A Distributed Architecture for Harvesting Metadata Describing Organizations in the Agriculture Sector

Valeria Pesce[1], Ajit Maru[1], Gauri Salokhe[2], Johannes Keizer[2]
[1] Global Forum on Agricultural Research, c/o Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla, 00153 Rome, Italy
{Valeria.Pesce, Ajit.Maru}@fao.org
[2] Knowledge Exchange & Capacity Building Division (KCE), Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla, 00153 Rome, Italy
{Gauri.Salokhe, Johannes.Keizer}@fao.org

Abstract. Providing easy access to updated, accurate and semantically meaningful information about organizations working in the agriculture sector is of primary importance in agricultural information management. Many databases of these organizations already exist but none are comprehensive and all differ in coverage (often overlapping), semantic organization, being up-to-date, quantity and quality of the information they provide. In addition, only very few information systems share and exchange data among themselves. This paper describes a distributed architecture which minimizes duplication in information storage and flow and improves quality of the information provided. In this architecture the data describing an organization are stored in a file as an XML description based on a specific metadata set, and access to these distributed files is facilitated by a central registry file. The proposed metadata set is also discussed, with special focus on those aspects that help to make the architecture coherent.

Keywords: organization metadata, XML, agriculture, information management, agricultural information systems, standards

1 Background

In 2005, the Expert Consultation on International Information Systems for Agricultural Science and Technology [1] was held in Rome with the goal of developing coherence in agricultural information systems. The Consultation set in place three taskforces that should carry on activities related to specific areas, one of which was identified for content management.

The activities of the Content Management Taskforce (CMTF) relate to information sharing, and in this context three of the participating organizations, the Food and Agriculture Organization of the United Nations (FAO), the Global Forum on Agricultural Research (GFAR) and Wageningen International, have carried on activities for developing a metadata set for a specific information object: organizations working in the area of agriculture.

The objective for designing a new metadata set was that of facilitating information exchange between the different information services that manage data about organizations.

An Application Profile (AP), or a set of standard terms, for describing organizations working in the agriculture sector (the Organizations Application Profile[1] or OAP) was presented by FAO and GFAR at the Content Management Taskforce meeting in Wageningen (March 2007) [2] and has been available for comments since then on the Agricultural Information Management Standards Web site[2] (AIMS).

During the same Taskforce meeting, GFAR and FAO proposed a special use case for the AP, with the objective of streamlining the management and flow of information on organizations, thus minimizing the duplication of data, work and costs. In this proposed use case, each organization describes itself using the AP, stores the XML/RDF description on a server and registers the URL of the description with a publicly accessible central Registry file so that information services can harvest the descriptions by accessing the Registry.

There was a general consensus on creating a project paper detailing the proposal. The document is available[3] for comments on EGFAR[4] Web site.

1.1 Rationale

Agricultural sciences and technology boast a large number of organizations, both in the developing and developed countries, mainly because agriculture is the primary industry in nearly all the countries. The demand for quality information services on "who is doing what" and "who is operating in which areas" is high. Many organizations (research institutes, research networks, all the organizations managing projects) are interested in information about funding agencies, farmers' organizations are interested in information about research institutes and government institutions, civil society organizations are interested in linkages with other organizations etc.

Information on agricultural organizations is now managed by many different information services[5] with independent databases. These systems differ considerably in coverage (often overlapping) and can therefore only provide partial answers. Besides, no cross-searches are possible as very few of these systems share data among themselves, very few export data in some agreed exchange format – so far, no metadata set has been promoted as a standard - and very few adopt (or map to) common controlled vocabularies for subject classification.

---

[1] Organizations Document Type Definition: http://www.purl.org/agmes/organizationap/dtd/

[2] Agricultural Information Management Standards Web site: http://www.fao.org/aims

[3] Proposed Architecture and Workflow for Managing Decentralized Information on Organizations: http://www.egfar.org/egfar/website/opensite/collabwebsite?contentId=1599

[4] The Website of the Global Forum on Agricultural Research Web site: http://www.egfar.org

[5] Some examples of databases with global coverage are: AROW http://www.isnar.cgiar.org/arow/index, InfoSys+ http://www.infosysplus.org/, WISARD http://www.wisard.org/, the FAO NARS db http://www.fao.org/sd/researchinstitutions

There is also the demand for visibility on the part of the organizations themselves: they have to submit their own data to the major information systems, who somehow become the owners of the information.

This situation translates into three major difficulties:

1. **on the part of the information services and the data owners**: maintenance is costly (and not cost-effective, since many similar datasets are maintained and updated by different services) and there are no standard exchange formats that allow to tap into external sources;
2. **on the part of the users**: similar information is available from several sources, none of which is comprehensive and all of which differ in: subject coverage, type coverage, semantic organization, quantity and quality of information; selecting only one source is limiting and no cross-searches are possible;
3. **on the part of the single organizations**: visibility in all the existing databases requires submitting the same data about the organization to multiple databases.

1.2 Approach

While the adoption of a common metadata set would facilitate information sharing and allow information systems to access data from each other, a distributed architecture where data are only managed by the owners and harvested by the information services would also minimize the problems related to maintenance.

The approach followed in this project is based on the design of both a metadata set and a distributed architecture.

Agreement on standards and technologies for easier interoperability. Powerful information services can be easily built if all the available data sources manage, or at least expose, data in the same way. Using the same metadata set and protocols allows to expose data in the same way; going even further, using (or mapping to) the same subject classification can make data all the more interoperable. This is why the agricultural community should agree on standards and common procedures in data management and exchange.

Some standards already exist for facilitating information exchange and thus have improved interoperability in the agricultural field. These are documented on the AIMS Web site which aims at coordinating and harmonizing "the decentralized efforts currently taking place in the development of methodologies, standards and applications for management of agricultural information systems". The proposed architecture adopts the OAP developed by FAO, which takes into account the needs of a distributed architecture. The AP also provides metadata elements and refinements for subject classification and the usage of controlled vocabularies.

Distributed storage for easier maintenance and greater reliability. Data are most easily managed (stored, updated and made accessible) where they were originally created and can be easily updated and where they will be used to deliver

desired services that meet local needs. In the case of organizations, the proposed architecture establishes that the data about an organization should be stored and managed by the organization itself and stored and maintained in one single place.

2 Proposed Architecture and Workflow

The proposed architecture is based on a central Registry File which stores the locations of the distributed organization metadata files.

The elements of the proposed architecture are:

- **Data Providers.** These are the organizations themselves: each organization should describe itself and the description should be in the form of an XML/RDF record (compliant with the OAP) stored in a file. If the organization has the capacities, it can create the XML/RDF record, validate it against the prescribed Document Type Definition (DTD), store it on a web server and register the URL of the file with the central Registry. Otherwise it can use the services of a gateway provider (see below). The description is always maintained and, if necessary, edited by the organization itself.

- **Registry.** Through a simple web application, URLs would be appended/updated to a central Registry File. The data in the Registry would be stored in XML/RDF format, so as to allow ease of access and reuse. For each organization, there will only two pieces of information stored: an identifier (discussions are still going on about the assignment of unique identifiers) and the URL for retrieving the XML/RDF description. An additional element could be a reference email, in order to avoid spamming and notify the organization in case of unavailability of the source.

- **Gateway Providers.** Organizations that can provide facilities for other organizations, like: a) a web tool for creating the XML/RDF record; b) web hosting for the XML/RDF file with related tools for updating the record and registering it with the Registry.

- **Service Providers (harvesters).** All the organizations/services that want to provide information services based on the descriptions of the organizations. Service providers would access the Registry file, either directly or via web services, and harvest all the URLs. With this information, they can then individually access the metadata records, read the data that they need and create desired value-added services. Since the XML/RDF descriptions, at each URL, are based on the OAP and created by the owners themselves, they will allow for the implementation of quality information services.

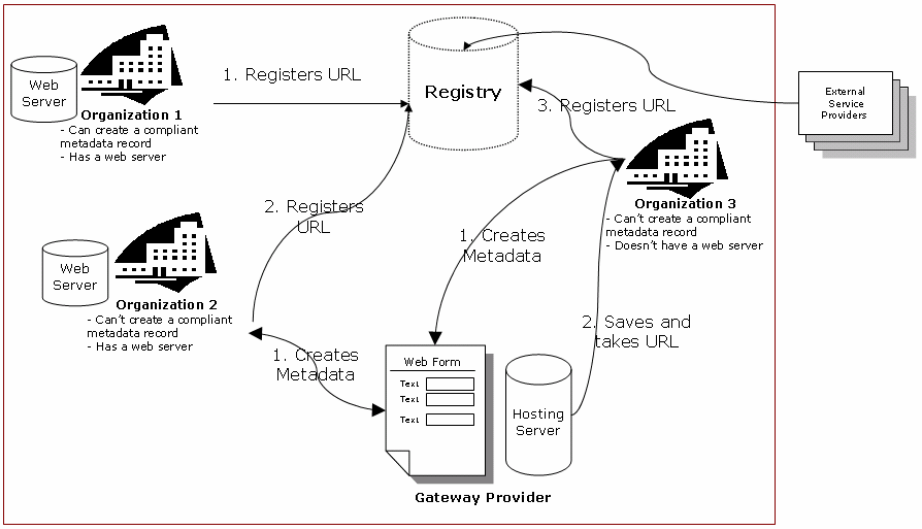

Fig. 2. Architecture and workflow.

3 The Organizations Application Profile

Studies, on the available systems that contain organization information, indicate that most of them have been created to meet their individual needs[6]. To allow standardized description of the organizations, there was a need to define a standard set of metadata elements. Metadata about an organization are a means to help identify regional, national and international organizations specializing in different agriculture-related domains. The AP does not aim to be an all inclusive and comprehensive standard for describing organizations, their units and their personnel. Rather, it limits itself to describing the type of organization and its location information: a business card for an organization.

As in the case of any AP, the OAP reuses metadata elements from existing namespaces, namely: Dublin Core[7] (DC), DC Terms[8] (DCTERMS) and Agricultural

[6] Some of the information systems mentioned earlier in this document use a set of fields for describing organizations, but have not designed a metadata set for general purposes. In a broader field than that of agricultural organizations, the Common European Research Information Format (CERIF) standard was released in 1991 to foster the diffusion of research information across Europe. This standard was created to exchange project information, however the documentation of the updated version (2000) includes other types of information including organizations. The standard is stable but has a complex structure with different information types integrated within. The result is a high number of elements of which most are mandatory. This makes the standard not flexible or easily interoperable with others standards.

[7] http://www.dublincore.org/

[8] http://www.dublincore.org/

Metadata Element Set[9] (AGS). The elements that were evaluated as necessary for describing an organization are as follows:

Table 1. Elements of the Organizations Application Profile.

Proposed Elements	Namespace	Controlled Vocabulary	Requirement 10	Cardinality 11
organizationName	AGS	no	M	R
- fullOrganizationName	AGS	no	M	N-R
- organizationAcronym	AGS	no	O	N-R
address	AGS	no	M	N-R
- streetAddress	AGS	no	M	N-R
- country	AGS	dcterms:ISO3166	M	N-R
telephone	AGS	no	O	R
fax	AGS	no	O	R
telex	AGS	no	O	R
email	AGS	no	O	R
identifier (scheme “dcterms:URI”)	DC	no	M	N-R
description	DC	no	O	R
subject	DC	no	M	R
- subjectThesaurus	AGS	yes [12]	O	R
organizationType	AGS	recommended	M	R
relation	DC	no	O	N-R
- isPartOf (scheme “dcterms:URI”)	DCTERMS	no	O	R
- replaces (scheme “dcterms:URI”)	DCTERMS	no	O	R
date	DC		M	N-R
- created	DCTERMS	dcterms:W3CDTF	M	N-R
- modified	DCTERMS	dcterms:W3CDTF	M	N-R

The elements are fairly self-explanatory and the metadata design is intentionally simple. The entity described by the metadata is the organization, and it was agreed that all elements should refer to the entity itself, except for the dc:date element, which refers to the record.

The design of the metadata was influenced by its foreseen use in a distributed architecture, which is particularly evident in the choices made regarding: a) which elements are mandatory and which are not; b) the usage of ags:subjectThesaurus with scheme refinements for different thesauri; and 3) the usage of dc:identifier and dc:relation to implement relations between entities.

[9] http://www.fao.org/aims/

[10] Mandatory (M) / Optional (O)

[11] Repeatable (R) / Not-Repeatable (N-R)

[12] Scheme refinements are provided for the most widely used thesauri in the agricultural field and in related fields: AGS:AGROVOC, AGS:CABT, AGS:ASFAT, AGS:NALT, DCTERMS:MeSH, DCTERMS:LCSH

3.1 Mandatory and Optional Elements

In a scenario where the organizations are describing themselves, more elements can be made mandatory than it can be done when a metadata set is used as an exchange format. Therefore, only those elements for which an organization might not have a value are optional.

3.2 Semantic Coherence

Considering that different organizations might be familiar with different vocabularies or classifications and that different information services harvesting the records might use different taxonomies or Knowledge Organizations Systems (KOS), a certain range of values for the scheme refinement of ags:subjectThesaurus is provided, in addition to the possibility of using free-text values in dc:subject. In order to promote coherence, the metadata guidelines will encourage the use of AGROVOC[13], the multilingual agricultural thesaurus produced by FAO, and the web interface for creating metadata will provide a browsing option for this, however other vocabularies are also allowed.

The envisaged architecture also offers other means of achieving coherence and integration, as the web tool for creating the metadata and/or the information services harvesting the descriptions can map terms between different vocabularies[14] and offer real added value.

3.3 The dc:identifier Element. Unique Identifiers and Relations

Theoretically, the Registry File could just consist of a list of URLs, with no need for unique identifiers, since the URLs, though not permanent, are unique. Using this approach, the dc:identifier element would be used for the URL of the website and only the URLs of the descriptions would be stored in the Registry file.

However, assigning a unique identifier to an organization allows to: a) change the URL of the record without creating a second entry in the registry file; b) (to a certain extent) avoid duplication; c) performing faster harvesting; d) most important of all, create and maintain relations between the records (impossible with URLs since they can change).

Since it was decided to implement relations between the organizations (and their parts: divisions, departments, branches etc.), the dc:relation element was included with nested DCTERMS elements for describing different types of relations with other records (organizations). Using this approach, one of the elements of the metadata set had to contain a unique and permanent - therefore location-independent - identifier. The dc:identifier element was used for this.

---

[13] AGROVOC Thesaurus: http://www.fao.org/aims/

[14] Of course, mapping between different KOS may be a long and difficult task, but some projects have already started, like the mapping between the NAL Thesaurus and the AGROVOC Thesaurus.

The dc:identifier is often used for the URL where a resource is available, but in this case it had to be persistent, so it was proposed to use Uniform Resource Names (URNs), which "are intended to serve as persistent, location-independent, resource identifiers"[15].

The "DCTERMS:URI" scheme refinement fits the URN syntax perfectly and will be used for both the dc:identifier element and the dc:relation sub-elements.

4 Conclusion

This proposed architecture and the applied metadata set aim at streamlining the management and flow of information on organizations - minimizing the duplication of data, work and costs - and improving quality, coherence and accessibility of the information. The final objective is to offer a coherent framework for the creation of useful and high quality services to the end user. However, it needs to be stressed here that the quality of the services created from this are highly dependent on the quality of metadata provided by the organizations. The central Registry File itself will be a Global Public Good on which everyone can leverage.

References

1. Expert Consultation. International Information Systems for Agricultural Science and Technology - Review of Progress and Prospects. FAO Headquarters Rome 19-21 October 2005 (2005). ftp://ftp.fao.org/gi/gil/consultations/final_report_10-02-06.pdf
2. Meeting Report: Content Management Task Force. International Information Systems for Agricultural Science and Technology. Technical Centre for Agricultural and Rural Cooperation (CTA) Wageningen. The Netherlands 1-2 March, 2007 (2007). ftp://ftp.fao.org/gi/gil/gilws/aims/publications/papers/20070416CMTF-Report.pdf

[15] http://tools.ietf.org/html/rfc2141