representation both in Web metadata and in user-facing Web sites. The distributed nature of the world described by AGRIS naturally fits a 'linked data' deployment model, in which AGRIS becomes more than a document discovery portal—it

les produits conventionnels (documents, données) et une littérature grise plus abondante (publications préliminaires, projets de textes) et des plateformes de communication (blogues, forums), qui aident les chercheurs à présenter des résultats officiels dans un contexte plus large.

Le présent document cherche à décrire le rôle en pleine évolution de la base de données bibliographiques AGRIS qui devient un centre de documentation scientifique agricole. Le réservoir gigantesque de 3 millions de sources d'informations agricoles, rassemblées par plus de 150 institutions depuis 35 ans, devient le point d'entrée pour accéder à la diversité des connaissances dans le domaine des sciences et technologies agricoles qui sont disponibles à l'échelle mondiale sur le Web.

**RESUMEN:** AGRIS dispone desde hace varios años de una extensa colección de referencias bibliográficas, como por ejemplo artículos de investigación, estudios y tesis, que contienen metadatos tales como conferencias, investigadores, editoriales, instituciones y palabras clave extraídas de varios tesauros, entre ellos el Tesauro AGROVOC.

El uso cada vez mayor de búsquedas de texto completo y la creciente disponibilidad de material de estudio en línea ha llevado a que los metadatos bibliográficos sean considerados cada vez más redundantes. Sin embargo, si estos metadatos bibliográficos son vistos como una forma de modelación que destaca las relaciones, conexiones y enlaces, el valor de éstos aumenta a medida que crece la conectividad Web, ofreciendo a los investigadores un mapa de la comunidad de investigación mundial, enlazando los productos formales (artículos, datos) con la literatura gris más amplia (tiradas preliminares, borradores) y con las plataformas de comunicaciones (blogs, foros) que permiten a los investigadores situar los resultados formales dentro de un contexto más amplio.

Este artículo busca describir la evolución y función de la base de datos bibliográfica AGRIS y cómo se ha convertido en el eje de la producción bibliográfica en materia de investigación agraria. El enorme silo de tres millones de recursos agrarios, reunidos por más de 150 instituciones a lo largo de los últimos 35 años, es el punto de acceso a múltiples conocimientos agrarios científicos y tecnológicos disponibles mundialmente a través de la Web.

becomes an entry point and map of the entire research landscape around some topic or theme.

Linked-data (Bouquet et al., 2008) techniques foster links between resources through the Web. This approach requires an emphasis on sharing identifiers, names and descriptions of real-world and abstract objects other than the bibliographic materials themselves: conferences, workshops, research centres, researchers, subject themes, homepages.

None of this is news to the bibliographic professional: Such concerns have been at the heart of metadata work for

years. What is new today is the presence of tools (standards, software) and community trends (open linked data, open archives, RSS/Atom syndication) that allow the full potential of such link-oriented metadata to be exploited.

## An entity centric approach to data aggregation

The idea to shift the web from a huge graph of documents to a huge graph of data has become more and more popular since the days when Tim Berners-Lee proposed the idea of the Semantic Web. Since then, scientists and practitioners have invested a lot of effort to realize this vision, often trying to adapt and reuse models and techniques originating in more traditional areas like databases and artificial intelligence. However, there is a very important difference between traditional knowledge-based systems, and current work that aims to reach semantic computing at web scale: The notion of global interlinking of distributed pieces of knowledge.

At the base of such interlinking—and the resulting semantic interoperability of fragments of data—is the notion of identity of and reference to entities. Systems that manage information about entities (objects/individuals/instances...) commonly issue identifiers for these entities, in the same way that relational databases issue primary keys for records. If these identifiers are generated by the information systems themselves, several issues arise that hinder interoperability and integration: (i) it can lead to a proliferation of identifiers because the same object is potentially issued with a new identifier in several information systems; (ii) one-to-one relations between identifiers and an entity cannot be achieved, i.e. one identifier can denote different entities in different information systems; (iii) reference to entities across information systems is very complicated or impossible because there is no way to know how an entity is identified in another system.

To overcome this lack of data-level integration, OKKAM (Heath et al., 2009) proposes a global, public infrastructure, called Entity Name System (ENS), which fosters the systematic creation and reuse of identifiers for entities in the global space of information and knowledge. This a-priori approach enables systems to reference the entities which they describe with a globally unique identifier, and thus create pieces of information that are semantically pre-aligned around those entities. Semantic search engines or integration systems are then able to aggregate information from distributed systems around entities in a precise and correct way. We call this the ‘entity-centric approach’ to semantic interoperability, and the resulting information/knowledge space is called the ‘Web of Entities.’

### The OKKAMization Process

‘OKKAMization’ is the process necessary to include an entity in existing information sources in the total web of entities. It involves the identification of entities inside

existing repositories and the creation of unique identifiers (OKKAM ID) for entities which are not already present in the ENS system.

Creating an OKKAM ID for an object involves getting a unique identifier which is a non-ambiguous way to refer to that object without ambiguity. An OKKAM ID is a well formed URI which enables a resource to be semantically connected to other global resources.

To allow the correct creation of an OKKAM identifier, it is necessary to collect a minimal set of information about an entity. This minimizes the risk of ambiguities (imagine creating an OKKAM ID about Mr. John Smith, just using his name. The result will be an OKKAM ID that refers to a person, but OKKAM will not be able to identify uniquely this entity because there are many ‘John Smiths’ in the world. Building an OKKAM ID with more information, such as state, city, work, allows the system to better recognize the right ‘John Smith.’)

The OKKAMization process of the AGRIS repository comprises four phases:

1. **Corpus entity recognition.** This focuses on entity recognition inside the AGRIS repository and related sources.
2. **Association of OKKAM IDs to extracted entities.** This task matches extracted entities against the OKKAM ENS. If a match can be found for an entity, then the identifier is reused. Otherwise a new entity profile is created and thus a new unique identifier is created for the extracted entity.
3. **Enrichment of the AGRIS repository with OKKAM identifiers.** The OKKAM identifier generated in phase two are included in the XML files of the repository as another type of metadata. This allows automatic identification and aggregation of entities inside the repository. The objective of this phase is to enable entity-based retrieval and to semantically connect entities in different contexts. Thus, from the user point of view, this will translate in an efficient retrieval that avoids information overload.
4. **RDF enrichment of the AGRIS repository.** During this step, the AGRIS repository is described using RDF notation. Publishing the repository using RDF makes the content of the repository understandable by external semantic search engines (SIG.MA, Google Project...). The hidden semantic connections among entities can be discovered and displayed to the users. Entities form the AGRIS repository can be described in RDF or microformats in other web resources such as the FAO website, and this will increase the semantic information that can be aggregated for the same entity.

### The AGRIS Linked-data Model

The AGRIS repository is a large and rich collection of bibliographic references encoded in a qualified Dublin Code XML format. Each XML document is structured in a metadata description for a publication which is sometimes available in a PDF format.

FIGURE 1 – Semantic Search by OKKAM id for a journal

**SIG.MA**  
SEMANTIC INFORMATION MASHUP

http://www.okkam.org/ens/id01dff3a2-cf11-4471-905d-18f9403493a2 Add More Info Start New

**artifact name:** Savremena poljoprivredna tehnika (Serbia), Contemporary Agricultural Engineering [?]

**country:** SERBIA [?]

**citation [2006]:** [?]

**chronology:** [2005] [?]

**citation identifier:** 0350-2953 [?]

**is citation of:** show 31 values [?]

**identifier:** www.okkam.org/ens/id01dff3a2-cf11-4471-905d-18f9403493a2

**is citation of:** Obrada zemljišta kombinovanim sredstvima [?]  
 Efikasnost rada pneumatske rejalice, pri setvi oronog luka [?]  
 The results of parallel testing of soy bean seed drill in strips and row [?]  
 Otpornost rasplivkača na petrosnju i značaj njihovog testiranja [?]  
 Contemporary combines in conditions of Vojvodina [Serbia] [?]  
 Mogućnost za identifikaciju korena secerne repe u zemljištu pomoću radarskog senzora [?]  
 Optimization of transport in harvesting [?]  
 Universal frame with elements for autumn-winter tillage of lasting planted [?]  
 Estimation of necessity heat energy quantum for plastic house heating of 0.5 ha in real conditions [?]  
 Test results for a 250 kW biomass energy boiler and suggested technical and organizational measures boiler installations [?]  
 Štopljanje, transport i skladištenje hala slame [?]  
 Ušteda energije kod panelnog sistema za grejanje prasadi [?]  
 Mikroklimatski automatski kontrolni sistem za povećanje energetske i ekološke efikasnosti objekata [?]  
 Mikroprocesorski termo-regulacioni sistem za održavanje temperature [?]  
 Geometrijski pokazatelji rada rotacione sibilnice sa suprotnosmernim obrtanjem radnih organa [?]  
 Neke mehaniske metode zaštite bilja u kontrolisanim uslovima za gajenje [?]  
 Technological and technical aspects of young potato production [?]

FIGURE 2 – AGRIS data linked to other web resources

Optimization of transport in harvesting Add More Info Start New Order Options Use It

**Optimization of transport in harvesting**

**comment:** From theoretical point of view for determination of optimal harvest model, for transport and storage theory of mass serving closed type with complex analysis among all of three manifolds is applied. In the beginning harvest and serving of carriers by combine harvester are being observed. Because of comparison of simulated parameters, software for simulation of problems is made on PC computer. [?]

**availability:** COBISS.SR-ID 5117698 [?]

**number:** [?]

**location:** RS [?]

**arn:** RS2007000008 [?]

**creator personal:** Turan, J. [Poljoprivredni fakultet, Novi Sad (Serbia)] [?]

**citation number:** v. 31(3) p. 136-142 [?]

**creator:** Simpozijum Poljoprivredna tehnika, 31, Zlatibor (Serbia), 22-29 Jan 2005 [?]

**conference:** [?]

**citation:** http://www.okkam.org/ens/id01dff3a2-cf11-4471-905d-18f9403493a2 [?]

**description notes:** Summaries (En, Sr) [?]  
 4 ref. [?]  
 2 graphs [?]  
 1 table [?]  
 1 ill. [?]

**date issued:** hide value just this value which sources reject sources [?]

**identifier:** agris.fao.org/agris-search/search/display.do?f=2007/RS/RS0701.xml;RS2007000008 [?]

**Sources (7) Approved (0) Rejected (0)**

- Untitled document 39 facts | 2010-02-03  
http://okkam2.din.com.rs:8080/Agris/RS0701.rdf
- Produktschap Tuinbouw - ... 1 facts | 2010-02-19  
http://www.narcis.info/organisation/Rece...
- Appropedia talk:Village ... 1 facts | 2010-01-30  
http://www.appropedia.org/Appropedia\_talk...
- Untitled document 1 facts | 2010-02-19  
http://www.acribd.com/doc/2364844/Comm...
- Untitled document 1 facts | 2010-02-19  
http://groups.yahoo.com/group/NTPD-EP/line...
- Alinate Bachelor thesis 1 facts | 2010-02-19  
http://www.acribd.com/doc/14933792/Alpina...
- Untitled document 1 facts | 2010-02-19  
http://groups.yahoo.com/group/barita-bh...

**AGRIS record**

**Record number:** RS2007000008

**Titles:** Optimization of transport in harvesting

**Original title:** Optimizacija transporta u zetvi

**Personal Authors:** Turan, J. (Poljoprivredni fakultet, Novi Sad (Serbia))

**Conference:** Simpozijum Poljoprivredna tehnika, 31, Zlatibor (Serbia), 22-29 Jan 2005

**Publication Date:** [2005]

**AGRIS Subj. Cat. AGRIS Terms:** Handling, transport, storage and protection of plant products; Agricultural machinery and equipment; HARVESTING, COMBINE HARVESTERS, TRANSPORT, COSTS, COMPUTER APPLICATIONS, RECOLTE, MOISSONNEUSE BATTEUSE, TRILLADORAS, TRANSPORTS, COSTOS, APLICACIONES DEL ORDENADOR

**Language notes:** Summaries (En, Sr), 1 ill., 2 graphs, 1 table, 4 ref.

**English Abstract:** From theoretical point of view for determination of optimal harvest model, for transport and storage theory of mass serving closed type harvest and serving of carriers by combine harvester are being observed. Because of comparison of simulated parameters, software for simulation of problems is made on PC computer.

**Type:** Conference, Summary

**Availability:** NSL:RS (Serbia)

**Availability number:** COBISS.SR-ID 5117698

**Serial Title:** Savremena poljoprivredna tehnika (Serbia); Contemporary Agricultural Engineering

**Serial ID - ISSN:** 0350-2953

**Serial number:** v. 31(3) p. 136-142

**Serial Date:** [2005]

**For document availability, please contact:**  
 NSL:RS (Serbia)  
 Adulice vrsake 2  
 21000 Novi Sad, Serbia  
**Contact:** Zorica VESOVIC  
 Tel: +381-21-525 859  
 Fax: +381-21-525-859  
 Email: zorica@insr.na.ac.rs

**Community's Contribution to Develop Sustainable Future**

Download Print Full Screen Share

**Scribd** Explore Community Upload Search Books, Presentations, Business, Academic

**The Book Depository** Big discounts on millions of books with free international shipping  
www.BookDepository.co.uk

**Most Executive Level Jobs** The No. 1 Site for Executives Access over 5 000 Headhunters  
www.Executive.com

**Looking For A Publisher?** Get published in just 30 days your rights. Earn 50% royalty  
www.AuthorHouse.co.uk

In this section, we report about a first experience enabling linked data in AGRIS using the OKKAM ENS infrastructure.

To create a linked-data model for AGRIS, we followed the OKKAMization steps described above. We initially focused on the journal entity type in order to show the immediate advantage of applying a linked-data model to the AGRIS corpus. Second, through the OKKAM ENS search API, we obtained unique identifiers for each journal. Third, we show how the unique identifiers are introduced in the original repository files and then how each file is translated in RDF format and submitted to a Semantic Web Search Engine.

Assigning unique identifiers to entities in the AGRIS repository leads to a light-weight data integration of entities and in this way enables inter-linkage among entities which can come from different information sources (see Figures 1 and 2). The result is that efficient information retrieval will be enabled within the AGRIS repository and globally by interlinking with other information sources.

Figure 1 shows the result of a search by unique identifier for a journal. The interface shows different statements about the journal resource; some of the attributes are clickable to enter a deeper level of detail.

In the example, clicking on the citation attribute displays all the article titles for that journal. Clicking on one of the article titles displays the details of the article (see Figure 2).

Information about the OKKAMized resource can be aggregated through different sources. These are displayed on the right side of the figure — a click on an attribute allows the user to explore more details of the attribute itself. If the value of an attribute is a URL, this can connect with external information sources. In Figure 2, the article details provide many ways to navigate the AGRIS website or to other external related sources.

The volume of inter-linkages with external resources grows with the use of the same unique identifier across the web.

As mentioned before, references in the AGRIS repository are encoded in a XML format. This type of file can be enriched with unique identifiers, allowing the future representation of the unique identifier on the AGRIS web page and to the wider Web of entities.

Below we present a snapshot of the XML of an AGRIS resource with an OKKAM unique identifier for the journal in which the article appears:

```
<ags:citation>
  <ags:citationTitle>Savremena poljoprivredna tehnika Serbia)
  </ags:citationTitle>
  <ags:citationTitle>Contemporary agricultural engineering
  </ags:citationTitle>
  <ags:citationIdentifier scheme="ags:ISSN">0350-2953
  </ags:citationIdentifier>
  <ags:citationNumber>v. 31(1-2) p. 29-37</ags:citationNumber>
  <ags:citationChronology>(2005)</ags:citationChronology>
  <OkkamID>http://www.okkam.org/ens/ida53b7142-5880-4684-aab3-f83c2a6d0ea8/</OkkamID>
</ags:citation>
```

Below is an example of the automatically-generated RDF file for an AGRIS resource article. The unique identifier for the journal is highlighted in bold. The journal attributes are described inside the *rdf:Description* tag for the resource with a specific unique identifier (*.rdf:about="okkam\_id value"*).

Whenever the journal is cited within the AGRIS record, the unique identifier is used to describe it. The link to the original AGRIS website is also displayed. This will connect through the AGRIS web search interface from any Semantic Web search engines.

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:j.0="http://models.okkam.org/ENS-core-vocabulary.owl#"
  xmlns:j.1="http://purl.org/dc/terms/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:j.2="http://purl.org/agmes/1.1/">
  <rdf:Description rdf:about="http://www.okkam.org/ens/id01dff3a2-cf11-4471-905d-18f9d03d93c7">
    <j.2:citationChronology>(2006)</j.2:citationChronology>
    <j.0:artifact_name>Savremena poljoprivredna tehnika (Serbia),
      Contemporary Agricultural Engineering</j.0:artifact_name>
    <j.2:citationChronology>(2005)</j.2:citationChronology>
    <j.2:citationIdentifier>0350-2953</j.2:citationIdentifier>
    <j.0:country>SERBIA</j.0:country>
  </rdf:Description>
  <rdf:Description rdf:about="http://agris.fao.org/agris-search/search/display.do?f=2007/RS/RS0701.xml;RS2007000023">
    <j.2:subjectThesaurus>PROPIEADADES TECNICAS
    </j.2:subjectThesaurus>
    <j.2:subjectThesaurus>TRITICUM</j.2:subjectThesaurus>
    <j.2:subjectThesaurus>http://www.fao.org/aos/agrovoc#c\_2572
    </j.2:subjectThesaurus>
    <j.2:subjectThesaurus>http://www.fao.org/aos/agrovoc#c\_2386
    </j.2:subjectThesaurus>
    <j.2:creatorConference>Simpozijum Poljoprivredna tehnika, 32,
      Zlatibor (Serbia), 28 Jan-4 Feb 2006</j.2:creatorConference>
    <j.1:abstract>The paper shows presentation of the exploitational
      examination results for the wheat drill sowing aggregates.
      Some technical-technological drill solutions and the results of
      the working quality (norm, drilling depth) and the exploitational
      parameters (working speed, output) have been shown.
    </j.1:abstract>
    <j.2:creatorPersonal>Mehandzic, R.(Poljoprivredni fakultet, Novi
      Sad (Serbia). Departman za poljoprivrednu tehniku)
    </j.2:creatorPersonal>
    <j.2:creatorPersonal>Malinovic, N.(Poljoprivredni fakultet, Novi
      Sad (Serbia). Departman za poljoprivrednu tehniku)
    </j.2:creatorPersonal>
    <j.2:subjectThesaurus>SEMIS EN LIGNE</j.2:subjectThesaurus>
    <dc:type>K</dc:type>
    <j.2:subjectThesaurus>SEMOIR</j.2:subjectThesaurus>
    <j.2:ARN>RS2007000023</j.2:ARN>
    <j.2:descriptionNotes>3 tables</j.2:descriptionNotes>
    <j.2:subjectThesaurus>TECHNICAL PROPERTIES
    </j.2:subjectThesaurus>
    <j.2:descriptionNotes>2 ref</j.2:descriptionNotes>
  <j.2:citation rdf:resource="http://www.okkam.org/ens/id01dff3a2-cf11-4471-905d-18f9d03d93c7"/>
```

## Conclusions

In the AGRIS 2010 work, we have been prototyping a redesign for AGRIS that brings these concerns to the core of the system. Both in our data modelling, and in the Web presence, AGRIS will better emphasise the networked, linked nature of the things it describes.

AGRIS has for many years provided a huge database of bibliographic references, such as research papers and theses, each including metadata such as conferences, researchers, institutions, and keywords from different thesauri as AGROVOC.

The solutions presented allow information retrieval systems to perform stronger automatic elaboration offering data identification and aggregation. OKKAM allows the AGRIS repository to acquire the added value of making its full content available to the global web and at the same time to combine and aggregate information between and outside organizational boundaries.

It offers an effective and innovative solution to diffuse global knowledge through semantic web technologies.

## References

Bouquet, P., H. Stoermer, J.C. Niederee, and A. Mana. 2008. Entity name system: The back-bone of an open and scalable web of data. *International Conference on Semantic Computing*, 0:554–561.

Brickley, D., S. Anibaldi, A. Picarella, and J. Keizer. 2009. Designing AGRIS 2010—Information linking and Agricultural Research. Poster at Dublin Core Meeting 2009. Rome: FAO. <ftp.fao.org/docrep/fao/012/ak689e/ak689e00.pdf>

Heath, T., M. Hepp, and C. Bizer. 2009. Special Issue on Linked Data. *International Journal on Semantic Web and Information Systems*, 1–22.

## About the authors

ANGELA FOGAROLLI is a Post-doctoral researcher in the context of the OKKAM project at the University of Trento. Her research interests are located in the multidisciplinary area between intelligent multimedia retrieval, Semantic Web and Data Integration. Since 2002 she has been working as a researcher and lecturer as well as forming part of the Information Systems Innovation Unit of the Health Care Service of the Province of Trento. — [afogarol@disi.unitn.it](mailto:afogarol@disi.unitn.it)

JOHANNES KEIZER is team leader ‘Knowledge Standards’ within the Office of Knowledge Exchange, Research and Extension at FAO. His team is facilitating initiatives to set common standards and methodologies for Agricultural Information and Knowledge Management. Johannes has a PhD in Biology and was working in biochemical research before joining FAO. — [johannes.keizer@fao.org](mailto:johannes.keizer@fao.org)

STEFANO ANIBALDI is responsible for the management and dissemination of the AGRIS Repository, whose data is collected from dispersed agricultural institutions world-wide. He works in the Office of Knowledge Exchange, Research and Extension at FAO; he holds a degree in Foreign Languages and Literature and a BSc degree in Information Science. — [stefano.anibaldi@fao.org](mailto:stefano.anibaldi@fao.org)

DAN BRICKLEY is best known for his work on Web standards in the W3C community, where he helped create the Semantic Web project and many of its defining technologies. Dan is currently working at the Free University of Amsterdam on the NoTube EU project, developing new approaches to interactive TV that build upon SKOS, FOAF and open Web technologies. — [dan.brickley@fao.org](mailto:dan.brickley@fao.org)