INFRA-2011-1.2.2. Data infrastructures for e-Science

A data infrastructure to support agricultural scientific communities
Promoting data sharing and development of trust in agricultural sciences

Review of existing agricultural data management practices, lifecycles and workflows

Deliverable number	*D5.1*
Dissemination level	*Public*
Delivery date	*3rd October 2012*
Status	*Final*
Author(s)	*Hugo Besemer (Food and Agriculture Organization of the United Nations), Philip Edge(Food and Agriculture Organization of the United Nations), Imma Subirats (Food and Agriculture Organization of the United Nations), Effie Tsiflidou (Agro-Know Technologies), Vassilis Protonotarios (Agro-Know Technologies)*

agINFRA is supported by the FP7 - Research Infrastructures, Capacities Programme.

Document Revision History

Version	Date / Contributor	Status	Summary of Changes
0.1	06.09.2012 / H.Besemer (Food and Agriculture Organization of the United Nations), P. Edge (Food and Agriculture Organization of the United Nations), I.Subirats (Food and Agriculture Organization of the United Nations)	Final draft 1	Final draft for review
0.2	26.09.2012 / G. Geser (Salzburg Research)	Final draft 2	Chapter 9
0.3.	26.09.2012 / H.Besemer (Food and Agriculture Organization of the United Nations), P. Edge (Food and Agriculture Organization of the United Nations)	Final draft 3	Revisions according to internal review by Salzburg Research Forschungsgesellschaft
0.4	28.09.2012 / I. Subirats (Food and Agriculture Organization of the United Nations)	Final draft 4	Revisions according to internal review by Salzburg Research Forschungsgesellschaft
0.5	30.10.2012 / E.Tsiflidou (Agro-Know Technologies), V. Protonotarios (Agro-Know Technologies)	Final	Input for chapter 10 and final formatting

This deliverable has been created under the scope of the AgINFRA project. It contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Executive summary

This study contributes to a larger workpackage that deals with community-driven policies and practices for agricultural data management and curation, in order to explore their migration to and support from a research infrastructure. For this purpose the workpackage addresses the harmonization and interoperability of agricultural (meta)data, semantics and ontologies through a linked data framework. This particular study was tasked to shed light on "existing policies, practices and lifecycles of data management and curation within the virtual communities ".

Although agINFRA may eventually provide services for many disciplines within the multidisciplinary agricultural sciences, we could address only a limited number of areas. After consultation with agINFRA partners, and other stakeholders like funding organizations, we have chosen:

1. To investigate the broader potential of the types of information artefacts that the agINFRA integrating services are dealing with, i.e.:

   a. Bibliographic resources for the exchange of documents
   b. Open Educational Resources for the exchange of learning.

2. To investigate the potential for other uses of the emerging infrastructure within specific subject areas:

   a. Geospatial information
   b. Germplasm collections
   c. Genomics and bioinformatics
   d. Agricultural economics

We could not use the normal concept of workflows and data lifecycles since we are dealing with communities and subject areas rather than specific organizations or computer systems. The following questions were addressed for the areas under study:

- What types of information are produced?
- What types of service exist for information exchange within the communities?
- Are there regulatory frameworks (standards, agreements on ownership and use rights)?
- What potential benefits could a new (semantically enabled) infrastructure bring?

As a thought model a scheme was used that has been developed by OCLC (Online Computer Library Center) for scientific information exchange. Some areas (like bibliographic resources and open educational resources) fit well into the model. But important parts of other

communities cannot be accommodated easily in the model, such as most of the work of Geospatial Information Systems (which currently collaborates most on standards and methodologies) and agricultural economics (especially where a simulation model rather than repositories are the mode of cooperation). The case of plant germplasm collections illustrates the data quality issues that may occur in a model of repositories with harvesters / aggregators.

Table of Contents

List of Figures

1 Introduction

This report as a whole focuses on the study of community-driven policies and practices for agricultural data management and curation, in order to explore their migration to and support from a research infrastructure like agINFRA so as to promote data sharing and development of trust. It will also address the harmonization and interoperability of agricultural metadata, semantics and ontologies through a linked data framework.

1.1 Scope

This report meets the requirements of D5.1 of Work Package WP5. Specifically, D5.1 is a "Review of existing agricultural data management practices, lifecycles and workflows: Study of data management and curation practices in the agricultural data providers covered from agINFRA, as well as identification of a set of directions/guidelines for mapping them into a generic workflow to be followed."

This review of broader domains of direct relevance to agriculture will then be important for developing an overarching framework (T5.2/5.3) that is necessary for a wider uptake of the agINFRA applications that are developed. The aim is to bring together the specific integration within agINFRA in the context of the data integration already in place in the broader information environment related to agriculture.

This broader information environment of direct relevance to agricultural science is defined, for the purposes of this study, by six domains. The domains, below, cover both areas of specific research focus (e.g. agricultural economics) and also areas where a particular type of information or data provides a platform for research activity in general (e.g. bibliographic resources).

- Bibliographic resources
- Open educational resources (OER)
- Geospatial information systems (GIS)
- Bioinformatics and genomics
- Plant germplasm collections
- Agricultural economics.

Each domain is studied, where possible, in the context of information systems, workflows and researcher behaviours which are currently prevalent in that domain.

Although each domain may have behaviours and workflows which are unique to itself, this might also be said of individual networks, organizations or even individual researchers or information managers. Nevertheless, it might be useful to generalize at this point to help to provide a conceptual framework for research activities across our six domains. Figure 1.1 below shows a generalized life cycle for scholarly research and communication. A description of this cycle is given in more detail in part 3 of this report.

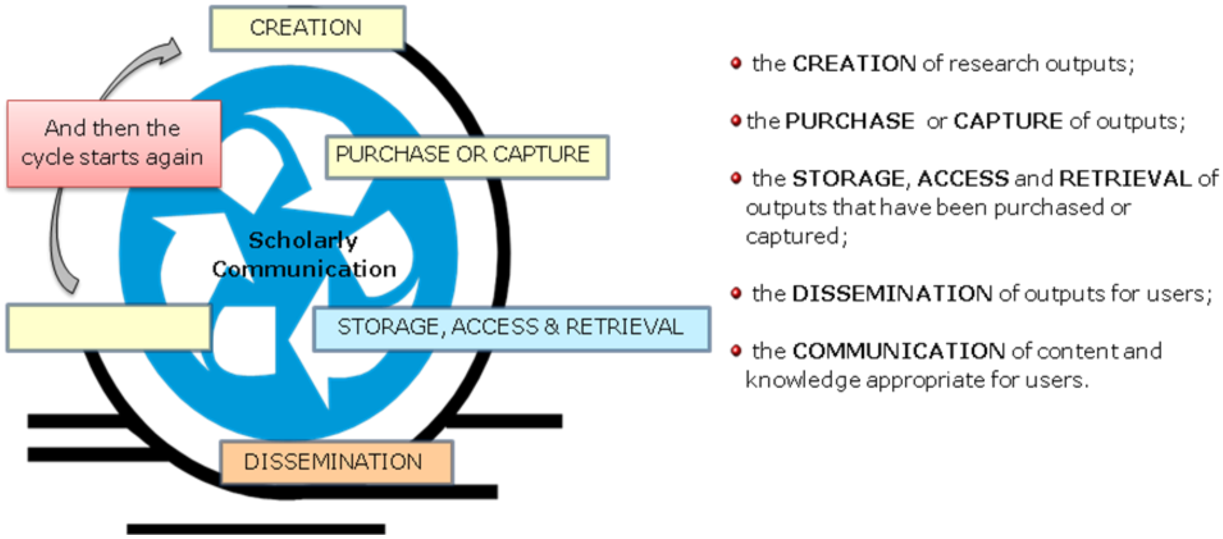

CREATION
PURCHASE OR CAPTURE
STORAGE, ACCESS & RETRIEVAL
DISSEMINATION
And then the cycle starts again
Scholarly Communication
the CREATION of research outputs;
the PURCHASE or CAPTURE of outputs;
the STORAGE, ACCESS and RETRIEVAL of outputs that have been purchased or captured;
the DISSEMINATION of outputs for users;
the COMMUNICATION of content and knowledge appropriate for users.

Figure 1.1: The scholarly communication cycle

1.2 Audience

This report is aimed primarily at the agINFRA participants and provides input to the direction of the agINFRA project as a whole. However, it is also hoped that the deliverable's outcomes are of relevance more broadly in communities where agricultural research and technical infrastructure development are working together to achieve enhanced openness and interoperability of information and data.

1.3 Structure

The structure of this report is presented in a sequence of domains, as listed above. Each domain is treated separately, though of course there are many issues in common between them. The study of each domain has a common structure, following generally, though not completely, the following sequence: background; the current environment; standards and metadata; communities, workflows and life cycles; implications for agINFRA; bibliography. At the end of the report is a synthesis of the issues raised by the studies of the six domains and their implications for agINFRA.

2 Objectives and methods

2.1 Objectives

This study was tasked to shed light on "existing policies, practices and lifecycles of data management and curation within the virtual communities that the five integrated services of WP-S2 will support". Before we started this study we needed to make a couple of choices:

> **Which communities?** agINFRA is an innovative exploration and therefore the work plan does not limit itself in advance to which communities could potentially be served and which not. But for this study we had to limit ourselves and a reasoned choice had to be made.
> **Which parameters?** "Practices", "life cycles" and "workflows" (a term that is also mentioned in the header of the workpackage) are probably to a degree fluid terminology, but some clarification is required to make sure the same issues are treated consistently across the different communities.

Of course these choices are not completely independent of each other: depending on what communities are chosen different parameters may be relevant.

2.1.1 Which communities

This question was discussed with agINFRA partners and others. How close should we stay to the integrative services that in a sense form the backbone of this infrastructure effort? The requirements of the direct stakeholders of these services have been studied in another study for workpackage 3. This D5.1 study should give a wider perspective. After the deliberations it was decided that we should:

1. Look at the broader potential of the types of information artefacts that the integrative services are dealing with, i.e.
   a. Bibliographic resources for the exchange of documents and related objects
   b. Open Educational Resources for the exchange of learning objects and their aggregations.

2. Look at the potential for other uses of the emerging infrastructure within specific subject areas. These areas should play a central role in modern agricultural science. Therefore related areas like biodiversity were not studied here (other infrastructure efforts are addressing those areas). We have chosen to cover areas where there are visible exchange activities within the research communities, and our choices were partly guided by the potential for investments by donors in those areas. In the end we decided to cover the following domains:
   a. Geospatial information
   b. Plant germplasm collections
   c. Genomics and bioinformatics
   d. Agricultural economics

2.1.2 Which parameters

There have been many attempts to formulate definitions for the concepts of workflow and data life cycle, and we did not feel that we would help agINFRA further with extended discussions of those issues. What is striking when reading such discussions is that workflows always are described as concatenated tasks of persons *within one or more specific organisations*, or, as the Wikipedia entry simply puts it "any abstraction of real work".

"Data lifecycles" often refers to lifecycles *within a specific system or connected systems*. We have chosen here to deal with classes of information artefacts (1.) or subject areas (2.) rather than specific organisations or systems. For all these categories we will describe a number of parameters:

- What types of information are produced?
- What types of service exist for information exchange within these communities?
- Are there regulatory frameworks (standards, agreements on ownership and use rights)?
- What potential benefits could a new (semantically enabled) infrastructure bring?

2.2 Methods

Information was gathered in these areas by:

- Searching the Internet, concentrating on the relevant portals and sites of important organisations in the subject area.
- Searching the scholarly literature especially for areas where that is the most important source (e.g. bioinformatics and genomics)
- Speaking with resource persons with an overview of the subject matter.

We did not aim to treat any of the areas exhaustively, but to gain the necessary understanding to provide agINFRA with relevant insights.

We adopted a diagram, see Figure 2.1, developed by OCLC as a thought model to test whether the experiences from the different domains fit in, and where they fit in.

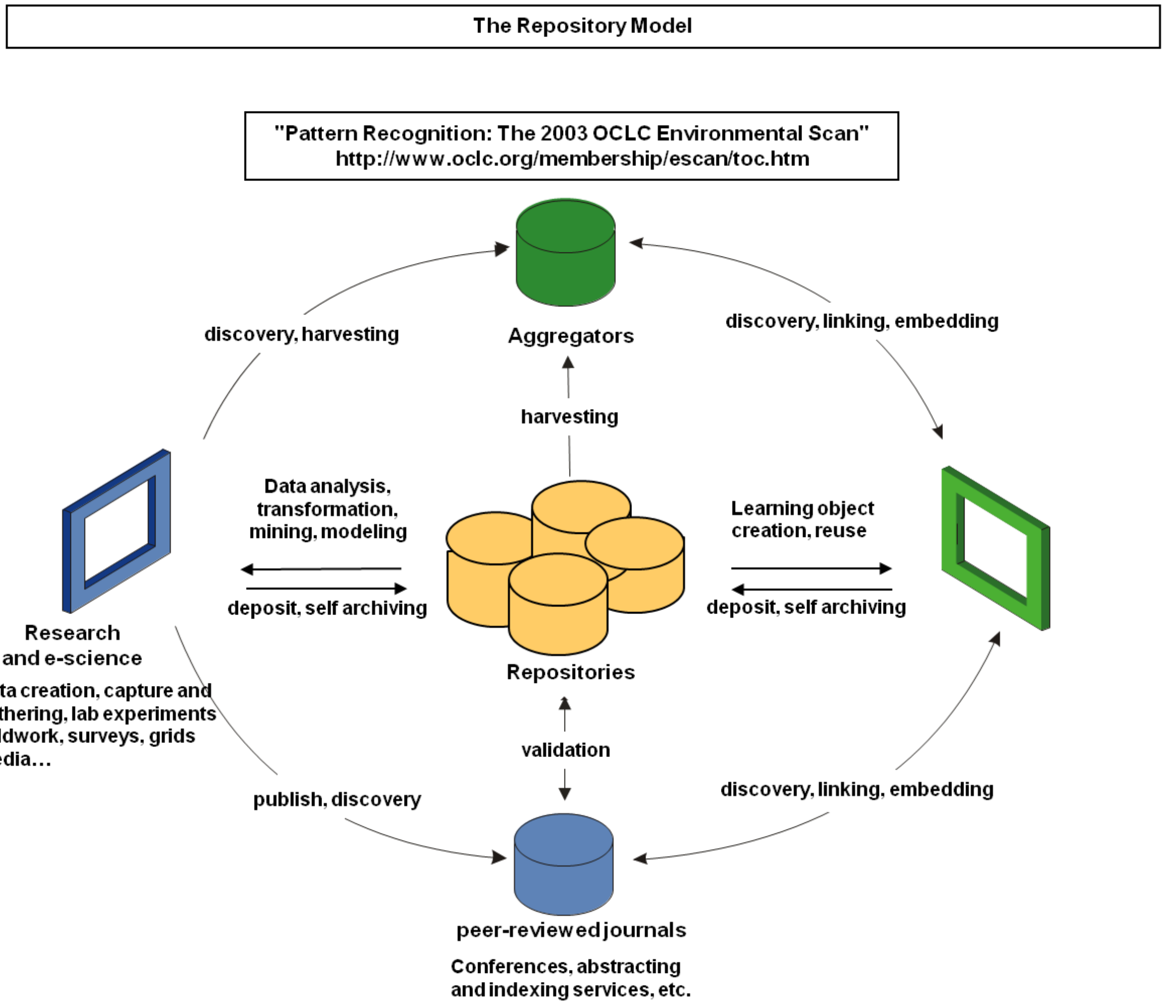

The Repository Model
"Pattern Recognition: The 2003 OCLC Environmental Scan"
http://www.oclc.org/membership/escan/toc.htm
Aggregators
discovery, harvesting
discovery, linking, embedding
harvesting
Data analysis, transformation, mining, modeling
deposit, self archiving
Learning object creation, reuse
deposit, self archiving
Research and e-science
Data creation, capture and gathering, lab experiments fieldwork, surveys, grids media…
Repositories
validation
publish, discovery
discovery, linking, embedding
peer-reviewed journals
Conferences, abstracting and indexing services, etc.

Figure 2.1: The OCLC repository model

Below is a brief introduction to the different domains and where they fit on the diagram, with the intention to help readers to select the areas that are most interesting for them.

2.2.1 Bibliographic resources

This area addresses this part of the diagram:

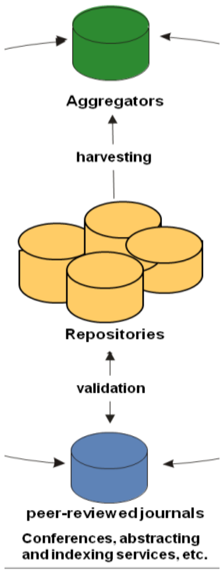


Figure 2.2: Bibliographic resources

Note that in the AGRIS network bibliographic information is not only seen as a means to discover pointers to relevant objects (i.e. documents), as the diagram implies, but also as a knowledge base in its own right.

2.2.2 Educational resources

This chapter deals with this part of the diagram:

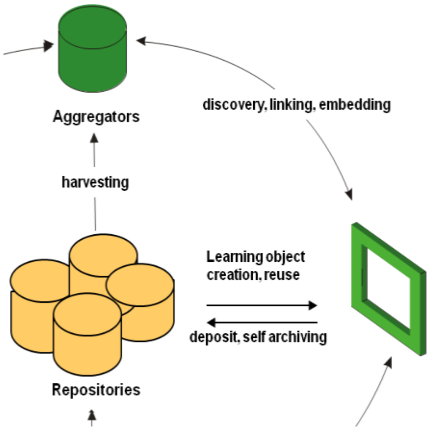


Figure 2.3: Open Educational resources

The discussion may give some insight into how these processes work in reality.

2.2.3 Geospatial Information Systems

The diagram does not cover all aspects that bring the GIS community together. These active communities centre around standards and exchanging methodologies rather than sharing objects in repositories. There are however current initiatives that may lead to a more prominent role for data exchange within the community. This may lead more clearly to a community that deals with this part of the diagram:

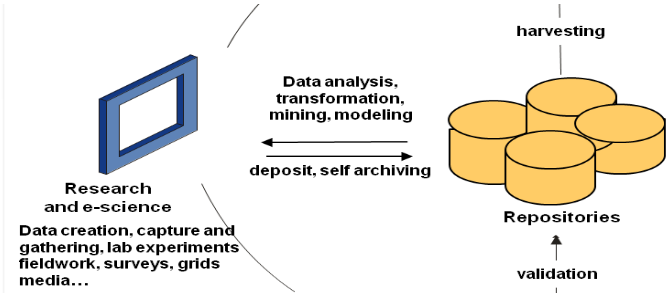

harvesting
Data analysis, transformation, mining, modeling
deposit, self archiving
Research and e-science
Data creation, capture and gathering, lab experiments fieldwork, surveys, grids media…
Repositories
validation

Figure 2.4: Geospatial Information Systems

2.2.4 Genomics and bioinformatics

In this area there are very crucial central services and indeed research is done by contributing sequences and comparing them to sequences from other research. There are no smaller services being harvested by aggregators, so this area fits with this part of the diagram.

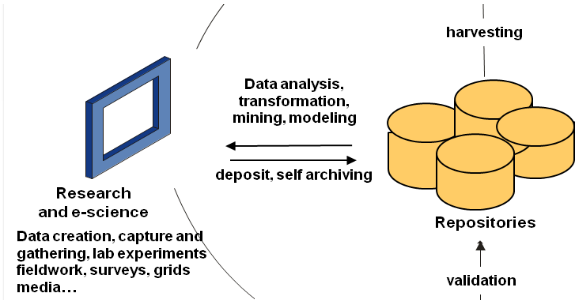

harvesting
Data analysis, transformation, mining, modeling
deposit, self archiving
Research and e-science
Data creation, capture and gathering, lab experiments fieldwork, surveys, grids media…
Repositories
validation

Figure 2.5: Genomics and bioinformatics

2.2.5 Agricultural economics

For agricultural economics we describe two quite different exchange models. At the macro level we describe a network where exchange is not done through repositories, but by collaboratively working at a simulation model.

At the micro level we describe how datasets from surveys are stored in harvestable repositories, so that community fits with this part of the diagram:

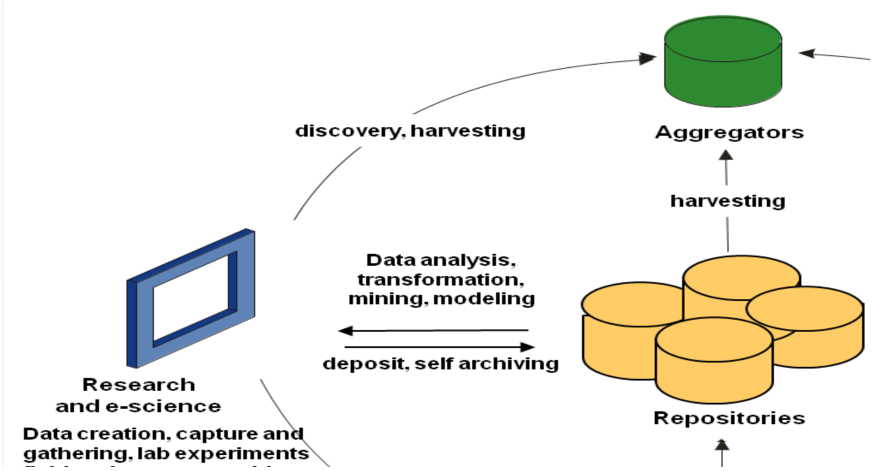

discovery, harvesting
Aggregators
harvesting
Data analysis, transformation, mining, modeling
deposit, self archiving
Research and e-science
Data creation, capture and gathering, lab experiments
Repositories

Figure 2.6: Agricultural economics

2.2.6 Germplasm Collections

Germplasm collections were one of the first types of data that were exchanged within the agricultural scientific communities. This area fits with this part of the diagram:

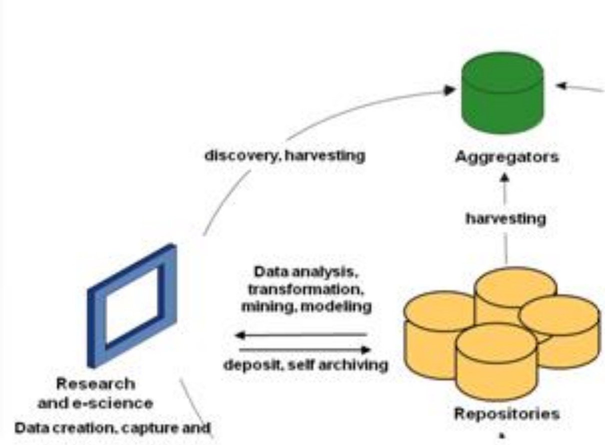


Figure 2.7: Plant germplasm collections

This chapter will illustrate how the options at the aggregator level can be limited by data quality at ‘ground level’.

3 Bibliographic Resources

3.1 Background

The emphasis of this piece concerns access to available Open Access (Suber, 2012) bibliographic resources or the metadata that describes them. However, resources which lie behind a ‘paid for’ access barrier are included where relevant because they form a significant part of this information environment. These resources have been created as part of, or are the result of, a piece of research or larger body of work. They represent the primary forms of scientific communication of the last 350 years and still have great significance in research communities. In this category we can include:

- ***Peer reviewed journal articles*** (pre- and post-publication), either paid-for or freely available through Open Access. They may be available through publisher or aggregator sites, or from an institutional or theme-based repository.
- As well as primary research articles, journals publish ***review articles*** summarizing progress in a field. These are included.
- ***Books*** and ***book chapters***, or other sub-units of books.
- Other materials, sometimes called ‘grey’ material, which include: ***theses, conference papers and presentations; Government, business and institutional research and other reports; and learning objects***. However, it can be argued that the term ‘grey’ material is being eroded by the increasing availability of documents that are now identifiable and accessible in an ordered way – much of it is no longer ‘grey’ but is prominently accessible.

The types of resource referred to above can all be seen as part of the scholarly communication cycle, which is applicable to researchers and developers in most fields of research activity. Figure 3.1 is adapted from a new Information Management Resource Kit (IMARK)[1] training module on ‘Strategic Approaches to Information’, Lesson 1.1.[2] In more detail:

> **Creation**: A researcher writes a paper or article for publication, or creates data from his or her research. A photographer will create images. A software developer creates a new piece of software.
>
> **Purchase or Capture**: An Information or ICT Manager will capture digital information from a variety of sources, from both within and outside an organization, to build them into a resource which is accessible for users. A librarian will purchase or obtain under license digital resources for the library’s collections.

[1] Information Management Resource Kit (IMARK) http://www.imarkgroup.org/

[2] The Module will be available in late 2012.

Storage, Access and Retrieval: Digital information content will be managed and stored by an Information or ICT Manager, or a publisher, or other distributor of information, in a way that makes it accessible to users and allows them to retrieve the information in an efficient way. An electronic publication or a dataset may be stored in a subject-based or institutional repository. An IT developer may create a repository interface to allow people to access and retrieve documents. A researcher may store and then access research data and results on his or her computer in an information management package. In the print world a publisher may store a paper book in a book distribution centre, or a librarian may catalogue and shelve new publications.

Dissemination: Information will be disseminated by a publisher, or by an information management system, or by an individual, and target the end users who have interest in using that information.

Communication: Information content may be transformed (or repackaged) by various intermediaries (such as video makers, developers of mobile phone networks, or publishers) into forms which are more directly usable by other end users such as policy makers, the media, farmers or poor communities. A developer may create an infographics app from a dataset.

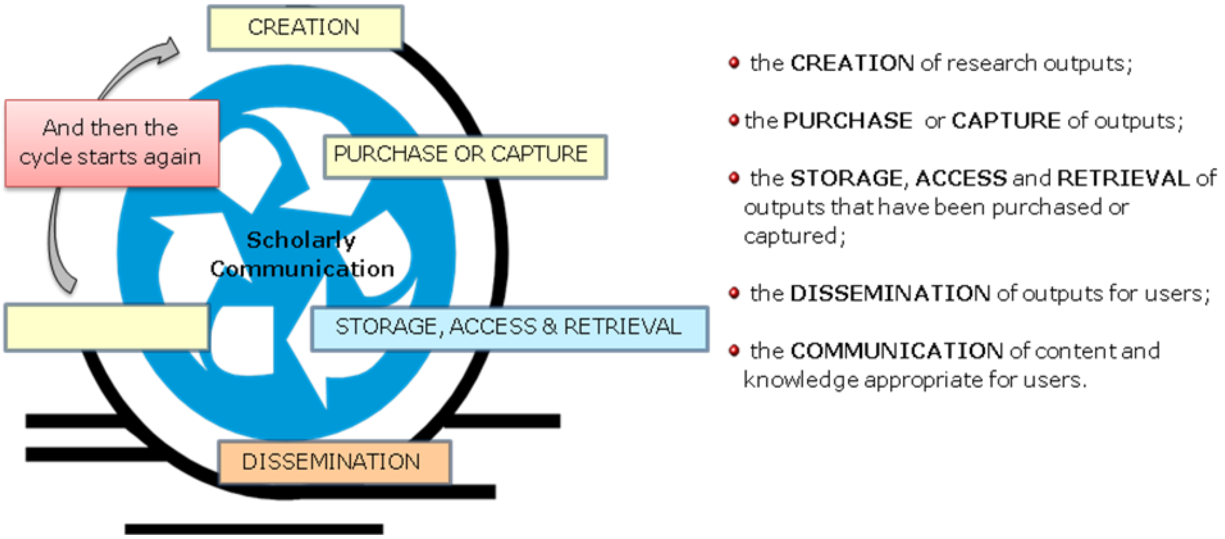

CREATION
And then the cycle starts again
PURCHASE OR CAPTURE
Scholarly Communication
STORAGE, ACCESS & RETRIEVAL
DISSEMINATION
the CREATION of research outputs;
the PURCHASE or CAPTURE of outputs;
the STORAGE, ACCESS and RETRIEVAL of outputs that have been purchased or captured;
the DISSEMINATION of outputs for users;
the COMMUNICATION of content and knowledge appropriate for users.

Figure 3.1: A scholarly communication cycle

An online worldwide survey of researchers in agriculture and related fields (Edge *et al*., 2011) was carried out in March 2011 by the Consultative Group on International Agricultural Research (CGIAR), Food and Agriculture Organization of the United Nations (FAO) and Global Forum on Agricultural Research (GFAR) on behalf of the CIARD (Coherence in Information for Agricultural Research for Development) initiative. The aim of the survey was to gain greater understanding of researcher behaviours and attitudes in relation to communicating research outputs and making such outputs open and accessible. Among other things, the survey indicated that for researchers the dissemination of research through journals and books (76%), conferences

(74%), and booklets, newsletters and pamphlets (47%), were still by far the most popular methods of reaching their peers and end users.

Why are these different types of research resource still so important for the development of research fields, particularly in the light of the opportunities now offered by Web 2.0 communication – such as blogs, Facebook, and so on? The answer is two-fold. Firstly, it lies in the long history of methods for summarising a piece or body of research. Secondly, these long-established behaviours have become the basis on which the quality and volume of research outputs are assessed – for instance through the Science Citation Index (SCI).

What is now changing rapidly is the way that most resources can be accessed on or via the internet. There is also growing interest and activity in providing research data with documents. The CGIAR, for instance, is making significant progress in this area. (Besemer *et al*., 2011) Such is the mass of available resources that the challenge now is to locate and retrieve material of interest in an effective and efficient way. This is a challenge for information management. Further, the entry into this field of search engines, such as Google Scholar and Scirus, is changing the pathways through which users obtain documents.

Many research outputs which used to be almost exclusively paid-for are now available freely through Open Access systems of various sorts. Although this proportion of OA is still too small, it is growing steadily. Materials are available from a range of sources, including publisher and aggregator sites, repositories, web sites, and also through some ‘social web’ sites. This report intends to be inclusive of all of these.

An issue of primary importance remains the needs and perceptions of researchers themselves. It is certainly still the case that many researchers, probably the majority, search for information on the Web using Google Scholar and Scirus and are happy to achieve ‘good enough’ results. Librarians and information specialists know better – in systems which have been developed around standards, the quality and consistency of metadata, and controlled vocabularies, the effectiveness of research information retrieval is greatly enhanced. The challenge is how to bring these divergent views and behaviours together.

3.2 The current environment: Open Access Publishing, Digital Repositories, and Search and Aggregation Services

3.2.1 Open Access Publishing

Open Access (OA) is the practice of providing unrestricted access via the Internet to peer-reviewed scholarly journal articles as well as theses, scholarly monographs and book chapters, and other research outputs. It may be Green OA (where the author is allowed to deposit a research article in a personal or institutional open repository, while it is also published in a paid-for journal disseminated by a publisher) or Gold OA (where the author, or his or her funder, pays a fee to have a research article published in a journal which is openly accessible to all.) Many publishers have now offered a so-called ‘hybrid’ Open Access option, whereby authors can pay a publication fee and have their article made Open Access within an otherwise subscription-based journal.

The number of openly accessible journals in the life sciences is increasing all the time. There are now a number of established services making available OA journals particularly with a focus on developing country titles and with a strong representation of agriculture and related fields. These include Scielo[3] (Scientific Electronic Library Online), Bioline International[4], Hindawi Publishing Corporation[5], Open J-Gate[6], and African Journals OnLine[7] (AJOL). ThomsonReuters recently announced that their Web of Knowledge service now includes Scielo in its coverage, which is a significant step forward for South American, and Open Access, research literature.

The Directory of Open Access Journals[8] (DOAJ), an indexed listing of quality-controlled Open Access journals from around the world, currently details over 200 in its ‘biology’ category, over 400 in ‘agriculture and food science’, and many more in the social and environmental sciences.

Biomedicine is one of the foremost fields in the availability of journal articles through OA. BioMed Central[9] (now part of the Springer Science publishing organisation), with 210 journals, deposits all its journal articles in PMC[10] (PubMedCentral) at the time of publication as well as hosting them on its own website. The Public Library of Science[11] (PLoS), another leading Open Access publisher, has developed some very high quality journals in biology and medicine (*PLoS Biology* and *PLoS Medicine*, among others).

3.2.2 Open Access Repositories

Repositories, whether institutional or subject-based, may contain several types of content, including preprints and postprints of journal articles, theses, conference articles, research data, training materials, images, and so on.

Agricultural information has lagged behind some fields which have benefitted from either more generous funding globally (biomedicine) or a pre-existing culture of e-information sharing (theoretical and particle physics). Biomedicine has developed PubMedCentral (PMC) in the USA and UK PubMed Central[12] (UKPMC) in the UK. These dominating, centralised repositories of biomedical information on a national scale indicate a much more centralised infrastructure than is the case in agriculture.

[3] Scielo http://www.scielo.org/php/index.php
[4] Bioline International http://www.bioline.org.br/
[5] Hindawi Publishing Corporation http://www.hindawi.com/
[6] Open J-Gate http://www.openj-gate.com/
[7] African Journals Online (AJOL) http://www.ajol.info/
[8] Directory of Open Access Journals http://www.doaj.org/
[9] BioMed Central http://www.biomedcentral.com/
[10] PMC http://www.biomedcentral.com/
[11] Public Library of Science http://www.biomedcentral.com/
[12] UK PubMed Central http://ukpmc.ac.uk/

However the CIARD Routemap to Information Nodes and Gateways[13] (RING) is a global registry of web-based services that will give access to any kind of information sources pertaining to agricultural research for development (ARD). The CIARD RING is the principal tool created through the Coherence in Information for Agricultural Research for Development (CIARD) programme to allow information providers to register their services in various categories and so facilitate the discovery of sources of agriculture-related information across the world. The essential feature of services in the CIARD RING is that they are exposing their metadata. There are currently 345 services being made available through the CIARD RING from 164 providers.

OpenDOAR[14], the Directory of Open Access Repositories, provides a listing of quality-controlled repositories around the world, currently a total of 2184. It also allows the user to search for repositories or search repository contents. The search facility is a Google custom search, the effectiveness of which depends on whether repository managers have effectively made their content accessible to Google web crawlers. OpenDOAR currently covers 81 repositories classified as ‘Agriculture, Food and Veterinary’ (Figure 3.2 below), with a further 101 for ‘Ecology and Environment’. The ‘Agriculture, Food and Veterinary’ repositories show a predominance in developed countries, but there is, overall, a worldwide distribution. Europe, North America and Asia make up over 80% of the total of 81.

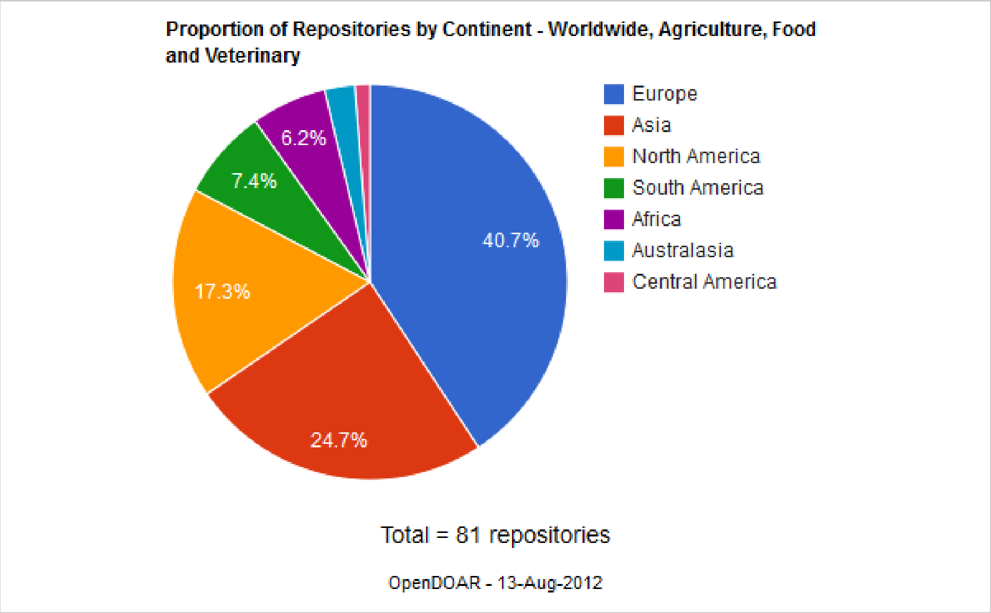


Figure 3.2: Global distribution of OpenDOAR repositories for Agriculture, Food and Veterinary

[13] CIARD Ring http://www.ciard.net/ciard-ring-0
[14] OpenDOAR http://www.opendoar.org/

It is also noteworthy that coverage by OpenDOAR of software type indicates that, for the Agriculture category, 67% of the total use either EPrints[15] or DSpace[16] software. A number of other softwares are identified, all at low levels of usage compared to these two, thus indicating that most repositories are making their data accessible using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[17].

In January 2010 the Knowledge and Capacity for Development Branch (OEKC) branch of FAO carried out a survey of Open Access repositories in the agricultural domain. The survey covered the areas of content, format and metadata, semantics, software and management. From 82 responses a valuable snapshot was obtained of infrastructure and behaviour related to these repositories at the time. Significant outcomes from this survey relate to the CIARD initiative's activities in both content management and advocacy. (FAO, 2010)

3.2.3 Aggregators

AGRIS[18] and CAB Abstracts[19] are perhaps the two most important bibliographic databases which are specific to international agricultural and related sciences. Though while AGRIS is freely available, CAB Abstracts is predominantly a paid-for product. One might add Food Science and Technology Abstracts (FSTA)[20] here, though it is focused on food science and nutrition. They are all abstracting and indexing services which expose to users very large amounts of highly structured (and therefore efficiently searchable) research information from the agricultural sciences. They enable searchers to quickly and easily narrow down their search options. Increasingly they now have links through to full text wherever it is located, whether in libraries, on publisher's websites, or elsewhere.

AGRIS: More than 150 national, international and intergovernmental centres are currently providing bibliographic metadata to AGRIS. Linkage to full text, wherever it is located in the world, is enabled. The records in AGRIS are automatically indexed by Google Scholar and consequently displayed in any Google search. Many different types of content/data provider contribute to AGRIS: journal publishers; national centres, some of whom also maintain their own database of national content; and others who contribute records to AGRIS as their sole output. A lot of this material is harvested from providers, but AGRIS generally does not harvest from other aggregators, though it does do so from DOAJ.

CAB Abstracts: contains approximately 8 million bibliographic records of the world's agricultural research information. Each year around 280,000 new research article references are added. CAB Abstracts links through to full text where possible using DOIs or URLs. A full text

[15] EPrints http://www.eprints.org/
[16] Dspace http://www.dspace.org/
[17] Open Archives Iniative Protocol for Metadata Harvesting http://www.openarchives.org/pmh/
[18] AGRIS http://agris.fao.org/knowledge-and-information-sharing-through-agris-network
[19] CAB Abstract http://www.cabi.org/default.aspx?site=170&page=1016&pid=125
[20] Food Science and Technology Abstracts http://www.ifis.org/

linkage rate approaching 60% is claimed, though many of these are to documents that are not OA.

AGRICOLA[21] (NAL Catalog): serves as the catalog and index to the collections of the USDA National Agricultural Library, as well as a source for world-wide access to agricultural information. Thousands of AGRICOLA's records are linked to full-text documents by URL.

MEDLINE[22]: is the openly accessible U.S. National Library of Medicine's (NLM) bibliographic database that contains over 19 million references to journal articles in the life sciences with a concentration on biomedicine. Search results provide an indication of free electronic full-text availability.

3.2.4 Indexing Services/Search Engines

A range of services are available for locating and accessing scholarly content, both paid-for and open access, on the internet. Scirus and Google Scholar depend on the indexing already done by content originators. The user can link from the results of a search to the original document wherever it is stored. If the article is in a paid-for journal the user will have to pay to gain full access. AGRIS is also partly a search engine that, among other functions, uses Linked Open Data (LOD)[23].

Scirus[24]: developed by publisher Elsevier, it is a science-specific search engine. With over 460 million scientific items indexed, it allows researchers to search for not only journal content but also scientists' homepages, courseware, preprint server material, patents and institutional repository and website information.

Google Scholar[25]: indexes "full-text journal articles, technical reports, preprints, theses, books, and other documents, including selected Web pages that are deemed to be 'scholarly'." The service also offers 'citing papers' and 'related work' options – which indicates backward and forward citation tracking like that provided by Web of Science.

Scopus and **Web of Science**: similar to Scirus and Google Scholar, but they are subscription products. Both Scopus and Web of Science provide federated search of content (abstracts and citations) from many publishers, concentrating predominantly on journal articles and conference materials. Scopus, owned by Elsevier, offers 19,000 titles from 5000 publishers. Both services also offer interoperability with Medline and other resources, and advanced search facilities.

[21] AGRICOLA http://agricola.nal.usda.gov/
[22] MEDLINE http://www.ncbi.nlm.nih.gov/pubmed/
[23] Linked Open Data (LOD) http://linkeddata.org/
[24] Scirus http://www.scirus.com/
[25] Google Scholar http://scholar.google.com

A comparison of search services was carried out by (Falagas, *et al*., 2008). The authors found Google Scholar unreliable and 'opaque' concerning material covered, but useful. "Google Scholar, as for the Web in general, can help in the retrieval of even the most obscure information but its use is marred by inadequate, less often updated, citation information."

Although the presence of these search and access services may not seem to be directly relevant to agINFRA, their presence and success are indicative of user behaviour. Researchers will use services which give them rapid and direct answers to questions – particularly if they are free and open. Bibliographic databases are perhaps not as important to them as we would like to think. Whether these services meet the standards of an information specialist is usually not in the mind of the researcher. The development of agINFRA may need to address this conundrum.

3.3 Current environment: standards and metadata

Whatever the type of bibliographic resource, and wherever it is located, it is important that it should be made available in a form which maximises accessibility and interoperability.

Accessibility requires meaningful metadata that makes use of comprehensive indexing. Metadata should be meaningful for the community in which it is exchanged and this often requires the development of specific Application Profiles (AP). Dublin Core[26] has been the standard for descriptive metadata for documents in online resources. It consists of fifteen information elements and was introduced in 1995 specifically for describing networked resources. XML and RDF have become important as syntaxes for expressing the Dublin Core.

Interoperability is achieved mostly either through harvesting or federated search. Harvesting is achieved by extracting metadata from a large group of resources and copying it into a central database. This central database then provides services to end-users. There are two important roles: *data providers*, i.e. repositories, that expose their data for harvesting; and *service providers* who harvest the repositories, combine them in a database, and develop value-added services using that data. Examples of such services are AGRIS and BASE[27] (which has a strong representation of agricultural material). OpenAIRE[28] harvests from repositories and journals, and is soon to include data. To control the flow of information between data providers and service providers the OAI-PMH protocol is widely used. In federated search metadata remains in the various resources but the federated search engine issues simultaneous queries and integrates the results. Search engines such as Google and Scirus work in this way.

The OpenURL[29] provides another interoperability mechanism by encoding metadata elements of a citation for a bibliographic resource as a URL. The OpenURL is, in effect, an actionable URL that transports metadata or keys to access metadata for the object for which the OpenURL is provided. The citation is provided by using either a global identifier for the resource, for example

[26] Dublin Core http://dublincore.org/
[27] BASE http://www.base-search.net/about/en/
[28] OpenAIRE http://www.openaire.eu/
[29] OpenURL http://openurl.ac.uk/doc/

a Digital Object Identifier[30] (DOI), or by encoding metadata about the resource, for example title, author, journal title, etc., or by some combination of both approaches.

The Semantic Web and Linked Data is starting to allow a more flexible approach to data linkage and there could be a gradual evolution away from the requirements of XML and APs such as the AGRIS AP. Linked data declares links between RDF (Resource Description Framework) data sets and creates links where semantic equivalents occur. The utilization of linked data using RDF and standard vocabularies (ontologies) provides a more flexible linking and interoperability environment.

3.4 Communities, life cycles and workflows

3.4.1 Where are the communities?

For a class of objects as broad and as ubiquitous as bibliographic resources it cannot be said that a single community exists. Communities of researchers usually form around particular subject themes – e.g. crop protection or animal genomics. There are also communities of technical developers, but it seems that they usually develop networks separate from the subject-focused ones. It could also be said that there is a third group of communities, made up of publishers and librarians, who relate both to the interests of the technical developers and the researchers. Many of these subject-based researchers are producing data and documents while publishing their core results in journals, without interacting at all with the growing interoperable and semantic world. It is possible that the technical developers/researchers are moving ahead at a speed which is not bringing the [agricultural] researcher along with them.

3.4.2 Creating new digital communities

Certainly, the new digital environment facilitates the development of communities of interest based around information resources. If digital document collections on websites are accessible on the Internet and likely to be indexed by search engines like Google, the community can create a specialized search engine, for example with the Google Custom Search Engines. An example of a community that has created a custom search engine relevant to development is Focuss[31].

Or, if the information is held in different databases, there are two basic options:
Create a joint database of the metadata from these different services through harvesting. An example of such a service in the development sector is AIDA[32] that brings together project and data registries from different development agencies.

This can also be approached through federated search. The library world has been at the forefront of developments of this sort which enable end-users to integrate access to different services (both

[30] Digital Object Identifier http://www.doi.org/
[31] Focuss http://www.focuss.info/
[32] AIDA http://www.aiddata.org/content/index

paid and free-of-charge). The CGVirtual Library[33] is a service that uses one of these products, Metalib[34], to integrate access to many services in different subject areas related to agricultural research for development.

So the opportunities to create communities of common interest have never been greater.

3.4.3 Research life cycles and workflows

The difficulty in considering research and document workflows and life cycles in general is that one must range across the many different ways in which a document or research output may have come into being, and the different reasons that it was developed. Probably the only thing that unites them all is origination within a research workflow, within a particular organization or community, and that they have at some point originated in data produced by research and/or analysis. A generalized scholarly communication cycle or workflow is shown in the Introduction to this document. An even more simplified life cycle for both research and data could be presented as:

> *SEARCH & DISCOVERY > CAPTURE > ANALYSE & EXPERIMENT > PUBLISH & DISSEMINATE > STORE & ARCHIVE > SEARCH & DISCOVERY (again), and so on.*

Because of the breadth of scope involved in looking at bibliographic resources as a whole, the starting point should probably be where a document or other output is being created and then described with meaningful metadata, with the intention of maximising dissemination and availability. In the context of agINFRA it will be useful to look in more detail at the AGRIS database/repository.

3.4.3.1 The AGRIS Network

The AGRIS Network is an international initiative based on a collaborative network of institutions whose aim is to promote free access to information on science and technology in agriculture and related subjects. For over 35 years the AGRIS Network has indexed and given access to bibliographic metadata used in the Food and Agriculture Organization of the United Nations (FAO) efforts to end world hunger. In this way it has served both developed and developing countries in order to give scientists and students free access to agricultural knowledge.

The goal of the AGRIS Network is to enhance the exchange of science and technology research outputs in agriculture and related subjects, especially grey literature. Until the late 1990’s, outputs mainly comprised a centralized bibliographic database - the AGRIS Repository - and associated products. Since 2000 the efforts have focused increasingly on building up decentralized capacities in its participating resource centres. From 2005 a new vision and

[33] CGVirtual Library http://www.aiddata.org/content/index
[34] Metalib http://www.exlibris.co.il/category/MetaLibOverview

strategy was developed emphasising partnerships, collaboration and networking, with the following objectives:

- **Decentralised** approach with greater emphasis on national partnerships
- Greater **diversity** of research-oriented organisations
- Strengthened role in **capacity building**
- Focus on management and availability of **full text** digital content in agricultural science and technology;
- Greater availability of **associated information** about activities, organisations, and people
- Continually improving set of **web-enabled** AGRIS methodologies and tools (with a focus on the establishment of standards).

Currently AGRIS has 130+ active centres all over the world.

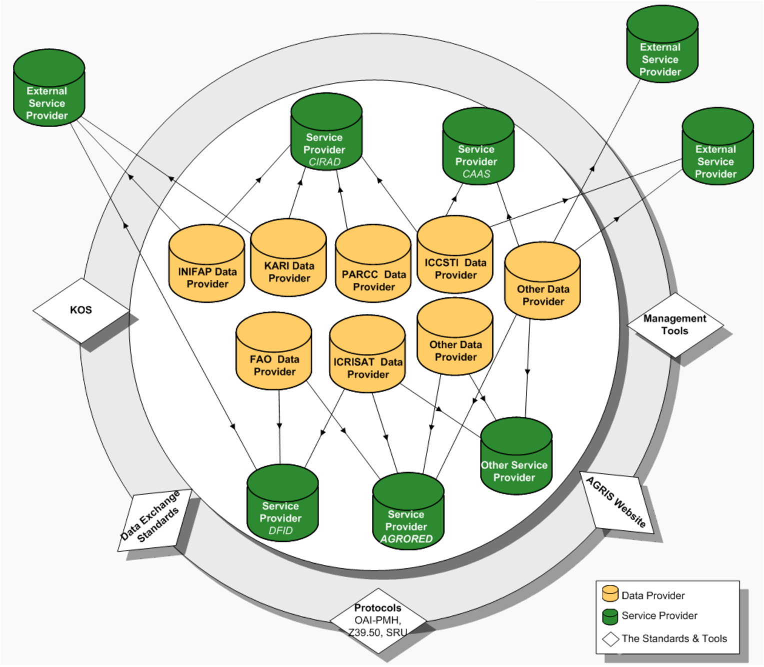


Figure 3.3: Infrastructure relationships within the AGRIS Network

The centres that participate in the AGRIS Network provide data to the AGRIS repository, a collection of nearly 2.9 million bibliographic references encoded in an XML qualified Dublin Core metadata format that eases sharing of information across dispersed bibliographic systems. Its high quality content description is enhanced by the AGROVOC thesaurus[35], extensively used by cataloguers to enrich data indexing in agricultural information systems. The map in Figure 3.4 below shows the current locations of these providers. In total the providers are made up of official AGRIS centres, other resources that are not necessarily officially part of the AGRIS Network (for example OceanDocs[36]), and also journal publishers such as Scielo.

[35] AGROVOC http://aims.fao.org/standards/agrovoc/about

[36] OceanDocs http://www.oceandocs.net/

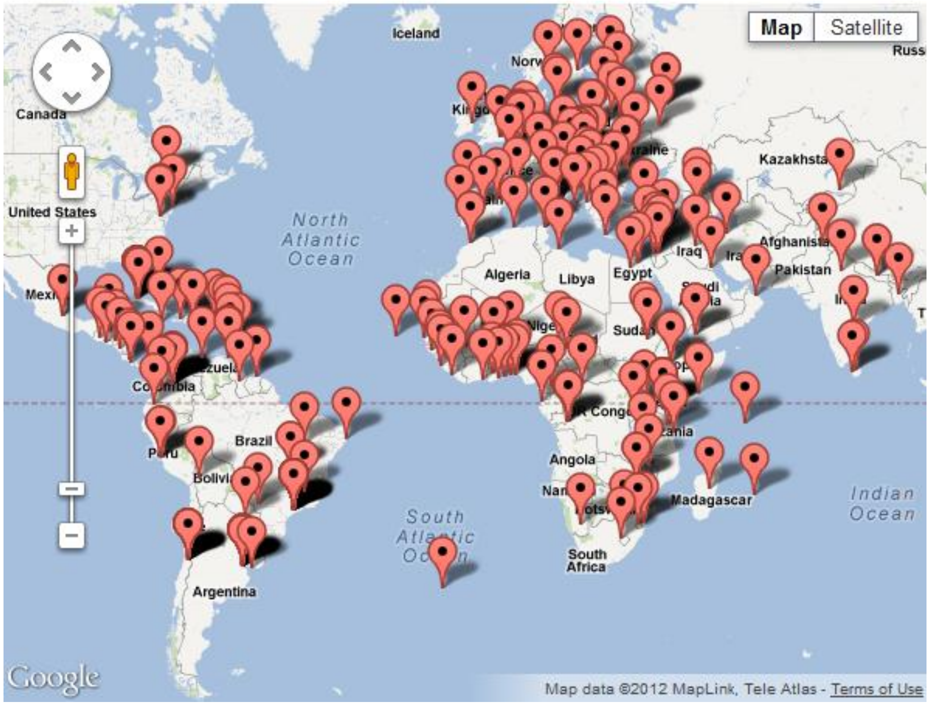


Figure 3.4: The AGRIS Network map

Figure 3.5 shows all the elements of the AGRIS Network and their connections in the repository model. The workflow is divided into two parts: Content Management (left) and Exposing Metadata (right). To summarize the most important relationships:

- the AGRIS Secretariat (FAO) provides standards and tools to the AGRIS Centres;
- the Information Management Specialists of the AGRIS centres are responsible for the Content Management - along with researchers they input full text documents and associated metadata using the standards promoted by AGRIS initiative;
- the AGRIS centres' databases can then be interoperable and become Data Providers allowing them to expose their metadata for harvesting.

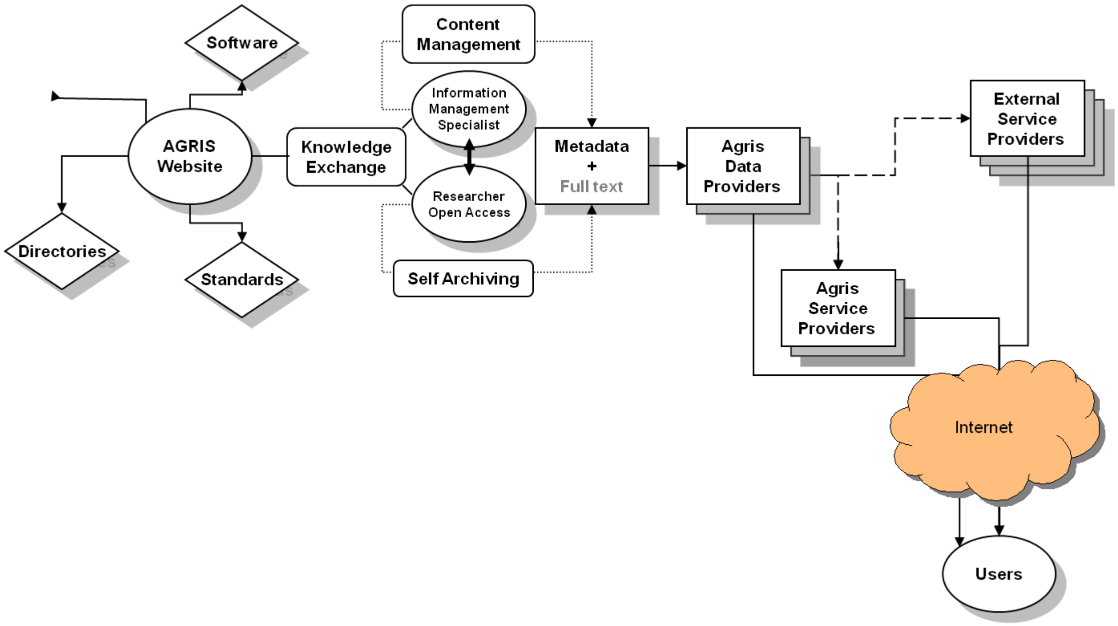


Figure 3.5: The AGRIS Network Repository Model

Since 2007, with the proliferation of full-text documents held in open repositories, and the increased awareness of repository managers of the need to develop full interoperability, the number of records with links to full text documents indexed by AGRIS is increasing year on year. Currently the AGRIS collection includes 2,690,844 bibliographic records, of which 82.24% are citations from scientific journals and 21% have links to full text documents. Figure 3.6 gives an overview of the evolution of records published in AGRIS from 2005 to 2012.

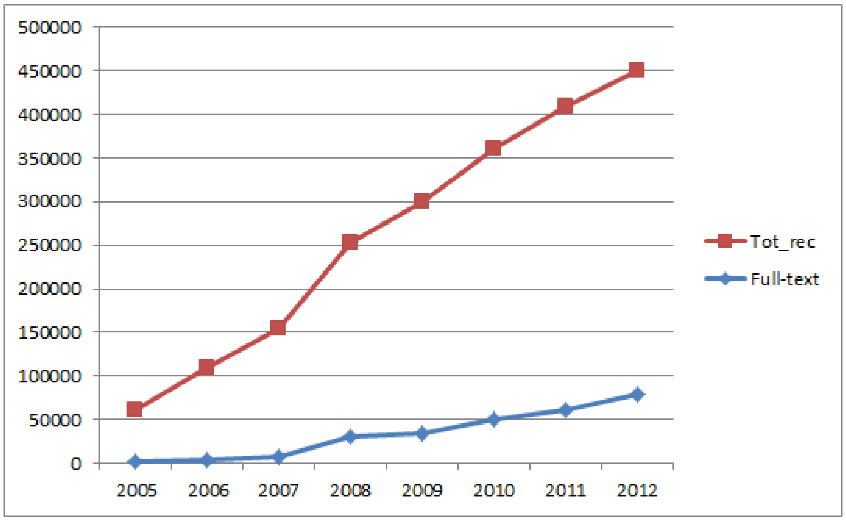
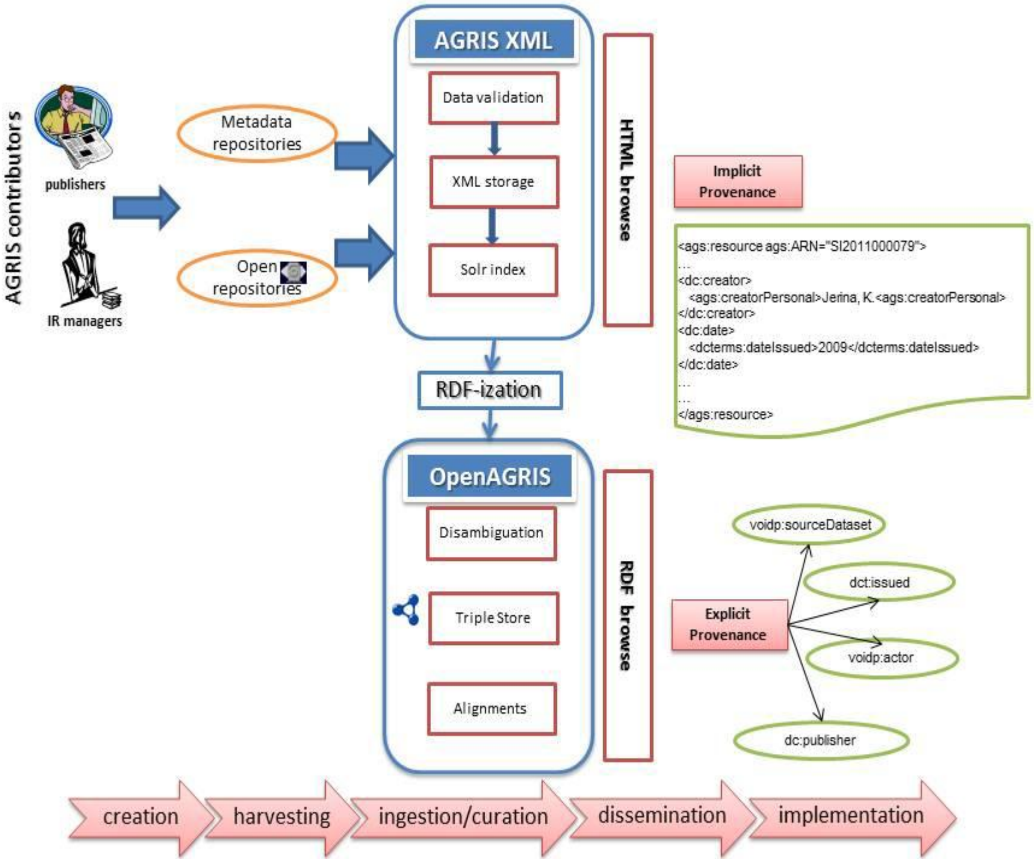


Figure 3.6: A cumulative representation of the increase of open access in AGRIS in the last 7 years

The life cycle of an AGRIS record has changed a great deal in recent years. 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 has changed consistently its "ingestion approach" and currently also indexes data harvested from service providers such as DOAJ (Directory of Open Access Journals) whose content comes from external publishers. Following the conversion of its indexing thesaurus AGROVOC into a concept-based vocabulary, the decision was made to express the entire AGRIS repository in Resource Description Framework (RDF)[37] as Linked Open Data. As part of this approach OpenAGRIS[38] has been developed to show semantic mash-up in operation. Figure 3.7 below shows the long flow of an AGRIS artefact, from genesis to dissemination, through both AGRIS XML and OpenAGRIS routes.



```
<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>
```

Figure 3.7: Flow of an AGRIS artefact, from genesis to dissemination, through both AGRIS XML and OpenAGRIS

[37] Resource Description Framework (RDF) http://en.wikipedia.org/wiki/Resource_Description_Framework
[38] OpenAGRIS http://aims.fao.org/openagris

In this sense AGRIS is at the centre of a community of data providers who are willing to expose and deliver their data for inclusion in the AGRIS repository.

The Norwegian University Library of Life Sciences provides data to AGRIS. It has a long history of database and repository management, having created BIBSYS, the Norwegian national database (covering all subject areas) in 1986. More recently they have developed Brage-UMB, their institutional repository. Brage captures scientific articles, doctoral theses, master's theses and other works. All content is openly available as full text documents on the internet. The primary purpose of the repository is to make the work produced at UMB more internationally visible and available. Raw data is also being stored, though it is not yet openly available.

At the moment the AGRIS provision has a workflow separate from that of Brage, though work is under way to integrate the systems thus allowing the workflows to be combined. The Library sees clear benefits, in terms of international exposure, in being made more visible through the AGRIS system.

Brage uses a generalized form of Dublin Core and uses DSpace to manage the repository content. Brage is indexed by Google as well as NORA, the Norwegian Open Research Archives search engine. It exposes data observing the OAI-PMH. Brage is also registered with OpenDOAR and OAIster.

3.5 Implications for the agINFRA infrastructure

It seems that the challenges to the development of openness, interoperability and semantic consistency are considerable in the area of bibliographic resources. These challenges relate more to researcher and institutional behaviour than to deficiencies in the availability of technical infrastructure. The survey of researcher behaviour referred to in the first section of this (Edge *et al.*, 2011) indicated that less than 30% of researchers were making their outputs available in a repository. Other forms of e-communication, such as Web 2.0 tools and e-newsletters, were being employed even less. There is a need for continuing advocacy to persuade practising researchers (and their organizations) of the value to them, and to their communities, of enhancing the accessibility, interoperability, and relevance of their research outputs. Without a clearer alignment of researcher behaviour with the opportunities available in the new digital environment, global progress toward 'openness' will indeed be slow. Note the aims and activities of the CIARD programme in this context.

There is a growing trend both within publishing, and in the general world of exchanging research information, toward providing data and other objects along with a resource. Aggregation of multiple objects of this sort will depend on high quality metadata. One approach to this is the OAI Object Reuse and Exchange (OAI-ORE)[39] specifications – defined as a standard for the

[39] OAI Object Reuse and Exchange (OAI-ORE) http://www.openarchives.org/ore/

identification and description of clusters of Web resources (known as "aggregations"). ORE provides an aggregation with a URI, a description of its constituents, and optionally the relationships among them. Further, because of the development of Open Access in publishing, it becomes important to know which version of a paper the user is accessing – final preprint (or an intermediate stage), the postprint, and so on. Initiatives such as ORE can address these gradations.

Finally, the Semantic Web and Linked Data are presenting new challenges for the optimal operation of linking and interoperability in the semantic environment. Such as:

- Automated discovery of links in documents
- Association between the indexes of bibliographic databases and other sources using, for instance, AGROVOC URIs.
- An integrated authoring environment for the content management systems, particularly for metadata.

Bibliographic resources are still fundamental to the way that most researchers work. The Semantic Web needs to provide ways of working that operate seamlessly in the background from the perspective of the typical user (in this case, the researcher). The researcher will want not just an uncomplicated user environment, but also a level of control and choice which gives them relevant information which relates also to the way they work. Without this there is a danger of the technical development communities running into the distance while leaving behind the user communities.

These issues are starting to be addressed by the continuing development of AGRIS into services such as OpenAGRIS. AGRIS already provides a framework through which agricultural research outputs from across the globe can be harvested and made openly available to users. Further, development of AGRIS is taking place which will allow it to operate in a Semantic Web environment in ways which will require constant innovation and development.

4 Open Educational resources (OER)

4.1 Background

The growth of the Internet offers many opportunities for improving access and transfer of knowledge and information from universities and other learning bodies to a wide range of users. One result of this is that the growth of the Internet over the last 20 years has seen the concomitant growth in online learning resources. In recent years a particular focus on ‘Open’ resources has developed along with the Open Access (OA) and Open Source Software (OSS) movements. This report focuses mostly on ‘Open’ as opposed to paid-for systems, and where possible on specifically agriculture-related resources.

The field has also seen the emergence of the learning object (LO) concept. Although there are many different definitions of the learning object (LO), it can be said that a learning object is any type of digital resource that can be reused to support learning. Learning objects and/or their associated metadata are typically organised, classified and stored in online databases referred to as learning object repositories (LORs) or open educational resources (OER). Online, objects used for learning exist and interoperate at different levels of granularity. They could be a simple text document, a photograph, a video clip, a three dimensional image, a Java applet or any other object that might be used for online learning. The object becomes useful for learners when a lesson is added to it. Different lessons might be created from one component. A learning object may have originated as an educational object, or it may have resulted from research but proved to be usable also in an educational context.

It has been recently proposed that open educational resources (OER) should conform to three main elements (Geser, 2012): that access to open content (including metadata) is provided free of charge for educational institutions, content services, and the end-users such as teachers, students and lifelong learners; that the content is liberally licensed for re-use in educational activities, favourably free from restrictions to modify, combine and repurpose the content, and, consequently, that the content should ideally be designed for easy re-use in that open content standards and formats are being employed; that for educational systems/tools software is used for which the source code is available (i.e. Open Source Software) and that there are open Application Programming Interfaces (open APIs) and authorisations to re-use Web-based services as well as resources (e.g. for educational content RSS feeds).

So learning objects enable and facilitate the use of educational resources online. Internationally accepted specifications and standards make them interoperable and reusable by different applications and in diverse learning environments. The metadata that describes them, for instance the IEEE LOM[40], facilitates searching and renders them accessible.

[40] IEEE LOM http://en.wikipedia.org/wiki/Learning_object_metadata

4.2 The current landscape of Open Educational Resources

The issue of whether there is a 'community' of educational resource developers and/or deliverers will be addressed later. It should be noted that the OER landscape is a very dispersed one, lacking central foci of activity, unlike for instance genomics or geospatial information. This is not necessarily a bad thing – it is a reflection of the many different needs of users around the globe. It may be that this provides a need for agINFRA to align itself with. This will be addressed later in this report.

Many organizations and services are operating in the OER arena, some focusing on technical and infrastructural development, and others on access for end users, sometimes in very specific areas of knowledge and with agriculture represented strongly. However, there are different levels of commitment to technical/infrastructural development. It may be that as many as 90% of discoverable educational resources are not interoperable or have metadata of inadequate quality. [This was a view expressed to the author during his research for this report.]

Across the whole spectrum of resources and communities, their main emphasis may be split into two categories: 1) where developers of standards and interoperability communicate and work together irrespective of subject focus, though it is likely that agriculture-focused resources will be accessible through them; 2) where there is a focus on agriculture and related areas for the end user. Agricultural resources do not display features which are fundamentally different from other fields, and generic issues for the whole of the OER field apply to agriculture in the same way that they do to other areas of application. Issues related to metadata and interoperability will be addressed in more detail later in this report, as will the issue of the development of communities.

4.2.1 Services/organizations primarily working on technical infrastructure development

This section highlights some of the services which are primarily progressing the application of standards, interoperability and the management of educational resources. Although they do not offer a particular subject focus, in many cases agricultural materials are accessible through them.

GLOBE (Global Learning Objects Brokered Exchange)[41]. GLOBE is a one-stop-shop for learning resource organizations, each of them managing and/or federating one or more learning object repositories. It should be noted that it is now moving toward fulfilling the role of a community. GLOBE makes a suite of online services and tools available to its members for the exchange of learning resources, and is set up as a worldwide Open Community guided by some key principles, particularly: providing open specifications and community source code as much as possible, openly shared among and beyond community members; and, using open standards, where appropriate, and contributing back to the development of these standards based on

[41] GLOBE http://globe-info.org/

experiences and best practices. ARIADNE (Europe), MERLOT (USA), LACLO (South and Central America), European Schoolnet, OER Africa, NIME (Japan) and others are members of GLOBE, making it a very significant gateway to OERs worldwide. Agriculture resources are accessible through GLOBE. A search on the term 'agriculture' produced 11,864 results, though not all of these are specifically objects or learning resources.

The ARIADNE Foundation[42]. The ARIADNE Foundation is a not-for-profit association that works to foster the sharing and reuse of learning resources. ARIADNE works to create a standards-based technology infrastructure that allows the publication and management of open digital learning resources. The aim is to provide flexible and efficient access to large-scale educational collections in a way that exceeds the capabilities of search engines. Agriculture-related resources are accessible through ARIADNE. It is a member of the GLOBE Alliance[43].

Latin-American Community on Learning Objects (LACLO)[44]. LACLO is an open community, made up of individuals and institutions interested in research, development and application of technologies related to learning objects in Latin American Education. Its main mission is to help to bring together the different initiatives in the Region to disseminate the advances and benefits of LO technology. The LACLO federation maintains a repository called FLOWER (Latin American Federation of Learning Object Repositories), which currently brings together 50,000 objects. It is also a member of GLOBE (see above), currently giving access to almost 1 million objects internationally. LACLO started life with a central focus on developing interoperability across regional resources but now it is moving further toward offering a community environment.

IMS Global Learning Consortium[45] focuses on standardising learning object metadata (LOM) and providing tools and solutions, but the service is not strictly 'open' – payment for membership is required. It is an international consortium that contributed to the drafting of the IEEE Learning Object Metadata (LOM), together with the ARIADNE Foundation, and endorsed early drafts of the data model as part of the IMS Learning Resource Metadata specification.

DLESE (The Digital Library for Earth System Education)[46]. DLESE is an example of a service which is attempting to create a community of technical developers working alongside educators and users (Marlino *et al*., 2009). Educators, students, and scientists work together to improve the quality, quantity, and efficiency of teaching and learning about the Earth system at all levels. DLESE provides:

- Access to quality-controlled collections of educational resources

[42] The ARIADNE Foundation http://www.ariadne-eu.org/content/about
[43] GLOBE Alliance http://globe-info.org/
[44] Latin-American Community on Learning Objects (LACLO) http://www.laclo.org/
[45] IMS Global Learning Consortium http://www.imsglobal.org/index.html
[46] DLESE (The Digital Library for Earth System Education) http://www.dlese.org/library/index.jsp

- Access to Earth data sets and imagery, including the tools and interfaces that enable their effective use in educational settings
- Support services to help educators and learners create, use, and share educational resources
- Communication networks to facilitate interactions and collaborations across Earth system education.

DLESE resources include materials for both teachers and learners, such as lesson plans, maps, images, data sets, visualizations, assessment activities, curriculum, online courses, and so on. The US National Science Foundation provided funding for the development of DLESE which is now operated by the National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory and the NCAR Library on behalf of the education community.

Connexions[47] is an educational resource platform consisting of an educational content repository and a content management system optimized for the delivery of educational content. It currently contains more than 17,000 learning objects or modules in its repository and over 1000 collections (textbooks, journal articles, etc.) are used by over 2 million people per month. A search on ‘agriculture’ gave 394 hits. Connexions aims to combine technical development with a community of authors who can convert and adapt information in the Connexions repository. Connexions promotes communication between content creators and provides various means of collaboration through author feedback and shared work areas.

OSCELOT[48]**,** the Open Source Community for Educational Learning Objects and Tools, brings together developers to collaborate on and share open source software related to e-learning. OSCELOT is an indicator of how the interest of the Open Source community in educational resources has grown in recent years.

In addition, there are a number of software packages which have become prominent by providing platforms for educational resource developers. These include:

> **Moodle**[49] is an Open Source content management system. for creating online web sites for students. There are also activity modules (such as forums, databases and wikis) to build collaborative communities of learning around particular subject matter.
>
> **SCORM**[50] is a set of technical standards for e-learning software products. It governs how online learning content and learning management systems communicate with each other. It is purely a technical standard.

[47] Connexions http://cnx.org/
[48] OSCELOT http://oscelot.org/
[49] Moodle http://moodle.org/
[50] SCORM http://scorm.com/

4.2.2 Services developing and/or providing access to primarily agriculture-focused resources

This section introduces a range of resources and initiatives which are primarily focused in the area of agriculture. Some may be centres for interoperability development as well as providing access to end user resources.

Agricultural Learning Repositories Task Force – AgLR-TF[51]. The Task Force was set up under the umbrella of FAO's AIMS programme. The aim has been to create a network of organizations that promotes the development of an open and interoperable global infrastructure to facilitate sharing and reuse of learning resources on topics related to agricultural and rural development worldwide.

Organic.Edunet[52]. The Organic.Edunet Web portal provides access to thousands of learning resources on Organic Agriculture, Agroecology and other green topics, including sustainability, ecology, biodiversity, environment and energy. It features a multilingual user interface and provides access to learning resources in various languages. The resources available through the portal are mainly targeted at school level (teachers and pupils) and university level (tutors and students). It also features a vocational training section providing access to related content, aiming mostly at adult/lifelong learning education. The resources are of various types, including reports and guides, handbooks, presentations, web resources (web sites), educational games, experiments, lesson plans etc. The portal provides four different search functions for content retrieval (text-based, tag-based, browse and semantic search) as well as a search mechanism for retrieving competencies. There are different sections of the portal dedicated to school material, vocational education and educational scenarios.

For a repository to participate in the Organic.Edunet network, it needs to follow the OAI-PMH protocol and expose its metadata records in a way that is compliant with the IEEE LOM standard, and more specifically the Organic.Edunet LOM Application Profile[53]. Authors and users can use either Confolio or MOLE as tools to author and manage learning resources. They have been adapted to operate with Organic.Edunet.

CGIAR Learning Resources Center[54]. Learning materials from across the Centres of the CGIAR are searchable within the ARIADNE repository where they are stored. Some resources, which have been developed using Moodle, are available directly on the CGIAR site.

OER in Agriculture[55] is an area of activity within WikiEducator[56]. WikiEducator is an initiative of Wikipedia and the Commonwealth of Learning[57] (CoL). It is a community intended for the

[51] Agricultural Learning Repositories Task Force (AgLR-TF) http://aglr.aua.gr/node/7
[52] Organic.Edunet http://portal.organic-edunet.eu/
[53] Organic.Edunet LOM Application Profile http://project.organic-edunet.eu/organic/files/document/OrganicEdunet_D5.1.2_final.pdf
[54] CGIAR Learning Resources Center http://learning.cgiar.org/
[55] OER in Agriculture http://wikieducator.org/OER_in_Agriculture

collaborative: planning of education projects linked with the development of free content; and development of free content on WikiEducator for e-learning; and work on building open education resources. WikiEducator's technical infrastructure is supported by a financial contribution from CoL to the Open Education Resource Foundation[58].

OER in Agriculture has compiled links to resources (largely from US universities) that are available online and openly licensed. There does not appear to be any interoperability across the recommended resources.

OER University[59] is another initiative under the umbrella of WikiEducator. It is a virtual collaboration of institutions creating pathways for OER learners to gain formal academic credit. The OER University aims to provide free learning to all students worldwide using OER learning materials and for them to gain credible qualifications from recognised education institutions.

AgriLORE[60] is part of a World Bank funded project, the National Agricultural Innovation Project (NAIP), being implemented by ICAR in India. The project has 3 objectives:

- to generate, review, manage and publish approved learning materials for wider use and re-use by distance learning institutions and interested rural and community organizations and extension agencies;
- to build a national pilot repository for digital content on agro-horticulture, for use in distance learning programs aimed at rural learners and extension workers;
- to assess the impact of new methods of ICT and extension approaches on rural livelihoods and on partnerships.

This project aims to be a proof of the value of OER in the extension environment.

There are also many services which provide search and access to resources at National, Regional or Global levels; and there are those which provide access to a single institutional resource. Although these services do not focus specifically on agriculture some of them are listed here because they provide access to *some* agriculture-related resources and therefore could ultimately be of relevance to agINFRA.

MERLOT (Multimedia Educational Resource for Learning and Online Teaching)[61] is a free and open online community of resources designed primarily for faculty, staff and students of higher education from around the world to share their learning materials. It aims to improve the

[56] WikiEducator http://wikieducator.org/Content
[57] Commonwealth of Learning http://en.wikipedia.org/wiki/Commonwealth_of_Learning
[58] Open Educational Resource Foundation http://wikieducator.org/OERF:Home
[59] OER University http://wikieducator.org/OER_university/Home
[60] AgriLORE http://agropedialabs.iitk.ac.in/agrilore/?q=node/575
[61] MERLOT http://www.merlot.org/merlot/index.htm

effectiveness of teaching and learning by providing peer reviewed online learning materials that can be incorporated into faculty designed courses. It currently has 104,000 members. MERLOT carries out federated searches of learning object repositories (LORs), and individual resources, including Connexions, MIT, OER Commons, ARIADNE and NIME. However, a search using the term ‘agriculture’ produced only 144 results.

OER Commons[62]: provides search and access to global resources. It is supported by ISKME (the Institute for the Study of Knowledge Management in Education). It contains a focused area, OER Commons Green, addressing sustainability and resource conservation. A search through OER Commons using ‘agriculture’ produced 428 hits.

OER Africa[63]: has been established by the South African Institute for Distance Education (Saide) to help to drive the development and use of OER across all education sectors on the African continent. Although it is involved in both some technical work (metadata production, search of repositories and other sources) and also collaboration and development of network support for groups in institutions, its’ primary focus is on making educational resources available to users across Africa. A search on ‘agriculture’ produced 140 results.

OER Asia[64]: is an Asian service sharing information, views and opinion, research studies and knowledge resources in addition to guidelines and toolkits on good practices on OER in the Asian region. It is hosted by Wawasan Open University. No resources were discovered when using the search term ‘agriculture’.

CORE (China Open Resources for Education)[65]: The China Open Resources for Education (CORE) is a non-profit organization with a mission to promote closer interaction and open sharing of educational resources between Chinese and international universities. CORE aims to provide Chinese universities with free access to global OERs and correspondingly to make high quality Chinese resources available globally. Currently there are 8 Chinese agricultural universities participating in the programme. The consortium is working now to open up the programme to hundreds more universities.

4.3 Infrastructure and interoperability

This section deals with metadata and search. Some of the services referred to in the previous section have a significant presence here also.

4.3.1 Metadata and Interoperability

The routes to the existence of a learning object or to educational resources are many and varied, and may in some cases even be serendipitous. The object may have had a formal educational

[62] OER Commons http://www.oercommons.org/
[63] OER Africa http://www.oerafrica.org/
[64] OER Asia http://www.oerasia.org/home
[65] CORE http://www.core.org.cn/en/

purpose from the beginning – or not. They may be very granular, they may not. But to be usable, flexible and effective they should have associated metadata that allows interoperability between LOs, repositories and other resources.

Learning Object Metadata (LOM) is a data model, usually encoded in XML, used to describe a learning object and similar digital resources used to support learning. The purpose of learning object metadata is to support the reusability of learning objects, to aid discoverability, and to facilitate their interoperability.

LOM (IEEE) – the IEEE standard: The IEEE 1484.12.1–2002 Standard for Learning Object Metadata is now an internationally-recognised open standard (published by the Institute of Electrical and Electronics Engineers Standards Association, New York) for the description of "learning objects". Relevant attributes of learning objects to be described include: type of object; author; owner; terms of distribution; format; and pedagogical attributes, such as teaching or interaction style.

Nilsson (2008) published a mapping of IEEE LOM into the Dublin Core Abstract Model. This mapping was also used within Organic.Edunet to annotate resources using the repository tool Confolio (Ebner *et al*., 2009). The mapping sets the basis for exposing both Dublin Core and IEEE LOM metadata by using a shared format.

The IEEE LOM provides a skeletal metadata framework which requires the development of an AP for effective implementation. For instance, both GLOBE and FAO have developed their own APs for learning objects (FAO, 2007).

Tzikopoulos *et al*., (2010) in their survey of 59 LORs reported on the application of metadata specification or standards for the description of the learning objects. With regard to the distribution of LORs, most of the examined LORs used either the IEEE LOM (29%) or the Dublin Core (22%) standards. Additionally 25% used IEEE LOM compatible metadata such as IMS Metadata or CanCore.

Among the other more important standards is the Shareable Content Object Reference Model (SCORM).

As OERs are extending to the global level to allow the exchange of metadata and learning objects as well as federating searches, more work is needed to ensure semantic interoperability. Semantic interoperability is related to, for example, vocabularies used to describe learning objects, their intended audiences, topics, and so forth that characteristically serve local needs. Harmonisation of these vocabularies on the local and global level and mapping between different concepts and vocabularies still remain issues for the field.

4.3.1.1 Search and other services (Federations and Harvesting)

Some repositories and services offer searching across multiple repositories through 'federated' searches. For example, ARIADNE and MERLOT cross-search each other's repositories for learning resources, a service that multiplies the availability of resources.

A number of services have been developed with the aim of providing interoperability between repositories and other content aggregators. However, federated search can be difficult to implement and to achieve effective results with. Harvesting is more efficient and manageable. This fact has impacted on the development of many services. Both GLOBE and MERLOT began life using a combination of federated search and harvesting, but are moving toward harvesting, and GLOBE is now harvesting 90% of metadata. GLOBE, Organic.Edunet and LACLO use the ARIADNE harvester.

Apart from efforts on search interoperability, development work has been conducted to connect repositories together in federations such as: EUN Federation of Internet Resources for Education by European Schoolnet[66]; and CORDRA[67]—the content object repository discovery and registration/resolution architecture. GLOBE is at the centre of a federation with founding members such as the ARIADNE Foundation, Education Network Australia (EdNA Online) in Australia, eduSource in Canada, MERLOT in the U.S. and National Institute of Multimedia Education (NIME) in Japan.

4.4 Workflows, Case Studies and Communities

In most areas of research and development, communities are formed around particular subject themes – e.g. crop protection or animal genomics. However, the infrastructure of OERs globally is not always subject-based but is developed around an object with a particular purpose, that of learning. The result is that in this field there are communities for developers which may be separate from the subject-based communities, as we have seen in some of the services introduced above.

How much are these services working as communities, in terms of producing interoperable, openly accessible services which are developing content and infrastructure in collaboration with educators? It would seem to be productive for communities of technical developers to be directly involved with communities that are also developing content for end users – whether subject-based or geographically-based. Organic.Edunet is attempting to do this. Some services which began with a predominantly technical focus are now starting to reposition themselves as communities, such as LACLO.

It does not seem viable to try to imagine workflows starting with any learning object or educational resource in any situation. These objects (however defined) may originate in many different ways. So in this report the starting point for a workflow is taken as the point at which a

[66] European Schoolnet http://fire.eun.org/
[67] CORDRA http://cordra.lsal.cmu.edu/cordra/

resource has been produced by a researcher, developer or educator and is ready to be integrated with or harvested by a site or repository of whatever sort where it can be validated, transformed, 'published' and so on.

It can be useful to think of workflows at four different levels:

- the workflow of an individual or a group developing LOs or larger resources
- institutional workflow, where for instance a university is bringing together workflow activities across a campus(es)
- the workflow of a network or aggregator of educational resource producers
- workflow for presentation at a service level, which brings together aggregations of networks.

Many of the examples shown here, as well as agINFRA itself, are relevant at levels 3 or 4.

A generalized workflow: Figure 4.1 below (from Ternier *et al*., 2010) generalizes the architecture that is currently in place in many learning repository networks. Metadata is gathered from various participating repositories through OAI-PMH or other protocols. Typically, within a network partners contribute metadata for a domain, e.g. organic agriculture. An LOM application profile relevant for the network will have been created. All partners that offer metadata according to this application profile will set up an OAI-PMH target to enable their metadata to be gathered. These partner repositories are represented by repositories on the right hand of the Figure. The metadata harvester is a component in this architecture that periodically checks the partner repositories for new metadata and updates the store of harvested metadata store, in this case using SPI. The store offers access to various search tools. New partner repositories can be added to the network through the registry. The metadata harvester uses this registry to decide which repositories to harvest from.

The workflow shown in Figure 4.2 was developed for the ARIADNE service. The ARIADNE Repository services allow for the management of learning objects in an open architecture and enable stable querying, publishing, and harvesting of learning materials.

The *Registry* service is a catalog service that provides up-to-date information on learning object repositories (LORs). It provides the information necessary for systems to be able to select the appropriate protocols such as OAI-PMH, SQI, SPI, SRU/SRW supported by a given learning object repository. The registry service facilitates interoperability between numerous learning object repositories.

The *harvester* uses the OAI-PMH framework for harvesting metadata instances from an OAI-PMH target and publishes them with the Simple Publishing Interface (SPI). Other services such as GLOBE use this harvester.

The *validation* service provides validation of metadata instances against predefined application profiles, for example based on IEEE LOM. To ensure that only compliant metadata are stored in the ARIADNE repository, the validation service is used to check both the syntactic and semantic validity of the instances used.

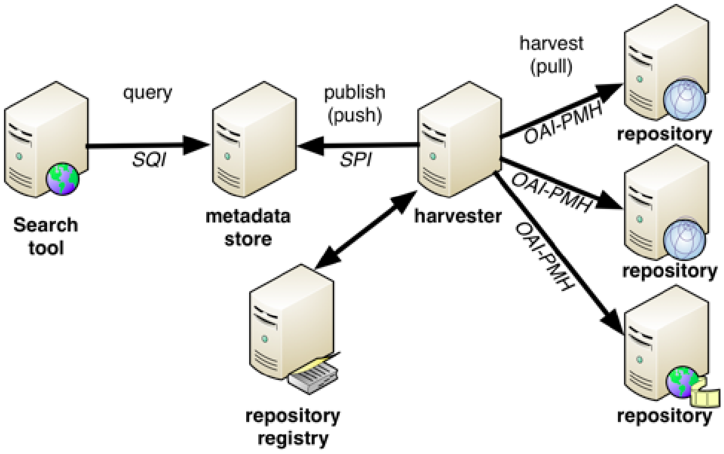


Figure 4.1: Generalized Federation of Repositories (Ternier *et al.*, 2010)

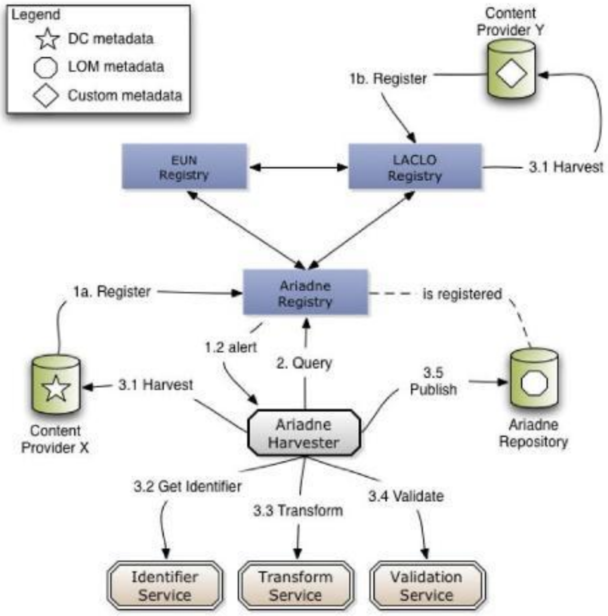


Figure 4.2: Tools and their relationships in the ARIADNE metadata workflow

The *transformation* service converts metadata in one format, e.g Dublin Core (DC), into another format, e.g. the ARIADNE application profile in LOM. This transformation service is needed because of the multiplicity of different metadata schemes that are used in various networks of learning object repositories.

The *Identification* service is used to provide persistent digital identifiers to resources in the ARIADNE infrastructure. The HANDLE system is used as the backend service to create globally unique, persistent and independent identifiers.

Other elements of the service include the *Federated Search Service*, the *Ranking Service*, and the *ALOCOM service* which supports two processes - the disaggregation of learning objects into their components (text fragments, images, definitions, diagrams, tables, examples, audio and video sequences) as well as the automatic assembly of these components in authoring tools.

4.4.1 Organic.Edunet

The Organic.Edunet Web portal[68] provides access to thousands of learning resources. It features a multilingual user interface and provides access to learning resources in various languages. The portal provides four different search functions for content retrieval (text-based, tag-based, browse and semantic search) as well as a search mechanism for retrieving competencies.

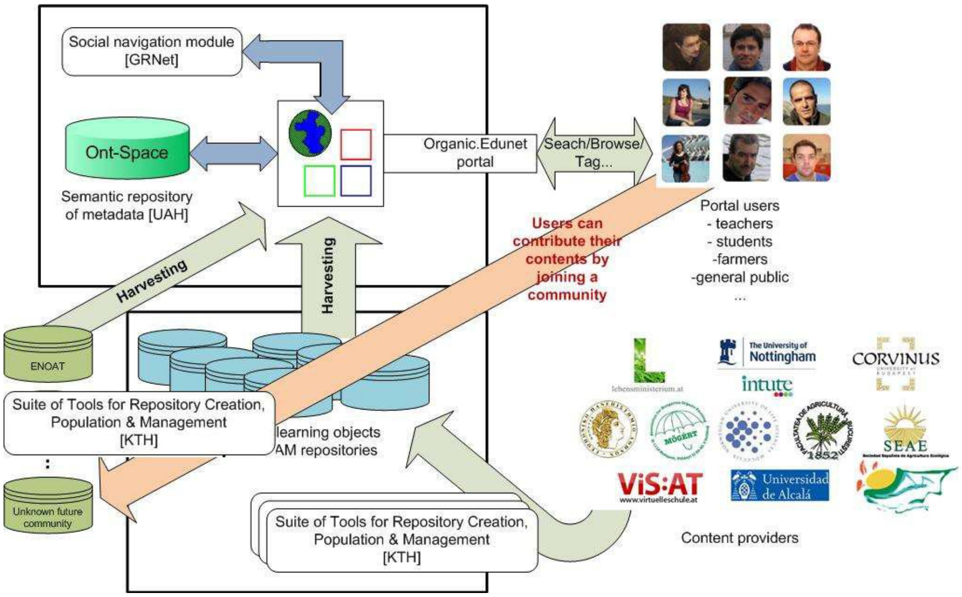

Social navigation module [GRNet]
Ont-Space
Semantic repository of metadata [UAH]
Organic.Edunet portal
Seach/Browse/ Tag...
Portal users
- teachers
- students
-farmers
-general public
...
Users can contribute their contents by joining a community
Harvesting
Harvesting
ENOAT
Suite of Tools for Repository Creation, Population & Management [KTH]
:
Unknown future community
learning objects AM repositories
Suite of Tools for Repository Creation, Population & Management [KTH]
lebensministerium.at
The University of Nottingham
intute
CORVINUS
SEAE
ViS:AT
www.virtuelleschule.at
Universidad de Alcalá
Content providers

Figure 4.3: Workflows and functional relationships in Organic.Edunet

[68] Organic.Edunet Web Portal http://www.organic-edunet.eu/

Figure 4.3 shows a generalised workflow for Organic.Edunet, from the provision or making available of content by providers, through to search and access of resources by end users. An additional function related to the provision of new content by end users is also indicated.

Figure 4.4 is a generalised workflow of Organic.Edunet as presented for Organic.Lingua, a European project developing a multilingual web portal for sustainable agricultural and environmental education. It shows the key functions of metadata and content acquisition, annotation and translation, and quality management.

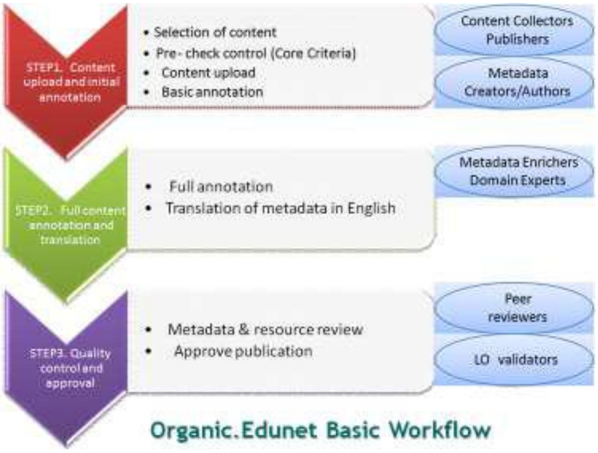

STEP1. Content upload and initial annotation
Selection of content
Pre- check control (Core Criteria)
Content upload
Basic annotation
Content Collectors Publishers
Metadata Creators/Authors
STEP2. Full content annotation and translation
Full annotation
Translation of metadata in English
Metadata Enrichers Domain Experts
STEP3. Quality control and approval
Metadata & resource review
Approve publication
Peer reviewers
LO validators
Organic.Edunet Basic Workflow

Figure 4.4: Basic workflow of Organic.Edunet

The figure below shows the workflow of the ingestion of XML files into the Organic.Edunet service, depending on whether or not the provider's process supports OAI-PMH.

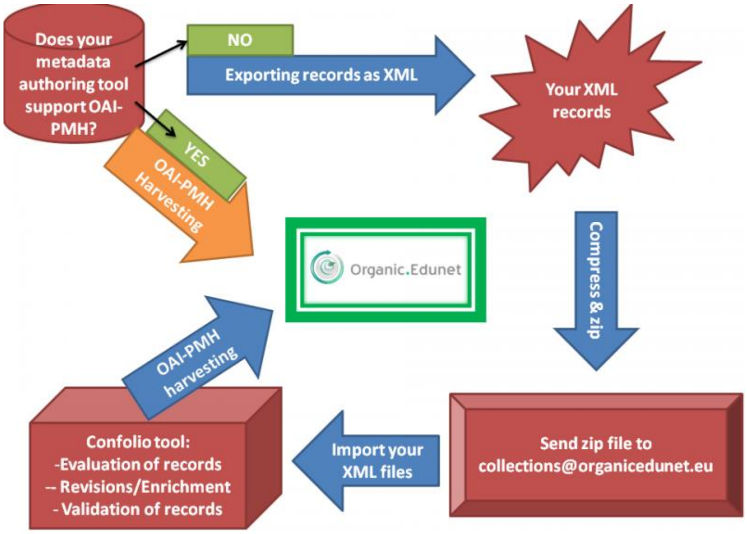


Figure 4.5: Organic.Edunet ingestion of XML files

Then in Figure 4.6 the workflow shows specifically the harvesting of metadata from repositories that make their metadata available through OAI-PMH. Organic.Edunet uses the ARIADNE harvesting tool[69] for this process. The metadata are published into a new repository using the Simple Publishing Interface (SPI)[70]. The metadata records available through the Organic.Edunet Web portal are created using the Organic.Edunet Metadata Application Profile[71], which is an adapted version of the IEEE Learning Object Metadata (LOM)[72].

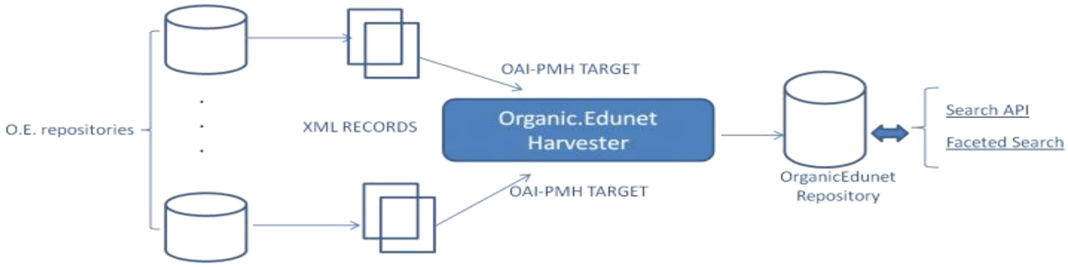


Figure 4.6: Metadata harvesting into the Organic.Edunet repository

4.4.2 AgLR Task Force (AgLR-TF)

The aim of the AgLR-TF is to create a network of organizations that promotes the development of an open and interoperable global infrastructure to facilitate sharing and reuse of learning resources on topics related to agricultural and rural development worldwide. Discussions are already taking place on how AgLR might provide key components of the agINFRA infrastructure. The basic components of AgLR are:

- an interface to search/browse through metadata descriptions of agricultural learning resources;
- the backend: providing periodical harvesting of metadata from various providers - through OAI-PMH protocol/targets, using mainly IEEE LOM-based metadata (but also some DC-based), and indexing metadata element values to use for search/browse.

The backend is based on ARIADNE infrastructure. The main software services for metadata aggregation are based on the Open Source Software of ARIADNE:

- Validation service providing automatic validation of metadata against predefined application profiles, e.g. IEEE LOM.

[69] ARIADNE harvesting tool http://ariadne.cs.kuleuven.be/lomi/index.php/Harvesting_Metadata#Harvesting_Tool
[70] Simple Publishing Interface (SPI) http://www.dlib.org/dlib/september10/ternier/09ternier.html
[71] Organic.Edunet Metadata Application Profile
http://wiki.agroknow.gr/organic_edunet/index.php/Organic.Edunet_Metadata_Application_Profile
[72] IEEE Learning Object Metadata (LOM) http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

- Repository service providing management of metadata records, open standards and specifications support (SQI, SPI, OAI-PMH, RSS), metadata schema support (IEEE LTSC LOM by default, or any schema with an XML binding), indexing, and an open source licence.
- Registry service. A catalogue service which manages up-to-date information on metadata providers.
- Harvester service.

AgLR is operating the steps taking place at the backend: OAI-PMH targets are first *validated* against some compliant metadata schema; if validated, the OAI-PMH target is included in a *registry* of AgLR providers; targets in the registry are *harvested* periodically and if needed are *transformed* into a core schema - invoking the Repository Service, the Harvester Service and the Transformation Service; the Web interface *indexes* selected elements to create facets and present them, invoking the *Finder Service*. These different components are shown in Figure 4.7 below in the context of potential agINFRA infrastructure layers.

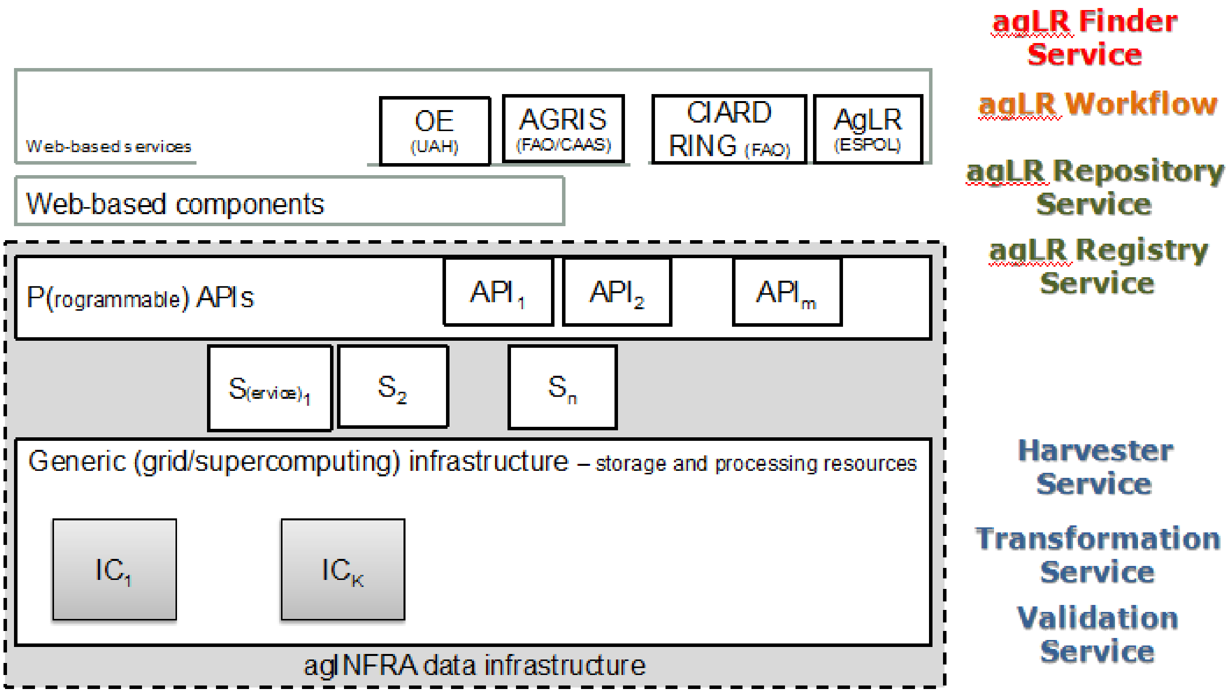


Figure 4.7: Allocation of different components invoked in AgLR workflow over the infrastructure layers

An example workflow is also shown below for AgLR metadata aggregation.

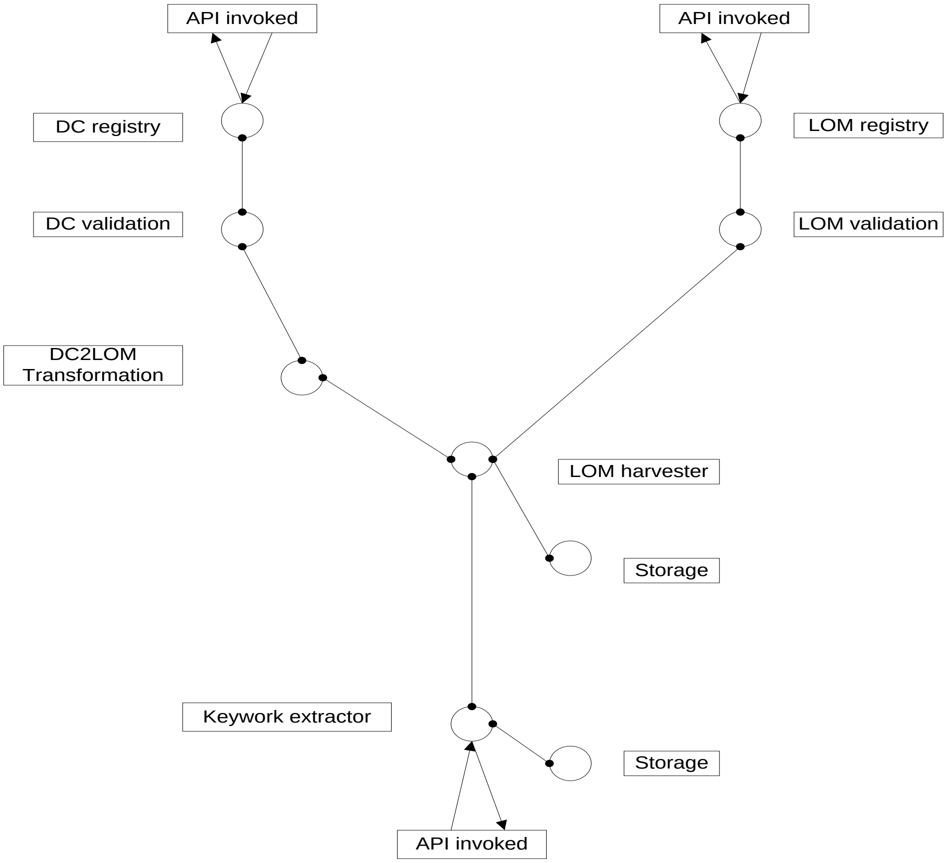


Figure 4.8: An example workflow for AgLR metadata aggregation

Other services:

There are particular problems associated with working in multilingual environments. The Organic.Lingua[73] project, referred to above in this report, is working on the management of multilingual content and associated metadata. The focus is on how a federated service such as Organic.Edunet can align to the different processes of a variety of content and metadata providers. Metadata workflows are available for a variety of services, including services of the Institut National de la Recherche Agronomique (INRA).

[73] Organic.Lingua http://www.organic-lingua.eu/

4.5 Implications for the agINFRA infrastructure

4.5.1 Interoperability

There are many learning resources available on sites that make no mention of metadata or interoperability. It is the view of one expert in the field that as many as 90% of online resources are not exposing metadata and are not interoperable. Within universities and other organizations resources are often stored for local use only.

The main barriers to effective interoperability across global resources are standardization and quality of metadata and APs (Manouselis *et al*.,2010). However, a positive feature noted by these authors in their analysis of LORs is that the IEEE LOM standard and Dublin Core are pervasive across the field. Perhaps one could conclude that the standards are in place and in principle interoperability should be a relatively trivial issue? Unfortunately there are many resources which are not participating in this interoperable environment.

As OERs develop globally semantic interoperability is seen as the required next step. Semantic interoperability is related to, for example, vocabularies used to describe learning objects, their intended audiences, topics, and so on, that characteristically serve local needs. Harmonisation of these vocabularies on the local and global level, and mapping between different concepts and vocabularies, is a major challenge here. Semantic interoperability becomes more important the more global the aggregation services become. Both IMS Global Learning Consortium and ISO SC36 are working actively in this area. IMS Global has created the Vocabulary Definition Exchange (VDEX)[74] specification that defines a grammar for the exchange of value lists of various classes, that is, vocabularies.

Although work is progressing in the OER field on Linked Open Data (LOD) and semantic interoperability (Sicilia *et al*., 2011) they are not yet adopted on a scale which will create significant impact in the short term. Nevertheless, LOD and the Semantic Web could have major benefits for OER in the future, particularly where the coverage of metadata is poor.

4.5.2 Communities

Organic.Edunet, ARIADNE, AgLR-TF and GLOBE have positioned themselves to be at the centre of technical development communities. Organic.Edunet is relatively unusual in bringing together a technical development community with an educator/user community. GLOBE is also now repositioning itself in this way, though its coverage of subject areas is general, as opposed to Organic.Edunet's focus. It seems that there are real benefits to be gained from bringing together educators (the creators of educational resources), users, and technical developers. In a community of practice of this sort the needs of the user (the ultimate beneficiary) are more likely to be met. (See also Geser, 2012.)

[74] Vocabulary Definition Exchange (VDEX) http://www.imsproject.org/vdex/

4.5.3 Other issues

Ownership (IPR) of objects or OERs is no greater an issue than it is in any other field of authorship or creativity. It is true that the move toward 'openness' on the internet does create nervousness among some researchers and authors. They feel that their property/creation is much more at risk than in a more closed environment. However, this is the world of the future and ownership in this open context can be asserted and to some extent protected through the use of Creative Commons licences (Hylen, 2007).

Quality management is an important issue that becomes more important as the scale of interoperability and global access grow (Stracke *et al*., 2007). This is a large topic in its own right and this report is not the right place to start to address it. However, the Commonwealth of Learning[75] has a microsite addressing this area.

[75] Commonwealth of Learning http://www.col.org/resources/micrositeQA/Pages/default.aspx

5 Geospatial information systems (GIS)

5.1 Background

The volume of scientific data being generated by highly instrumented research systems (sensor networks, satellites, seismographs, etc.) is so great that it can be captured and managed only with the use of information technology. The need to manage very large volumes of data is one of the main drivers of e-Science and cyberinfrastructure developments. If these data can be stored in reusable forms, they can be shared over distributed networks. Data are becoming an important end product of scholarship, complementing the traditional role of publications.

The management of geospatial data gains great benefits from an infrastructure that can support geospatial data processing within and across scientific domains. Geospatial Cyberinfrastructure (GCI) or GIS refer to infrastructure that supports the collection, management, and utilization of geospatial data, information, and knowledge for multiple scientific domains (Yang *et al*., 2010). Much progress has been made in defining standards by the Open Geospatial Consortium (OGC) and the International Organization for Standardization (ISO). Also, in 2007, the Infrastructure for Spatial Information in the European Community (INSPIRE) directive entered into force and laid down a general framework for a Spatial Data Infrastructure (SDI) to support European Community environmental policies and activities.

The use of geospatial information to address agricultural issues is growing. Geospatial information is used, for example, in: precision agriculture, remote sensing and geographic information systems, finding the best location for new enterprises, predicting potential threats from weeds, pests and diseases, using airborne geophysics for salinity management, soil and landscape assessment, and so on. There are many opportunities where the use of geospatial information can result in better decision-making that will lead to higher productivity, reduced costs and reduced environmental impacts.

5.2 A general summary of the current GIS landscape

Since 2000, many organizations collecting, processing, managing, disseminating and using geospatial information have increasingly moved towards integrating Internet services into their operational environment. Wireless and mobile applications, location-based products, services, and solutions initiated at the start of the new millennium with the promise of an increasing need for locational functionality via the Internet by not just the geographic community, but the world at large.

5.2.1 GIS resources in the current landscape

GIS includes many different categories of resources within a flexible framework. This framework can be seen to have three elements:

(1) *Functions*, which include both generic cyberinfrastructure functions (computing, networking, and hardware) and those that are geospatial-specific. These functions include the following:

> (a) a middleware layer to bridge geospatial functions and resource management, monitoring, scheduling, and other system-level functions;
> (b) a geospatial information integration layer to integrate geospatial data, information, and knowledge flow as supported by observations, geospatial processing, and knowledge mining; and
> (c) geospatial functions to provide various analytical functions for end-users.

(2) *The community* represents the virtual organizations and end user interactions within specific communities including geographic, environmental, Earth, and other science domains.
This dimension also provides feedback channels for knowledge collection functions to leverage scientific community and citizen participation.

(3) *Enabling technologies* provide technological support to invent, mature, and maintain all GIS functions, such as collecting data through observations and collecting and utilizing knowledge through a semantic web. The architecture and integration of GIS benefits from numerous enabling technologies, many of which contributed to the birth of GIS, for instance: Earth observation and sensor networks, SDI (Spatial Data Infrastructure), Distributed geographic information processing (DGIP), and so on.

5.2.2 GIS Standards

The definition and development of standards in the GIS field are dominated by two organizations, the ISO and the Open Geospatial Consortium. These standards are highly pervasive globally and their role and importance is summarised here.

5.2.2.1 ISO/TC 211 Geographic information

ISO/TC 211[76] (and the Open GIS Consortium – now called the Open Geospatial Consortium (OGC) – see below) emerged by the mid-1990s with GIS standards becoming a highly visible and prominent part of the international geographic agenda. The scope of ISO/TC 211 is standardization in the field of digital geographic information. It aims to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth.

In general, the OGC develops software interface specifications, while ISO/TC 211 develops geographic data standards. ISO/TC 211 has a programme of work that includes the concurrent development of an integrated set of twenty standards for geographic information. These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing,

[76] ISO/TC 211 http://www.isotc211.org/

presenting and transferring such data in digital/electronic form between different users, systems and locations.

5.2.2.2 Open Geospatial Consortium (OGC)

The Open Geospatial Consortium (OGC)[77] is an international industry consortium of 454 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards. OGC® Standards[78] support interoperable solutions that "geo-enable" the Web, wireless and location-based services and mainstream IT. The standards empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications. For instance, the OGC Web Map Server interface specification has been commercially implemented by over 130 of the GIS industry's 200 software vendors. Under the cooperative agreement between the OGC and ISO, the Web Map server interface (ISO 19128) is now being progressed as an International Standard within ISO/TC 211.

5.3 Some services and initiatives in GIS

Standards are very coherent across the GIS field, as are GIS's core practitioners who came initially from satellite-related areas such as remote sensing. The field is growing rapidly in terms of its range of applications. This section indicates some key initiatives, both general in their fields of relevance and also those related particularly to agriculture.

5.3.1 GeoNetwork (FAO)

GeoNetwork's purpose[79], particularly for agriculture related applications, is: to improve access to and integrate use of spatial data and information; to support decision making; to promote multidisciplinary approaches to sustainable development; and to enhance understanding of the benefits of geographic information.

GeoNetwork is a catalog application to manage spatially referenced resources. It provides metadata editing and search functions as well as an embedded interactive web map viewer. It is currently used in numerous Spatial Data Infrastructure initiatives across the world. GeoNetwork has been developed on the principles of Free and Open Source Software (FOSS) and international and open standards for services and protocols (from ISO/TC211 and OGC). GeoNetwork opensource allows users to share geographically referenced thematic information between different organizations. Its purpose has been specifically to focus on interoperability.

GeoNetwork provides:

- Search access to local and distributed geospatial catalogues
- Up- and downloading of data, graphics, documents, pdf files and any other content type

[77] Open Geospatial Consortium (OGC) http://www.opengeospatial.org/
[78] OGC Standards http://www.opengeospatial.org/standards/
[79] GeoNetwork http://www.fao.org/geonetwork/srv/en/main.home

- An interactive Web Map Viewer to combine Web Map Services from distributed servers around the world
- Online editing of metadata with a template system
- Scheduled harvesting and synchronization of metadata between distributed catalogs
- Support for OGC-CSW 2.0.2 ISO Profile, OAI-PMH, Z39.50 protocols
- Fine-grained access control with group and user management
- A multilingual user interface.

5.3.2 United Nations Spatial Data Infrastructure (UNSDI) Centre of Excellence (CoE)

Over 30 UN Organizations participating in the United Nations Geographic Information Working Group (UNGIWG)[80] identified the need for an international and coordinated approach in order to obtain geographical information required for addressing and solving global issues. To this end they have initiated UNSDI, supported by a number of international partners. The UNSDI's Centre of Excellence is based in the Netherlands. Specific concerns of UNSDI are:

- The identification of common geo-information needs (core geo-data sets) worldwide
- The manufacturing and / or increased accessibility of these core geo-data sets
- The identification of gaps in required core geo-data sets.

As part of the UNSDI's ClearSite Project, the Geospatial Data Warehouse (GDW) is led by FAO and is currently being initiated. The GDW aims to:

- strengthen and extend the network of geospatial information management
- implement standardized geospatial data-sharing practices and provide a common software platform based on open standards
- provide the hosting foundation for a visualization facility as well as a centrally accessible data repository for agency-produced or procured geospatial content such as maps, GIS data, remote sensing imagery, Global Navigation Satellite System logs, crowd-sourced data and other geo-referenced information

The GDW will result from the dynamic composition of different software components grouped into a few functional areas. The IT backbone of the GDW will be available to all UN agencies.

5.3.3 The INSPIRE Directive

The INSPIRE Directive[81] came into force on 15 May 2007 and will be implemented in various stages, with full implementation required by 2019. Through the implementation of a framework of technical standards it aims to create a European Union (EU) spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations

[80] United Nations Geographic Information Working Group (UNGIWG) http://www.unsdi.nl/

[81] INSPIRE Directive http://inspire.jrc.ec.europa.eu/

and better facilitate public access to spatial information across Europe. INSPIRE is based upon the ISO and OGC standards.

INSPIRE is based on a number of common principles:

- Data should be collected only once and kept where it can be maintained most effectively.
- It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
- It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.
- Geographic information needed for good governance at all levels should be readily and transparently available.

The INSPIRE Geoportal will provide the means to search for spatial data sets and spatial data services, and subject to access restrictions, to view spatial data sets from the EU Member States within the framework of the INSPIRE Directive.

5.3.4 The Open Source Geospatial Foundation (OSGeo)

OSGeo[82] is a not-for-profit organization whose mission is to support the collaborative development of open source geospatial software, and promote its widespread use. The Foundation provides financial, organizational and legal support to the broader open source geospatial community. It also serves as an independent legal entity to which community members can contribute code, funding and other resources. OSGeo also serves as an outreach and advocacy organization for the open source geospatial community, and provides a common forum and shared infrastructure for improving cross-project collaboration.

OSGeo as a community promotes interaction between users, developers, and community participants. It provides links to events, documentation, websites, and other information of interest to the open source web mapping community.

Under the umbrella of the above is the Geospatial Data Abstraction Library (GDAL), which is a translator library for raster geospatial data formats that is released under an X/MIT[83] style Open Source license by the Open Source Geospatial Foundation[84].

5.3.5 Global Spatial Data Infrastructure Association (GSDI)

The GSDI Association[85] is an inclusive organization of organizations, agencies, firms, and individuals from around the world. The purpose of the organization is to promote international cooperation and collaboration in support of local, national and international spatial data infrastructure

[82] OSGeo http://inspire-geoportal.ec.europa.eu/
[83] X/MIT http://trac.osgeo.org/gdal/wiki/FAQGeneral#WhatlicensedoesGDALOGRuse
[84] Open Source Geospatial Foundation http://www.osgeo.org/
[85] GSDI Association http://www.gsdi.org/

developments that will allow nations to better address social, economic, and environmental issues of pressing importance.

The Joint Board of Geospatial Information Societies (JB GIS)[86] is a coalition of leading international geospatial societies which can speak on behalf of the geospatial profession at international level, especially to the United Nations and other global stakeholders. Its' second goal is to coordinate activities within the geospatial society and organisations. The JB GIS is a co-operation network and there are no obligations to the membership. The current members of the JB GIS are:

- Global Spatial Data Infrastructure (GSDI) Association[87]
- IEEE Geoscience and Remote Sensing Society (IEEE-GRSS)[88]
- International Association of Geodesy (IAG)[89]
- International Cartographic Association (ICA)[90]
- International Federation of Surveyors (FIG)[91]
- International Geographic Union (IGU)[92]
- International Hydrographic Organization (IHO)[93]
- International Map Trade Association (IMTA)[94]
- International Society of Photogrammetry and Remote Sensing (ISPRS)[95]
- International Steering Committee for Global Mapping (ISCGM)[96]

5.3.6 Geographic Data Portals and Repositories (linked to from GSDI)

Numerous online facilities provide access to a variety of geographic data sources, both commercial and freely available, but availability and access vary considerably from nation to nation. There is no single meta portal that leads to an all-encompassing listing of geographic data offerings, but a few are: UNGIWG Data Links[97]; CIESIN World Data Center[98]; UNEP Geodata Portal[99]; Geospatial One Stop[100]; The Geography Network[101]; and GWSP Digital Water Atlas[102].

[86] JB GIS http://www.fig.net/jbgis/
[87] GSDI http://www.gsdi.org/
[88] IEEE-GRSS http://www.grss-ieee.org/
[89] IAG http://www.iag-aig.org/
[90] ICA http://icaci.org/
[91] FIG http://www.fig.net/jbgis/
[92] IGU http://www.igu-online.org/site/
[93] IHO http://www.iho.int/srv1/
[94] IMTA http://www.imtamaps.org/
[95] ISPRS http://www.isprs.org
[96] ISCGM http://www.iscgm.org/
[97] UNGIWG Data Links http://www.ungiwg.org/data.htm
[98] CIESIN World Data Center http://sedac.ciesin.columbia.edu/wdc/
[99] UNEP Geodata Portal http://geodata.grid.unep.ch/
[100] Geospatial One Stop http://gos2.geodata.gov/wps/portal/gos
[101] The Geography Network http://www.geographynetwork.com/
[102] GWSP Digital Water Atlas http://atlas.gwsp.org/

It should also be noted that DLESE, the Digital Library for Earth System Education[103], is a community which brings together learning resources and geospatial information. It is described in more detail in the educational resources part of this report.

5.4 Communities, life cycles and workflows in GIS

5.4.1 Communities

The established world of geospatial information is more centralised in its' focus on a core set of standards and interoperability mechanisms than most other fields. This might be related to the fact that it is a relatively small field in terms of the numbers of researchers involved and the focus of their work. The fields which use geospatial information as an adjunct to their work, such as environmental sensor networks or the development of agricultural practices, are of course more diverse. As a result of this, communities of geospatial research will tend to be based around core GIS practitioners. Otherwise communities develop around areas of, for instance, agriculture or ecology where GIS is becoming important as a tool to further research.

For the core GIS researchers a number of services or organizations have created a community environment. The Open Source Geospatial Foundation (OSGeo) promotes interaction between users, developers, and community participants. It provides links to events, documentation, websites, and other information of interest to the open source web mapping community. OSGeo aims to streamline the coordination of community development efforts, which it sees as crucial to the success of open source web mapping.

The INSPIRE Geoportal also aims to create a community of geospatial information researchers and developers, though there is as yet unclear evidence for this.

The Joint Board of Geospatial Information Societies (JB GIS) is a coalition of leading international geospatial societies which can speak on behalf of the geospatial profession at international level, especially to the United Nations and other global stakeholders. Its' second goal is to coordinate activities within the geospatial society and organisations.

There is also the International Geospatial Society which is under the umbrella of the Global Spatial Data Infrastructure Association (GSDI).

5.4.2 Life Cycles

Generalised approaches to data life cycles and workflows are hampered by the fact that each service, each research programme, perhaps even each experiment, has its own individual requirements. Nevertheless, it might be helpful to look at some generalised data lifecycles developed in the area of embedded sensor networks, a growing area of activity for GIS (Borgman *et al.*,2007). The example shown below proposes eight stages that are common to scientific data. The order of the steps is not absolute, as some stages are iterative.

[103] Digital Library for Earth System Education http://www.dlese.org/library/index.jsp

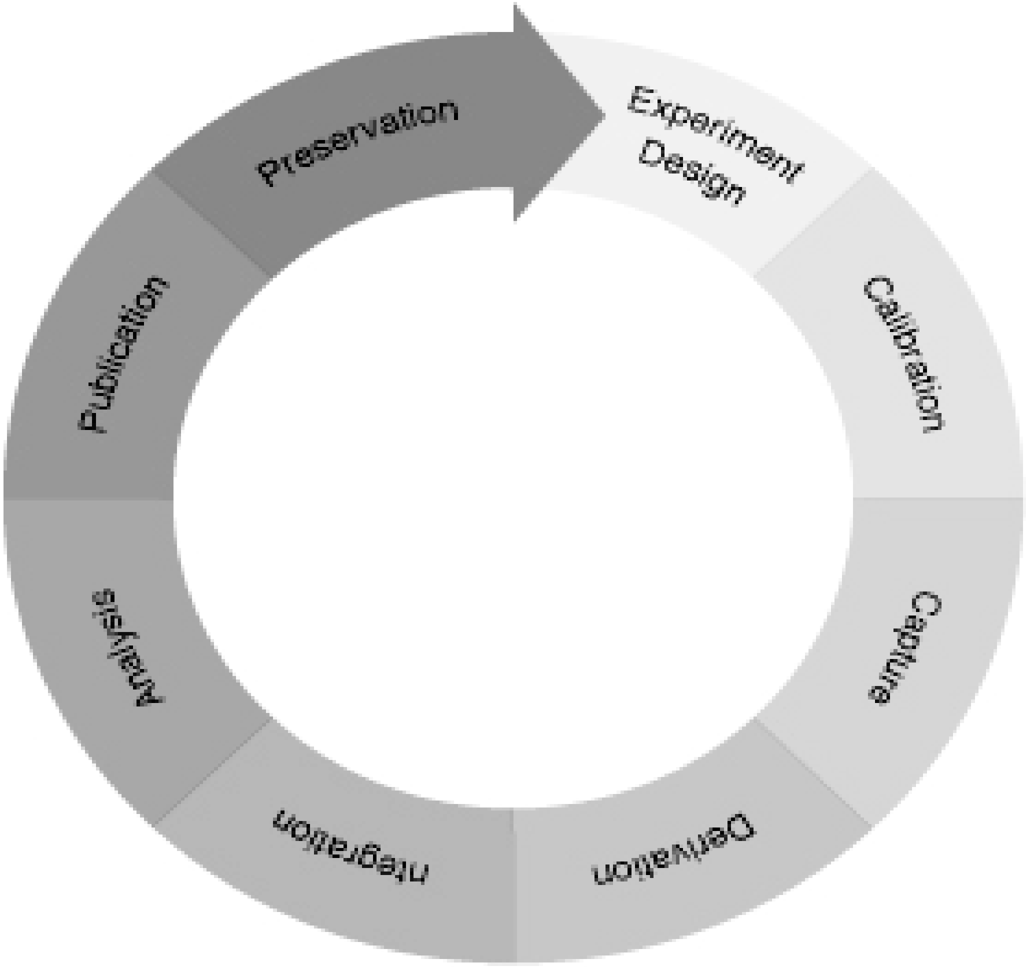


Figure 5.1: Data life cycle of scientists working on embedded sensor networks

The figure below (Pepe *et al*., 2010) integrates the whole life cycle of environmental sensing research data, as shown above, with more detailed cycles within each activity. The inner circle represents the life cycle of scientific research in environmental sensing.

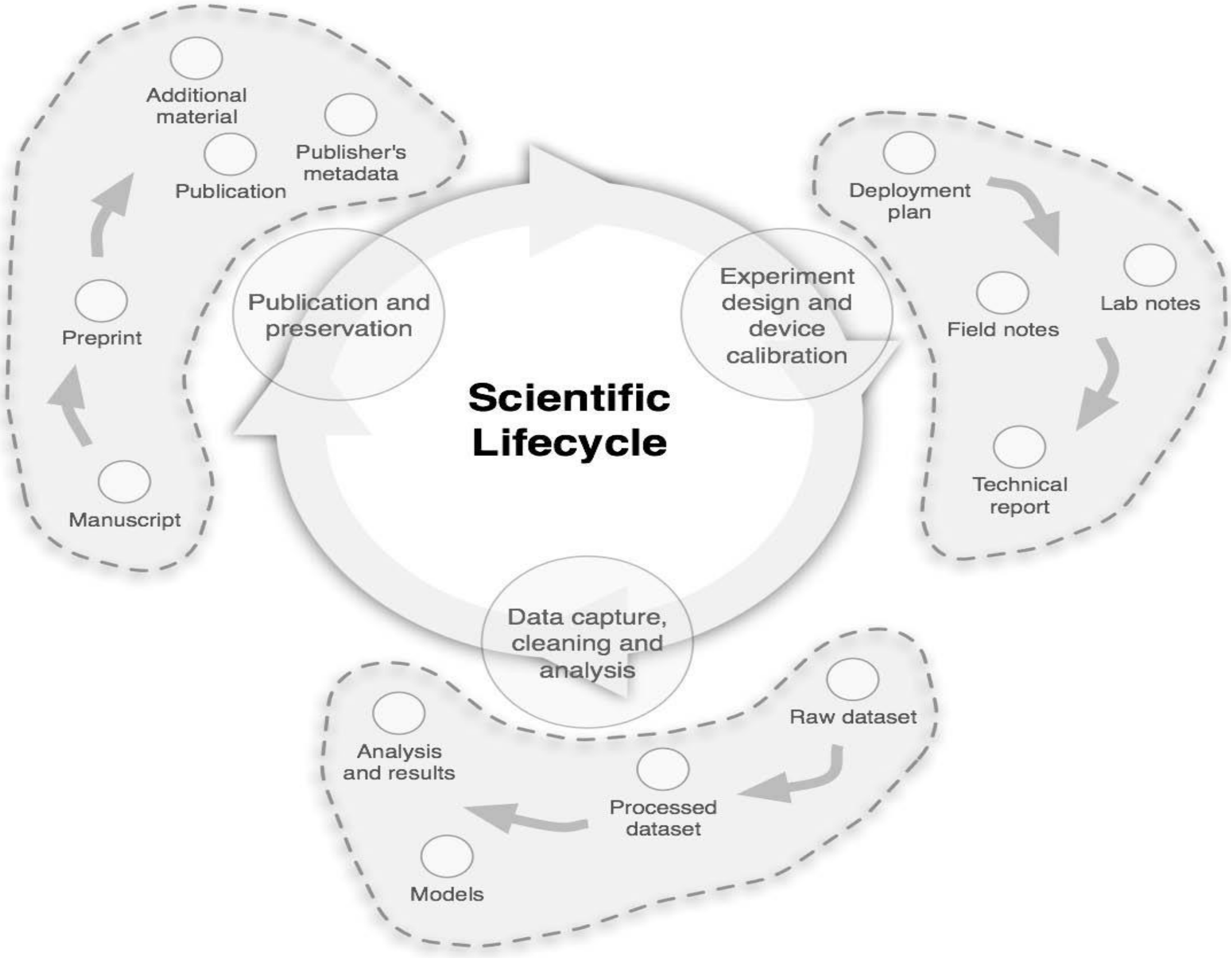


Figure 5.2 The integrated scientific life cycle of embedded networked sensor research (Pepe *et al*.,2010)

It is worth reiterating that the integrated life cycle presented in Figure 5.2 is based on the social, cultural, academic practices, and workflows of a specific scientific domain: embedded networked sensing research. Clearly, life cycles will vary by type of scientific practice, from laboratory to field, by research methods, and by research questions. The volumes of data being produced by embedded sensor networks and other scientific technologies are transforming the field research methods of the environmental sciences. For this community, data that accumulate in ad-hoc computer files on individual and communal servers cannot easily be leveraged for analysis. Improved levels of interoperability between digital objects will not only improve the reuse and long-term preservation of sensor data, but also augment the quality and extent of scholarly communication of the disciplines.

5.4.3 Workflows: ‘Big’ and ‘Small’ Science

In the GIS arena the range of research environments making use of geospatial data has been growing. Studies of scientific data practices have indicated that in only a few fields do researchers predictably contribute their data to shared repositories (Zimmerman, 2007). ‘Big’ science (such as satellite imaging, particle physics and astronomy) has a history of doing so. However, there are areas of geospatial information, such as environmental sensor networks, that are more recent entrants to this environment of managing large data volumes. Repositories often do not offer the tools and services that these scientists appear to need, such as the ability to store data for personal analysis and use. The majority of scientific researchers, who are not established in the ‘big ‘science environment, save data and reuse those data when applicable to future research. (Borgman *et al*., 2007)

5.4.3.1 Some examples of developments in ‘small’ science

Scientists faced with these growing volumes of data, and the need to perform complex calculations on them in distributed and collaborative environments, are turning to the concept of scientific workflows. Geospatially enabled scientific workflows are addressed by well-known tools such as SEXTANTE[104] in the FOSS4G world. These tools, however, are focused on core GIS while many scientists are simply using GIS as an adjunct to their investigations, implying the need to introduce geospatial functionality into generic scientific workflow environments such as Kepler[105].

Kepler: is designed to help scientists, analysts, and computer programmers create, execute, and share models and analyses across a broad range of scientific and engineering disciplines. Kepler can operate on data stored in a variety of formats, locally and over the internet. Using Kepler's graphical user interface, users select and then connect relevant analytical components and data sources to create a "scientific workflow"—an executable representation of the steps required to generate results. For example, Kepler has been used in the REAP (Realtime Environment for Analytical Processing) project[106] to facilitate the quantitative evaluation of sea surface temperature datasets. See Figure 5.3 below.

Multisite agricultural trial database for climate change analysis (agtrials.org): The online database developed at agtrials.org is the development platform for the CGIAR research programme on Climate Change and Food Security (CCAFS) Global Trial Sites Initiative. It shows the result of discussions between plant breeders running the agricultural trials and the geographers from a spatial data background. Agtrials.org is a development organised through the community working within the CCAFS and emphasises a pragmatic approach to the collection of metadata and data which reflects the realities of the diverse research environments involved.

[104] SEXTANTE http://www.sextantegis.com/
[105] Kepler https://kepler-project.org/
[106] REAP http://reap.ecoinformatics.org/Wiki.jsp%3Fpage=WelcomeToREAP.html

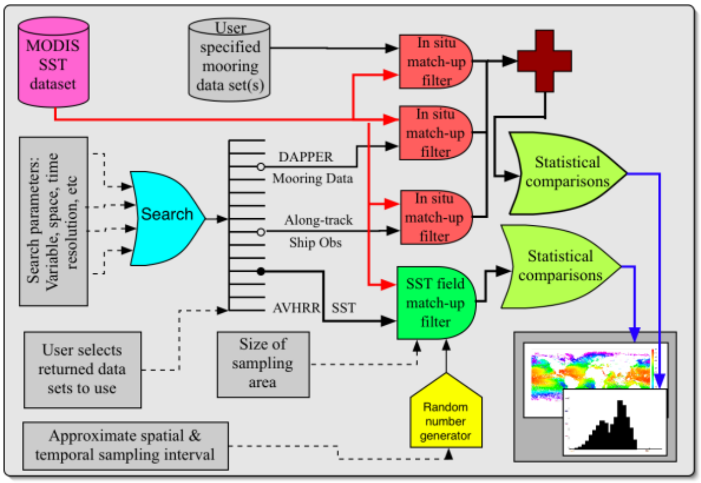


Figure 5.3: Kepler workflow application to facilitate the quantitative evaluation of sea surface temperature data sets (REAP project)

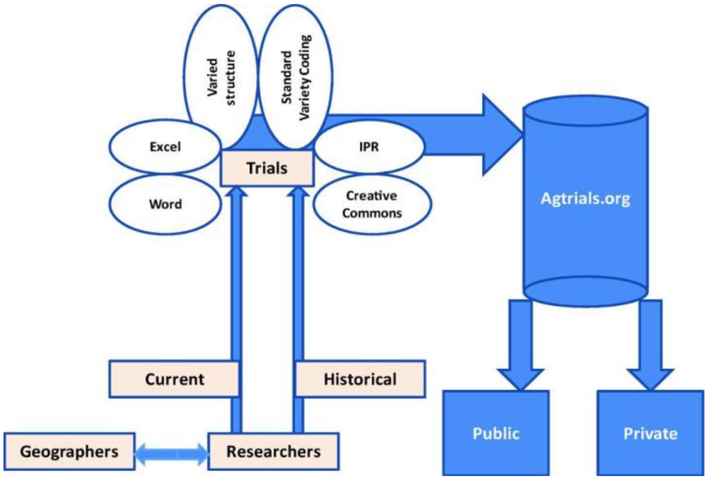
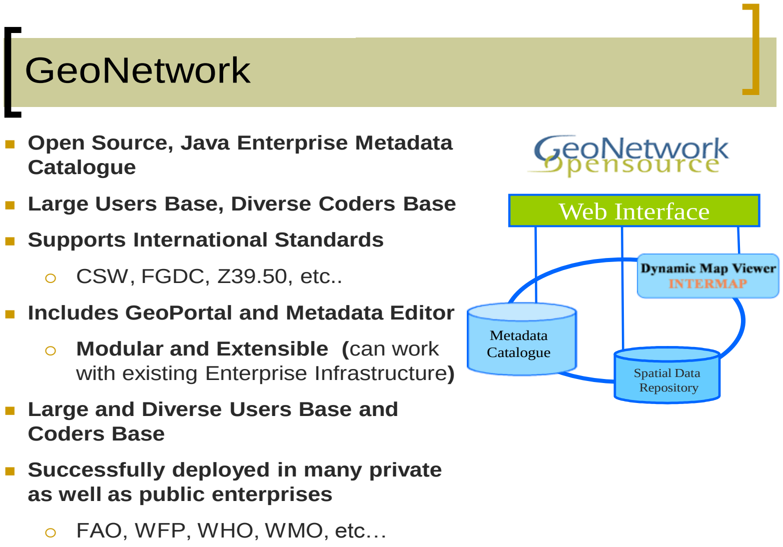


Figure 5.4: General workflow of the agtrials.org repository

A series of trials were identified which could be easily incorporated into the database with emphasis on what was possible within existing time and resource constraints. The application development focused on providing a data repository application where users could easily load historical trial metadata and information on current trials within the CCAFS programme. It needed to provide both private and public access. It built on experience on previous systems which were purely location based and incorporates the requirements of the plant breeders. Data is provided in a variety of formats and development of the application is continuing to accommodate the design of the database and metadatabase, which can cope with the different types of user. Researchers also provide, where available, information on weather conditions during the trial and soil characteristics.

5.4.3.2 Some Developments in 'Big' Science

GeoNetwork[107] (FAO). The GeoNetwork site is powered by GeoNetwork opensource[108]. FAO and World Food Program (WFP), United Nations Environment Programme (UNEP) and more recently Office for the Coordination of Humanitarian Affairs (OCHA), have combined their research and mapping expertise to develop GeoNetwork opensource as a common strategy to share their spatial databases, including digital maps, satellite images and related statistics. The three agencies make extensive use of computer-based data visualization tools, using GIS and Remote Sensing (RS) software mostly to create maps that combine various layers of information. GeoNetwork opensource provides them with the capacity to access a wide selection of maps and other spatial information stored in different databases around the world through a single entry point.



GeoNetwork

- **Open Source, Java Enterprise Metadata Catalogue**
- **Large Users Base, Diverse Coders Base**
- **Supports International Standards**
  - CSW, FGDC, Z39.50, etc..
- **Includes GeoPortal and Metadata Editor**
  - **Modular and Extensible** (can work with existing Enterprise Infrastructure**)**
- **Large and Diverse Users Base and Coders Base**
- **Successfully deployed in many private as well as public enterprises**
  - FAO, WFP, WHO, WMO, etc…

Figure 5.5: Key features of GeoNetwork

GeoNetwork opensource is a part of the current UNSDI development. Figure 5.5 shows an outline of GeoNetwork's features in this context.

The UNSDI (UN Spatial Data Infrastructure) project aims to create an infrastructure and increase the accessibility of existing and new geoinformation worldwide. Humanitarian response, economic development, environmental protection, peace, safety and security, but above all the threat of food and water shortages, requires a well-coordinated international approach. Geo-information is to be used in managing and monitoring of development processes. The initiative has establish a UN Centre of Excellence with the following tasks:

- Organizational and technical infrastructure development (ICT tools on the Internet and in users hands through mobile phones)

[107] GeoNetwork http://www.fao.org/geonetwork/srv/en/main.home
[108] GeoNetwork opensource http://geonetwork-opensource.org/

- Coordination of production of and access to core geo-data sets.

Working within this framework the ClearSite project:

- Aims to strengthen and extend network of geospatial information management.
- Implements standardized geospatial data-sharing practices and provides a common software platform based on open standards.
- Provides the hosting foundation for a Visualization Facility as well as a centrally accessible data repository for agency-produced or procured geospatial content such as maps, GIS data, remote sensing imagery, Global Navigation Satellite System logs, crowd-sourced data and other geo-referenced information.

Figure 5.6 shows ClearSite's geospatial data warehouse components.

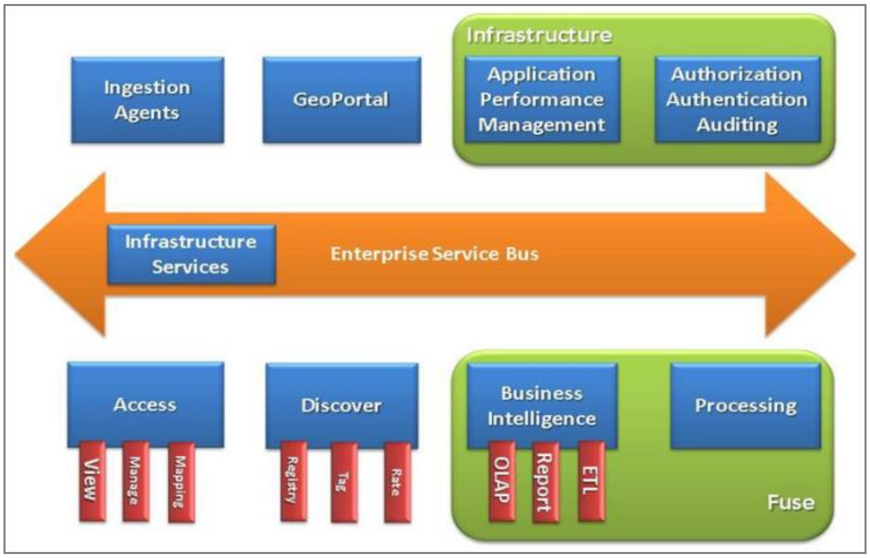


Figure 5.6: Geospatial Data Warehouse components and the Service BUS

5.5 Implications for the agINFRA infrastructure

It can be argued that the interoperability standards for GIS are widely accepted and used in the GIS arena. Interoperability is critical to the integration process and needs to be maintained at the data and information levels to avoid building silos. To support place-based policy and other national and international initiatives, various GIS's should be integrated into a global system. Standards organizations (e.g. the OGC and the ISO/TC 211) have increasingly led to cross-cutting interoperable specifications and prototypes. This recent trend is helping to advance sharing and interoperability within and across disciplines. There are also continuing developments within ISO, such as the recent part of ISO 19144 which specifies a Land Cover Meta Language (LCML) (ISO, 2010). However there are other issues, both positive and negative, which arise from a variety of technical and human factors. Such as:

Policy environments, such as the INSPIRE Directive. The Directive should be very beneficial in that it will help progress toward a GIS-centred research platform that enables discovery in multidisciplinary science and knowledge sharing, and fosters more meaningful analyses of data and the visualization, modeling, and simulation of real-world phenomena.

Quality assurance: the quality of information and resources within a GIS is not yet easy to evaluate. It remains a major task to assess and manage quality as systems become more interoperable globally.

The potential of the Semantic Web to support building knowledge and semantics into the next generation of scientific tools is considerable. It could support 'intelligent' processing of geospatial metadata, data, information, and knowledge for virtual communities and multiple scientific domains. How to capture, represent (visualize), and integrate knowledge within and across geospatial domains are all ongoing challenges. Transforming and integrating informal ontologies into formal community-accepted ontologies is a further challenge.

Data preservation and accessibility: A GIS should support the entire data life cycle, including the acquisition, verification, documentation for subsequent interpretation, integration from multiple sources, analysis, and decision support. In a collaborative environment, a GIS is widely needed to manage data and serve as a tool for managers and practitioners to access, analyze, and determine data management needs. Results representation and visualization are especially critical when using semantic technology to interpret datasets and to develop attractive end-user interfaces (Gahegan *et al*., 2009).

As pointed out in this report, there are issues related to the different communities who are now working in or alongside the GIS field. There are communities with behavioural origins in the realms of, for instance, ecology and environmental science, whose work practices may still be founded in a world of small data volumes. Or there are those for whom geospatial information is an adjunct to the fundamental elements of their work. For these the move to interoperability and management of large data volumes may be taking place slowly or ineffectively. Perhaps it is in relation to these communities in agriculture-related fields that a community initiative is needed? It is in this light also that the DLESE community is relevant. The study by Marlino *et al*.. (2009) addresses business planning for sustainability in the context of the development of the service.

Ownership: Both in the case of satellite images and in the case of on-farm surveys, the question arises of who owns the data, as data is collected and value is added along the chain. Data may be acquired directly through observations but may also be acquired from other parties. For example, satellite images or digital maps are used for research with a spatial component.

6 Bioinformatics and Genomics

6.1 Background

Mankind has been working for thousands of years to improve the genotypes of useful plants and animals by selecting for useful traits and against harmful ones. Often it depends on the environment whether such a trait is expressed in the phenotype. Breeding was first done by selecting in an intuitive way, but later the work of Mendel (genotype and phenotype) and Darwin (selection and hybridization) gave it its initial scientific basis in the 19th century. Useful and harmful traits show a continuous variation and are usually controlled by more than one gene. Therefore the development of quantitative genetics was a necessary step to apply Mendelian genetics for agricultural breeding.

In the 20th century the principles for the molecular basis of genetics and gene expression were clarified. At the end of that century the techniques to determine the nucleotide and peptide sequences - initially a rather laborious process preferably done for very simple organisms like bacteriophages – improved enormously and commercial platforms for parallel sequencing ("high throughput sequencing") came on the market. Enormous amounts of sequencing data became available and gave rise to a new area of research: bioinformatics, with different branches like genomics, proteomics, and transcriptonomics. This area combines experimentation and digging out useful information from these results by comparing sequences with other documented sequences from experiments elsewhere. Bioinformatics can be used in various ways to improve agriculture:

- Better understand the backgrounds of diseases
- Easier selection by using molecular markers
- Introducing genes from other species.

The last method is sometimes considered controversial, but this is not the reason why we have chosen not to include these techniques in this report. They come with their own set of data and intellectual property issues (Dunwell, 2005) and it would not be feasible to treat those here

6.2 Current landscape

There are good overview articles about the background of molecular breeding of plants (Moose & Mumm, 2008) and animals (Hu *et al*., 2009).

Mapping the complete genome of an economically important organism is usually a collaborative effort between research institutions from different parts of the world. The first reported map in livestock was for the chicken (Gallus gallus) and now maps are available at least for cattle (Bos taurus), Pig (Sus scrofa), Sheep (Ovis aries), Goat (Capra hircus), rabbit (Oryctolagus cuniculus) and Duck (Anas platyrhynchos). For plants in general the PlantGDB[109] site for comparative plant genomics lists 29 genomes including 16 crops. The list does not include the complete Musa (banana, plantain) genome that has recently been published (D'Hont *et al*., 2012). We will further illustrate how such a collaboration between laboratories may work in chapter 6.4.

6.3 Standards and services

The most important integrative services in this field cooperate in the International Nucleotide Sequence Database Collaboration (INSDC)[110]. The collaboration describes itself as follows (Cochrane, Karsch-Mizrachi, & Nakamura, 2011):

> *"The International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org) represents one of the most celebrated global initiatives in public domain data sharing. Growing from efforts in the early 1980s to capture and present the increasing volumes of sequence and annotation that arose from the emerging application of sequencing techniques, by 1987, the INSDC had taken shape with the stable three party membership that persists to this day. The parties to the collaboration are the DNA Databank of Japan (DDBJ) at the National Institute for Genetics in Mishima, Japan; the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK; and the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, USA. Together, the INSDC partners set out to provide a globally comprehensive collection of public domain nucleotide sequence and associated metadata. Coverage includes the spectrum of data, ranging from raw reads, through assembly and alignment information, to submitted functional annotation of assembled sequences. [..] Routine data exchange, standard formats and, increasingly, the sharing of technology, provide global synchrony across the collaboration."*

These services are best known for offering access to raw data *Sequence Read Archive* (SRA), assembled sequences and annotations, but the collaboration started with the development of standards, both syntactically and semantically, for annotations through the INSDC Feature Table Document that is updated twice a year.

The network also establishes internal rules for the availability status of a piece of data in the form of the INSDC Status Convention[111]. Data can be fully public, confidential prior to publication, or suppressed (as updated improved data become available). This status is assigned by the institute that submits the data. The INSDC partners are only hosting the data and, they make it very clear that they do not own the data

[109] PlantGDB http://plantgdb.org/prj/GenomeBrowser/

[110] INSDC http://www.insdc.org/

[111] INSDC Status Convention http://www.insdc.org/insdc_%20status.html

Within the network a bespoke XML format is used for the metadata of submissions (information relating to a sample, experimental design, library creation and machine configuration), SRA XML[112].

Upon submission a sequence gets an identification number[113]. There is a general convention in the field, enforced by many publishers, that these sequence ID`s are used to refer to sequences in publications.

Services store the sequences themselves in a binary format, but it is usually submitted and retrieved as flat files, usually consisting of one or more header lines and followed by the sequences themselves (where each nucleotide or peptide is represented as a Latin script letter). Flat file formats stem from a database where a sequence is submitted, or a commercial sequencing platform that is used to generate the data. Examples of such flat file formats are:

- EMBL – The flat file format used by the EMBL to represent database records for nucleotide and peptide sequences from EMBL databases
- FASTA – The FASTA file format, for sequence data. Originally from the FASTA software package but now a more generic standard.
- FASTQ – The FASTQ file format, for sequence data with quality.
- GenBank – The flat file format used by the NCBI to represent database records for nucleotide and peptide sequences from the GenBank and RefSeq databases

Within these networks software tools are made available as well, like Basic Local Alignment Search Tool, or BLAST, an algorithm for comparing primary biological sequence information. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Different types of BLAST are available according to the query sequences. For example, following the discovery of a previously unknown gene in Brassica, a scientist may choose to perform a BLAST search of the Arabidopsis genome to see if it carries a similar gene.

6.4 Life cycles and workflows

There is a wide range of software packages to support workflows in bioinformatics[114]. These systems support the data integration rather than the experimentation. We have chosen not to attempt presenting the processes for mapping a genome as it would require more background in molecular biology than we can expect from the readers of this report. We will describe here two international collaborative efforts to unravel the genomes of useful species, Brassica (a.o. cabbages) and Sus scrofa (pig) . There are several steps involved in this:

[112] SRA XML http://www.ncbi.nlm.nih.gov/books/NBK56555/
[113] Sequence Identifiers: A Historical Note http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html
[114] Bioinformatics workflow management systems
http://en.wikipedia.org/wiki/Bioinformatics_workflow_management_systems

Developing a BAC library. “BAC" is an acronym for 'Bacterial Artificial Chromosome. A short piece of the organism's DNA is amplified as an insert in a bacterial chromosome, and then sequenced. Finally, the sequenced parts are rearranged in a computer application (“in silico”), resulting in the genomic sequence of the organism.

The following steps involve the identification of markers to develop a physical map that gives the physical, DNA-base-pair distances from one landmark to another, in contrast to a genetic linkage map that illustrates the order of genes on a chromosome and the relative distances between those genes. Linkage maps are the results of crossing experiments that may predate bioinformatics.

Some of the possible steps in this process are:

> **Identifying SNPs.** A single-nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring when a single nucleotide in the genome (or other shared sequence) differs between members of a biological species.
>
> **Identifying QTLs.** Quantitative trait loci (QTLs) are stretches of DNA containing or linked to the genes that underlie a quantitative trait (a phenotype characteristic that varies in degree and that can be attributed to more than one gene). Mapping regions of the genome that contain genes involved in specifying a quantitative trait is done using molecular tags such as SNPs.
>
> **Identifying ESTs.** An expressed sequence tag or EST is a short sub-sequence of a complementary DNA sequence. They may be used to identify gene transcripts. The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "tags" to fish a gene out of a portion of chromosomal DNA by matching base pairs.

There are more methods, but for the purpose of this report it is relevant to note that most of thes information that is collected in these collaborative efforts is selected sequences deposited in INSDC databases. Other resources are pointers to the physical BAC libraries of microorganisms, and pointers to linkage maps.

6.4.1 The Brassica Genome

The Multinational Brassica Genome Project[115] was established in 2002, following discussions amongst members of the international research community who were at that time involved in developing a number of genomic resources. The Steering Committee for the MBGP selected Brassica rapa as the first species to be sequenced, as it has the smallest genome (ca. 550

[115] Multinational Brassica Genome Project http://www.brassica.info/resource/sequencing/BrGSP.php

Megabase), the lowest frequencies of repetitive sequences, and communal BAC[116] libraries and mapping populations are available. This picture illustrates the division of efforts between international partners. Laboratories in different countries have agreed to work on different chromosomes.

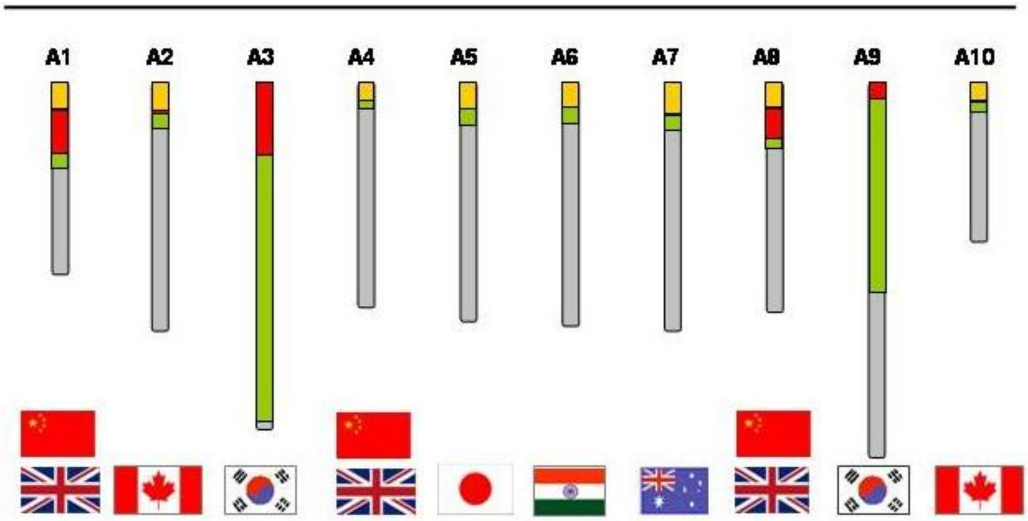


Figure 6.1: Brassica genome: International cooperation organized by chromosome

The international consortium worked together by end-sequencing BAC libraries, consisting of ca. 130,000 clones, by groups in Korea, Australia, Germany, Canada, France, USA and UK, in support of the strategy using BAC end sequences to identify overlapping clones.

The efforts (Iniguez-Luy, *et al*., 2009) to construct the genome involved the use of different types of marker and other sequences from different public sources, and of analysis with genomic blocks (from public databases) of another species (Arabidopsis thaliana) that acts as a reference organism.

The resources that the consortium makes available to the research community support the discovery of clone libraries (usually the result is a pointer to a contact), published maps (where the result is a pointer to a publication), or markers (where the result is a list of sequence identifiers).

6.4.2 The Sus scrofa (pig) Genome

The work on the pig genome is more advanced than the work on Brassica and several pig genomes have been released (Fan, Gorbach, & Rothschild, 2011). The physical map is now being refined. Coordinated efforts to better understand the pig genome were initiated in the early

[116] The term 'BAC" is an acronym for 'Bacterial Artificial Chromosome’. A short piece of the organism's DNA is amplified as an insert in a bacterial chromosome, and then sequenced. Finally, the sequenced parts are rearranged in a computer application (“in silico”), resulting in the genomic sequence of the organism.

1990s with gene identification and mapping efforts. The first two pig linkage maps were generated by the PiGMaP and USDA-MARC genome mapping projects in the mid-1990s. After entering the new millennium, a ‘White Paper’ outlining the roles pigs play in agriculture and as biological models for humans was announced, with the objective to sequence the whole swine genome. The Swine Genome Sequencing Consortium (SGSC) was established. Work then started to identify molecular markers. The ArkDB database combines information from various linkage maps, and has information on nearly 1,600 genes and 3,300 markers[117]. An extensive summary of pig QTL is available at the PigQTLdb[118]. Efforts are now moving to developing the pig-human comparative map and integrating the linkage, physical and cytogenetic maps.

The resources being made available to the research community are comparable to the resources for the Brassica community. In addition there are a number of specific disease related resources like the Swine Leukocyte Antigen (SLA) website[119] that brings together on the nomenclature and DNA sequence data for the associated genes.

6.5 Outlook for a supportive infrastructure

Data integration is the main challenge for research in the fields of genomics and bioinformatics in general. Commercial methods for high-throughput sequencing technologies parallelize the sequencing process, producing thousands or millions of sequences at once. As a rule of thumb it is often said that in a typical project in this field the experimental observations take about 20% of the research effort. Most of the efforts of researchers in this area are spent comparing this experimental data with nucleotide or peptide sequences retrieved from various sources in various formats found through skilful querying of different services. In an overview (Zhang, Bajic, & Yu, 2011) a number of different methods for data integration are distinguished:

- Data warehousing: all data from disparate sources is copied and transformed to offer a one-stop-shopping service. In fact the services discussed before International Nucleotide Sequence Database Collaboration (DDBJ in Japan, EBML-EBI in Europe and NCBI in the USA) are such data warehouses.
- Federated databasing: a user`s query is translated to queries against disparate data sources and the responses are displayed to the user as an integrated result. Examples are Biomart (OICR and EBI) and Discoverylink (IBM).
- Service oriented integration: Individual data sources agree to open their data through Web Services. Examples are Biomoby and Taverna, that uses a bespoke XML based lingua Franca (SCUFL, Simple Conceptual Unified Flow Language).
- Semantic integration: sets of tools like RDFizer and Sesame are used in studies to transform existing data sources into RDF. Examples are Bio2RDF and a special interest group of the W3C Semantic Web Health Care and Life Sciences (HCLS) Interest Group that recently issued guidelines for the conversion of biological and biomedical data into

[117] The ArkDB Database http://www.thearkdb.org/arkdb/do/getChromosomeDetails?accession=ARKSPC00000001
[118] PigQTLdb http:// www.animalgenome.org/QTLdb/pig.html
[119] Swine Leukocyte Antigen (SLA) http://www.ebi.ac.uk/ipd/mhc/sla/index.html

RDF and the use of relevant ontologies. (Splendiani, Burger, Paschke, Romano, & Marshall, 2011)

All these approaches are confronted with their limitations in view of the sheer amounts of data that is becoming available. The members of the INSDC collaboration ask pertinent questions with regard to their future role (Cochrane *et al.*, 2011):

- Technical: The “yield-doubling time” has gone down from 18 months to 5 months as a result of the development of next-generation parallel sequencing platforms. Affordable storage under a sustainable economic model is required
- Social and organizational: as a result of the advent of next-generation sequencing technology it is no longer necessary to understand all the intricacies of the sequencing work. Hence the user-base of integrating services is broadening and new needs for specialized support and new standards like the community developed MIGS (Minimum Information about a genome sequence).

The technical challenge applies specifically to the data warehouse approach. This approach also faces the challenge of how to synchronize data with the sources and make sure that changes at the disparate sources are reflected. Other approaches are also faced with specific issues. For the federated databasing approach and the service oriented integration approach the knowledge base about query interfaces and webservices respectively needs to be maintained. The semantic integration initiatives at this moment copy data from disparate sources ad transform it into RDF format. So in a sense this is a specific form of data warehousing and with that come issues of synchronization with the disparate sources.

For agINFRA the question is relevant what are the specific requirements for a European data infrastructure to support agricultural genomics and bioinformatics. The answer will require much more consultation with the relevant communities, but we can draw a number of lessons from this this section:

- The collaboration in this field is truly global and further developments should advance a European contribution to a global infrastructure rather than creating a European infrastructure.
- Agricultural genomics cannot have an effective infrastructure that is separate from other fields like biomedicine, if only because non-agricultural reference organisms are used, like the rat, the human and Arabidopsis.
- One of the important challenges in this field is to link sequencing information and more traditional, agronomic information about useful traits. An infrastructure could create conditions to make such information more usable for modern genomics.

7 Agricultural Economics

7.1 Background

It is beyond the scope of this report to give a comprehensive view how data is generated and used in agricultural economics as a whole. We will concentrate on two examples of data collections and two models for data exchange on different levels.

- At the **micro level** economists and social scientists do observations, often in the form of surveys. The results of these surveys have always been exchanged in the form of publications (see chapter 3 on Bibliographic Resources). Until recently the underlying data was not archived systematically for verification and re-use. We will give an example of a repository as an archive for such underlying datasets (IFPRI Dataverse).
- At the **macro level** economists study, for example, the effects of policies on the sector as a whole. They may acquire data from statistical offices and use those data to test economic models. Economists from different institutes and countries may work together, each of them providing part of the data that is used in the model. So here a model rather than a repository is the vehicle that brings scientists together to exchange data. We will give the example of the AGMEMOD network that has developed collaboratively a model to predict the effects of reforms in EU Common Agricultural Policies (CAP`s).

7.2 Current landscape

7.2.1 Macro level

Data for studies at macro level is often acquired from national statistical offices or international offices like Eurostat and FAOSTAT. For other sectors there are also commercial databases like ABI/Inform or Thomsonone but they are not used on a large scale in the agricultural sector.

FAOSTAT[120]: The 'Food and Agriculture Organization Corporate Statistical Database' (FAOSTAT) website disseminates statistical data collected and maintained by the Food and Agriculture Organization (FAO) on a number of agricultural subject domains such as production, trade, food balance sheets, and price statistics. Data are provided as a time-series from 1961 in most domains for over 200 countries.

EUROSTAT[121]: Eurostat is the statistical office of the European Union situated in Luxembourg. Its task is to provide the European Union (EU) with statistics at European level that enable comparisons between countries and regions. In fact it predates the European Union as a whole: the office was established in 1953 (to meet the requirements of the Coal and Steel Community).

[120] FAOSTAT http://faostat.fao.org/

[121] EUROSTAT http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Over the years its role has broadened and developing statistical systems in candidate countries for EU membership is becoming more important.

Offices like EUROSTAT and FAOSTAT depend on member states to supply the necessary data and do not always have the mandate to correct data from those offices, for instance if their input is not consistent over the years. The example of the AGMEMOD network illustrates the importance of correcting and harmonizing input from different sources by imputation (ranging from simple methods like a phone call to obtain a missing value, taking last year's value if that is acceptable for the model, to interpolation or more refined statistical methods).

7.2.2 Micro level

Until recently the underlying data of survey studies was usually stored locally, and although codes of conduct often require that scientific data should be available at least for verification purposes, in reality underlying data may have been lost when a research project was finished. There have been initiatives to build data archives where datasets are stored together with metadata and with data documentation (giving details about files, parameters and research methods) to enable re-use. Examples like the UK Data Archive[122] and DANS[123] in the Netherlands have their origin in the social sciences and humanities. We will illustrate the workflows for data archiving and data curation with the example of the Dataverse network that offers facilities to support data curation.

7.3 Standards and services

7.3.1 Macro level

An important development on the macro level is the development of the SDMX standard. It has been developed by international statistical offices like EUROSTAT, IMF, OECD, and the WorldBank, but it is now also being adopted for data exchange in research networks. The following description is from the SDMX User Guide ("SDMX USER GUIDE," 2009)

> *The SDMX message formats have two basic expressions, SDMX-ML (using XML syntax) and SDMX-EDI (using EDIFACT syntax and based on the GESMES/TS statistical message). The standards also include additional specifications (e.g. registry specification, web services). Version 1.0 of the SDMX standard has been recognised as an ISO standard in 2005.[2]. The latest version of the standard - SDMX 2.1 - has been released in April 2011.*
>
> *The stated aim of SDMX was to develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an*

[122] UK Data Archive http://www.data-archive.ac.uk/
[123] DANS http://dans.knaw.nl/

> *IT architecture for exchange of data and metadata. Organisations are free to make use of whichever elements of SDMX are most appropriate in a given case.*
>
> *SDMX aims to ensure that metadata always come along with the data, making the information immediately understandable and useful. For this reason, the SDMX standards and guidelines deal with both data and metadata.*
>
> *Structural metadata are those metadata acting as identifiers and descriptors of the data, such as names of variables or dimensions of statistical cubes. Data must be associated to some structural metadata, otherwise it becomes impossible to properly identify, retrieve and browse the data.*
>
> *Reference metadata are metadata that describe the contents and the quality of the statistical data (conceptual metadata, describing the concepts used and their practical implementation, methodological metadata, describing methods used for the generation of the data, and quality metadata, describing the different quality dimensions of the resulting statistics, e.g. timeliness, accuracy). While these reference metadata exist and may be exchanged independently of the data (and its underlying structural metadata), they are often linked ("referenced") to the data.*

For the integration of statistical data in models, platforms like GAMS[124] are used and there are modules to acquire data in different proprietary formats as well as SDMX. (Dol, 2009)

7.3.2 Micro level

Increasingly datasets are stored in institutional or thematic data repositories. The data repositories themselves are often run on the same software platforms as document repositories, such as Fedora or Dspace. Specific element sets for metadata to describe datasets are under development, like the Datacite metadata kernel[125]. For the datasets themselves there is a longstanding initiative for the social sciences to develop standards to document datasets on project, file and parameter level, the Data Documentation Initiative (DDI)[126].

We will describe here the Dataverse[127] service in more detail as it combines different functions for archiving, curation and sharing of datasets. The Dataverse Network is an open source application to publish, share, reference, extract and analyse research data. A Dataverse Network hosts multiple dataverses. Each dataverse contains studies or collections of studies, and each study contains cataloguing information that describes the data plus the actual data files and complementary files.

[124] GAMS http://www.gams.com/
[125] Datacite metadata kernel http://schema.datacite.org/meta/kernel-2.1/index.html
[126] DDI http://www.ddialliance.org/
[127] Dataverse http://thedata.org/book/about-project

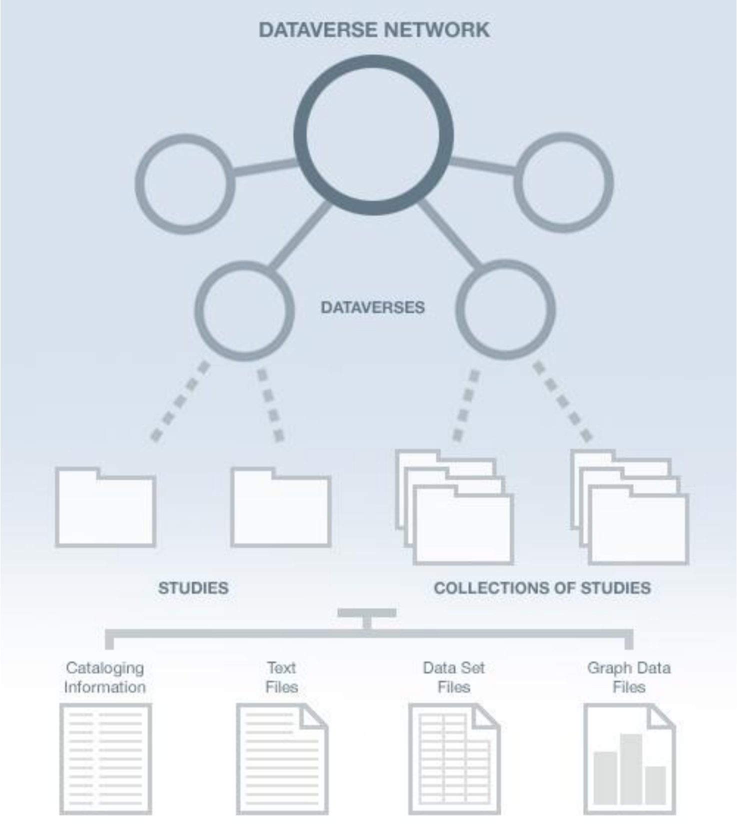


Figure 7.1: The Dataverse system

To the originator of a dataset a dataverse offers a central repository infrastructure with support for professional archival services, including backups, recovery, and standards-based persistent identifiers, data fixity, metadata, conversion and preservation. At the same time, it offers distributed ownership for data authors. It provides scholarly citation, custom branding, data discovery, control over updates, and terms of access and use. This combination of open source, centralized, standards-based archiving and distributed control and recognition makes the Dataverse system unique across data sharing solutions.

7.4 Life cycles and workflows

7.4.1 Micro level: IFPRI@Dataverse

The following description is an excerpt of a case study that was done in 2011 for the EU OpenAIRE program. (Besemer, *et alii*., n.d.).

The IFPRI is an international agricultural research centre working on informing national agricultural and food policies to find sustainable solutions for ending hunger and poverty. Much of the Institute’s research work relies on data collected through socioeconomic surveys and experiments. This has changed recently with the adoption of new technologies for the recording of information and new approaches to capture data. The IFPRI Mobile Experimental Economics Laboratory (IMEEL) was established in 2007. Its primary objective is to collect data through economics experiments in the field to better understand the behaviour of smallholders and the poor in rural areas, especially in Africa, Central America and the Caribbean, Latin America and south-east Asia. These experimental data are usually combined with survey data to understand farmers' decisions on the adoption of new technologies, and participation in marketing activities, contracting arrangements and farmer groups.

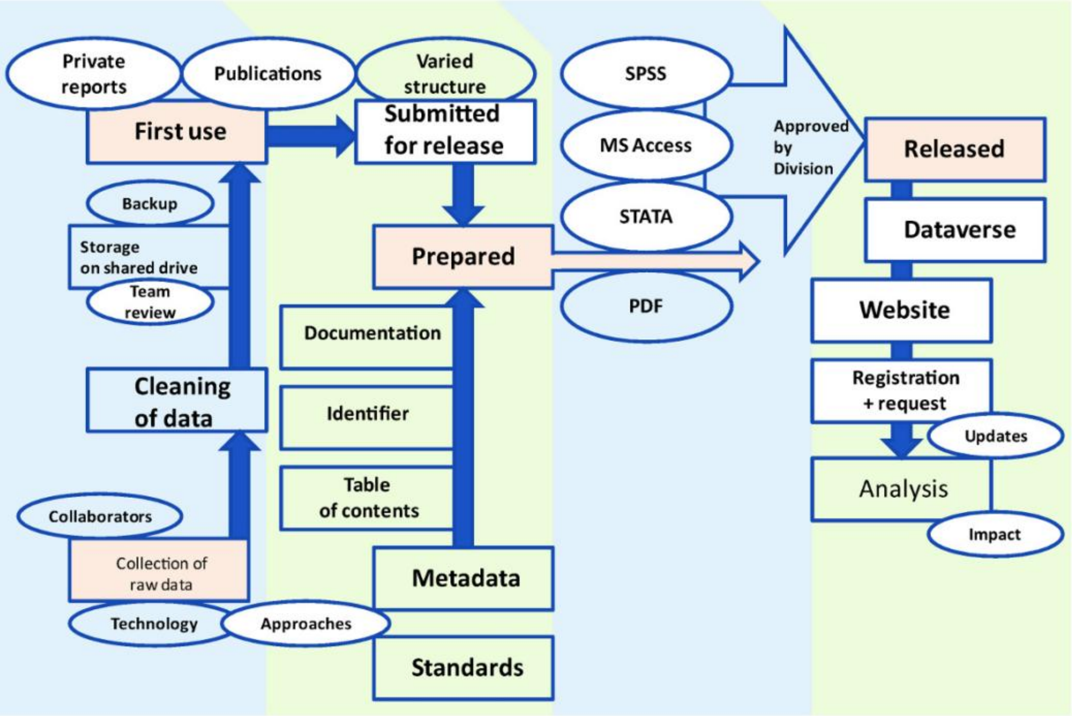


Figure 7.2: Workflow for IFPRI datasets and publications

A number of methods are used for collecting data including a variety of personal digital assistants, cell phones and tablets. Whilst there may be different risks in digital collection, the advantages of software to improve data collection provide increased efficiency in the collection and reduce the need for processing. For example, the software includes controlled responses and range checking, thus reducing errors in collection. The output in each case is a rectangular data file readable into statistics packages or Microsoft Excel. The choices of handheld devices for data capture is based on their battery life, ease of use and their durability. The data captured is cleaned by the research team and then stored in a shared area for review and validation. Whilst the data is held on the shared drive it is regularly backed up from the Institute's servers.

The data will then be used within the organisation either for the production of a donor report or limited distribution report or for a publication. The software used to analyse the data during this stage is SPSS, Stata, Excel or Access. Any models produced or developed during this stage are held on the researcher’s machine or the shared drives. Several of these models will be worked into a knowledge product and shared with the public through the institution’s website. The data is not released until the derived research is published. Once used for a publication, the publications review committee will require the author to submit the supporting data set. This may be submitted in several forms: STATA, SPSS, Excel, Access and PDF. It is then tidied, documented and packaged by the Library and Knowledge Sharing Unit in discussion with the researcher. A table of contents will be produced to indicate the various supporting components of the data set which comprises original questionnaires and resultant data sets. Attention is paid to ensuring anonymity of survey participants, standard formats for files where applicable and the addition of appropriate metadata. Once approved by the Division, the resultant files and records are then published in an external Dataverse repository and through web services it is made visible on the institute’s website as well.

7.4.2 Macro level: AGMEMOD

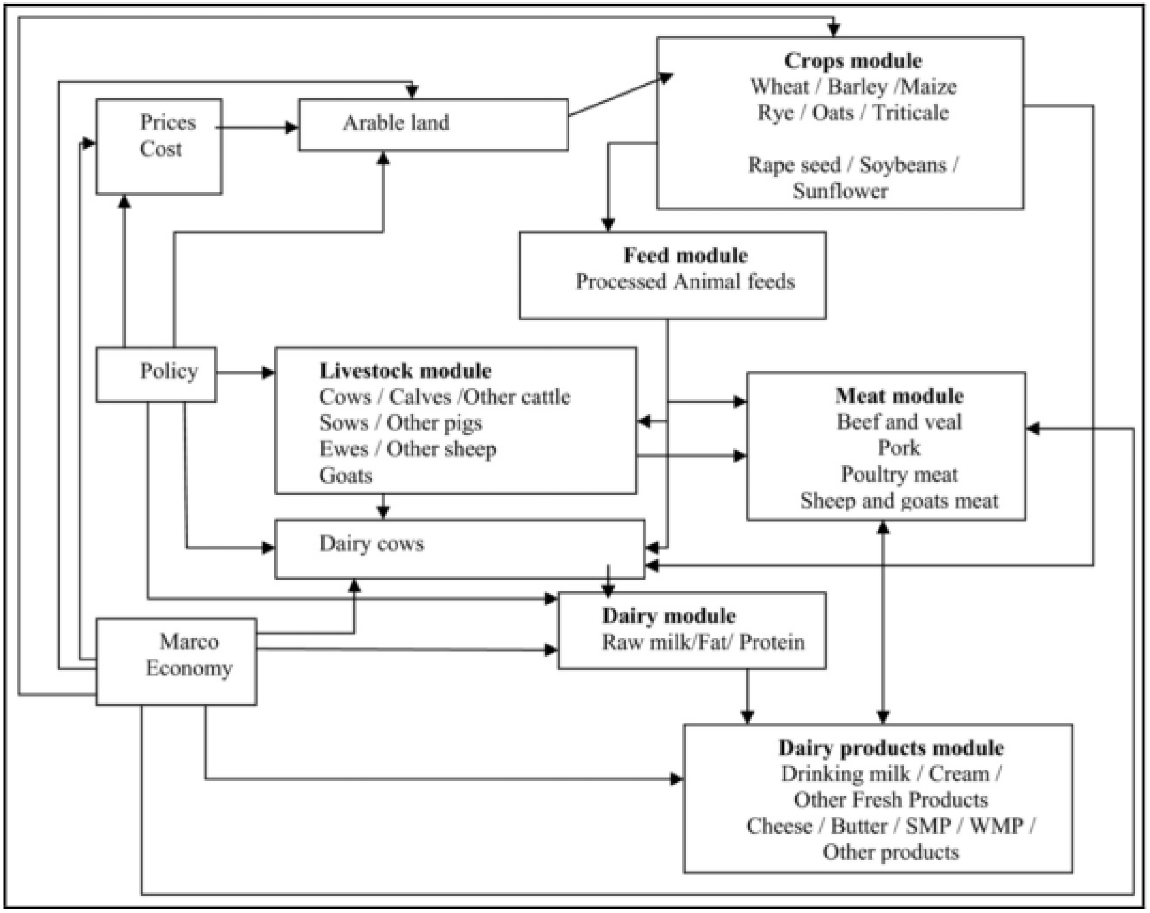
The following description is based on two publications about this collaborative modeling effort: (Bartova, 2008) and (Salamon, Thünen-institut, & Chantreuil, 2008).

AGMEMOD stands for “Agri-food projections for the EU member states”. The integrated AGMEMOD1 model links national partial equilibrium (PE) models for each Member State, possible Accession countries, and important neighbouring countries, into a combined model. This model is aimed at capturing the heterogeneity of European agriculture across EU Member States, while enabling, at the same time, simulations of the Common Agricultural Policy (CAP) and national agricultural policies in a consistent and harmonized way for the whole EU. In the process multidisciplinary teams in each of these countries were involved in building and verifying their own country models which were established on agreed rules for data, model design and underlying assumptions.

In the initial stages, the Partnership decided to replace the existing Excel country models with GAMS models to overcome PC memory problems at that time. Model revisions, an integral part of the model review and evaluation process, led to the need to constantly disassemble and re-combine country models, with all the associated problems and difficulties. In the process, guidelines on desired model building practices were formulated and later have been integrated into a tool to be used by project partners:

- Models should be reproducible to meet scientific standards;
- Other researchers should also be able to handle the models;
- Models are required to be flexible to meet the needs of different projects;
- Models should be reviewed by experts in order to enhance their overall quality; and
- Models should be easily amended and connected to other models.

Figure 1. Linkages between sub-models in the AGMEMOD national country models



Source: AGMEMOD Partnership

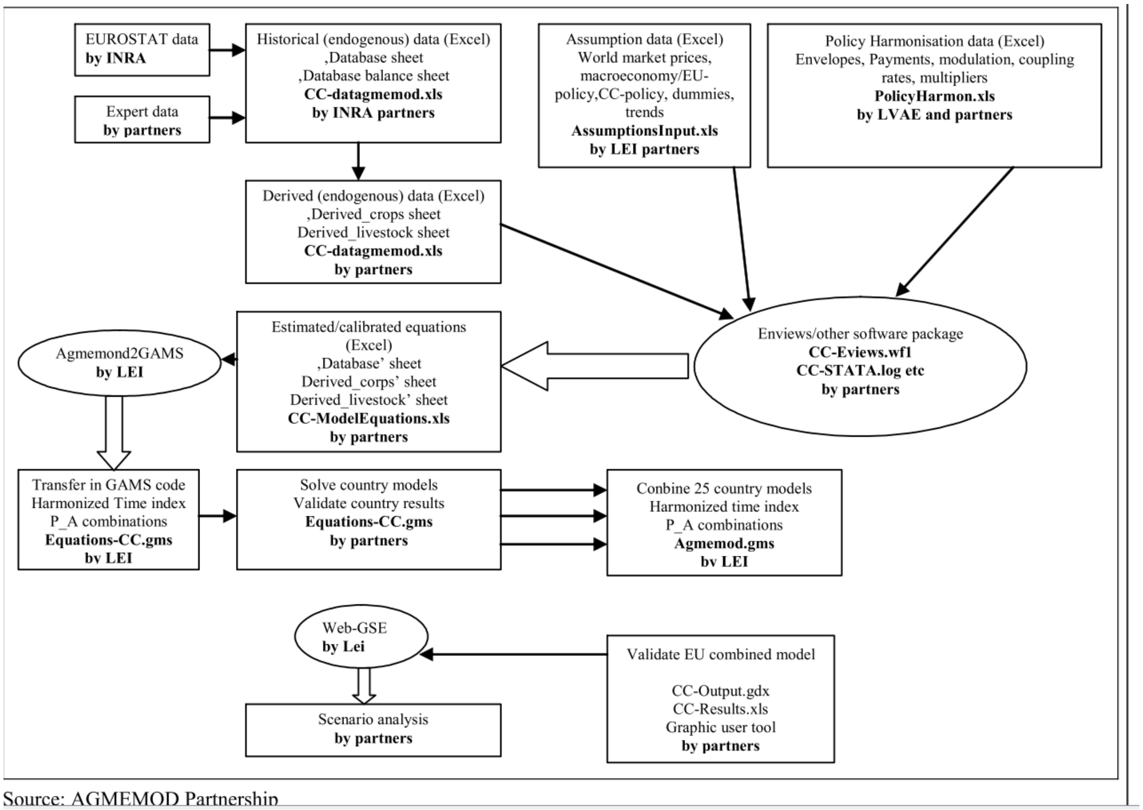
Figure 7.3: Linkages of AGMEMOD sub-models

At present not only the data, but also model equations, are converted from Excel or even directly from the econometric estimation into GAMS code overcoming deficiencies in the former GAMS code as controls have been established on, for instance:

- The existence of a full set of equations per country;
- The declaration of variables as being both exogenous and endogenous.

Data requirements for the AGMEMOD modeling approach are generally high, as time series for the parameter estimation purpose are requested to cover not only the supply side of agriculture but also different types of usages as well as processing. Each country model is based on an aligned database of annual time series, covering, in principle, a period from 1973 to the latest available year.

Where possible the AGMEMOD Partnership uses Eurostat sources.



Source: AGMEMOD Partnership

Figure 7.4: AGMEMOD data flows

Although, ideally, all data would be drawn from the same database, in practice, however, these may be incomplete or inconsistent or reflect some errors. Where there are such gaps or errors, the recommendation is to derive comparable data from different sources. If frequent database revisions are not taken into account through re-estimation of the respective equations, the model results will not reflect such changes in the database.

Length of the time series available may vary a lot from the standard for particular countries. Furthermore, national borders for some Member States may have changed in the course of time. In advance of and during the EU accession agricultural, market regimes may have changed, often combined with a harmonisation of the related statistics.

The resulting AGMOMOD models can be downloaded from the consortium`s website.[128]

[128] AGMOMOD http://www.agmemod.eu/the-models.html

7.5 Outlook for a supportive infrastructure

The question for agINFRA is how a semantically enabled infrastructure could support these exchanges of research data at micro and macro level.

7.5.1 Micro level

Datasets as they are published on data repositories (including 'dataverses') are in a sense publications and the metadata is often stored on platforms that can be semantically enabled. There is no adequate discovery service for datasets yet, and semantic technologies can help to create such a service. We refer here to the discussion in the chapter about bibliographic resources.

The data itself is stored in formats to be processed by the spreadsheet, database and statistical software package that scientists have at their disposal. It is not realistic to expect that the infrastructure will be able to take over the tasks that scientists now perform on their local computers. So we see no task for the infrastructure to deal with the datasets themselves.

7.5.2 Macro level

The format and protocols that are used by statistical offices to deliver statistics to researchers and other end-users are decided by those statistical offices, and a research infrastructure has no role to play here. There are certainly LOD initiatives for statistical data, like the RDF Data Cube vocabulary, but it is up to the data providers to decide if and when such vocabularies will be used.

We have seen in the AGMEMOD case study that scientists use whatever software packages are available, but often MS Office products, to work further on the data. As for the data itself in datasets, as we described them at micro level, we do not see a role for a semantically enabled infrastructure here.

8 Plant Genetic Resources (Germplasm)

8.1 Background

Plants often produce seeds that are able to survive unfavourable conditions and are meant to be distributed to other, hopefully more favourable, conditions. Agricultural scientists have made use of this property of plants by building collections of seeds as a resource for further research and development. In 1894 Professor A.F. Batalin, Director of the Saint Petersburg Botanical Garden, made the initiative to organize the Bureau of Applied Botany. During 1901 and 1902, requests were distributed throughout the Russian provinces to collect and return seeds of local cultivars (landraces) of agricultural crops. This was the start of the first collection and now there are all over the world collections of plant propagation materials. They are sometimes referred to as seed banks but we will use here the more generic term “germplasm collections” as they may also contain propagation materials from plants that are not propagated through seeds.

The union catalogues of these germplasm collections are amongst the oldest data systems that have been automated in the agricultural sciences. Therefore they can be considered as mature, but with this maturity comes a certain conservatism (or lack of investment in innovation). As one author remarked (TJL van Hintum, 2010) : "The community involved in the ex situ conservation of Plant Genetic Resources (PGR) is traditionally and by nature a conservative community. Conservation implies keeping what you have, preventing loss or change. However, it is the PGR that need to be conserved, and not the methodology to do so."

8.2 Current landscape

GENESYS[129] is a global portal to information about Plant Genetic Resources for Food and Agriculture. It was launched in 2011 as a one-stop access point to the information provided by three important genetic resources communities:

- Eurisco (European Plant Genetic Resources Search Catalog)[130]
- SINGER (System-wide Information Network for Genetic Resources) of the CGIAR[131]
- USDA-GRIN (The Genetic Resources Information Network of the United States Department of Agriculture).[132]

It offers access to more than 2.3 million germplasm collection accessions out of the estimated 7.4 million accessions existing worldwide.

[129] GENESYS http://www.genesys-pgr.org/
[130] Eurisco http://eurisco.ecpgr.org/
[131] SINGER http://singer.cgiar.org/
[132] USDA-GRIN http://www.ars-grin.gov/

Soon after the germplasm collection documentation systems became computerized, scientists were tempted to combine the information of different systems and analyze the result to determine the coverage of the gene pool in the combined collections, but also to determine the redundancy between collections and to try to coordinate activities of germplasm collections. This led to the establishment of the European Central Crop Databases (ECCDBs). There was initial EU funding but the creation and management of the ECCDBs was a voluntary input in kind contribution of voluntary institutes or scientists. A review in 2008 listed 62 ECCDBs in Europe covering most species maintained in European germplasm collections. Taken together they comprise nearly 750,000 accessions. However, only 12 databases currently contain a limited number of data for characterization and evaluation. In general, the ECCDBs vary widely with regard to their completeness, data quality, age of datasets, and inclusion of data on useful traits, but also the possibility to search or download them via the web

8.3 Standards and services

Data in exchanged in a wide variety of file formats that we will further discuss below. Technically the EURISCO database and the SINGER database are using the same platform that is being maintained by Bioversity International in Montpellier, France. There has been extensive work to develop standards for the content of the data to be exchanged. The minimal data elements to describe an accession have been laid down as the FAO/IPGRI Multicrop-passport Descriptors[133] that were agreed in 1997 and updated in 2001. For specific needs to describe the traits of particular crops there are extensions like the guidelines for developers of crop descriptor lists[134].

The purpose of these databases is the discovery of accessions and it should lead to a transaction whereby a scientist requests seeds or other plant propagation materials for further research. These transactions, the documentation to come with the material and the obligation to share results with the originators of the material are governed by the Standard Material Transfer Agreement (SMTA).[135]

8.4 Life cycles and workflows

8.4.1 Eurisco

The EURISCO system has over 1.1 million accessions from at least 239 holding institutions in 35 participating countries. Data are submitted to the central system by the national focal points. The central system is managed by Biodiversity International in Montpellier, together with the SINGER database. However data collection starts at the institutional level. In an overview of this

[133] FAO/IPGRI Multicrop-passport Descriptors http://www.bioversityinternational.org/nc/publications/publication/issue/faoipgri_multi_crop_passport_descriptors.html

[134] Guidelines for developers of crop descriptor lists http://www.bioversityinternational.org/index.php?id=19&user_bioversitypub%20lications_pi1%5bshowUid%5d=3070

[135] Standard Material Transfer Agreement (SMTA) http://singer.cgiar.org/index.jsp?page=smta

particular information landscape (T. V. Hintum & Begemann, 2008) the authors describe the wide variety of systems used. Almost all of them are computerized, but for smaller collections an Excel file will do.

They send the static “passport” data to the national focal points, but often not the dynamic emergence[136] data and the evaluation data that they receive back according to the SMTA agreement if an accession is used. The National Focal Points use a wide variety of database management systems (DBMS). Data from the different gene banks are sent in by the different gene banks in a wide variety of technical formats, like MS/Access databases, CSV files, etc.

Conversion to the central database is performed on case-by-case bases by the technical coordinator. Around 2005, attempts were made to develop a more standardised updating method using the WSDL/UDDI web services technology that was developing at that time. The Biocase2 protocol was developed in conjunction with Singer/Eurisco, and was deployed at six CGIAR centres. However, there were two bottlenecks: the performance (speed) of the system that implemented the protocol and the relative difficulty of producing the flat files that were required by the system from the various database implementations with which the participating gene banks are managed.

The Biocase protocol[137] had been implemented successfully for the Global Biodiversity Information Facility (GBIF). In 2010, GBIF, NordGen and Bioversity International initiated a feasibility study to evaluate how the GBIF infrastructure can meet the needs of the European germplasm collection community. A version of the software (GBIF Integrated Publishing Toolkit (IPT version 1.0)) with the DwC germplasm extension was installed at five institutions in the European network (Endresen & Knüpffer, 2012).

Figure 8.1 depicts the flow of information within the communities.

Not surprisingly it is very hard to control data integrity in such a decentralized system based on ad-hoc conversion of data to other platforms. A study of the consistency of taxonomic names in the Eurisco database (Theo van Hintum & Knüpffer, 2010) one example is that after the renaming of tomato from Lycopersicon esculentum to Solanum lycopersicum in 1993, one quarter of the tomato accessions in European germplasm collections were being called Solanum, while the rest were still named Lycopersicon.

[136] Emergence data is less relevant for external users than evaluation data. Emergence is tested annually and if the seeds are below a certain emergence level the accession is “regenerated”. This is not done more often than necessary to avoid genetic drift (and probably the cost of regeneration)

[137] Biocase protocol http://www.biocase.org/products/protocols/index.shtml

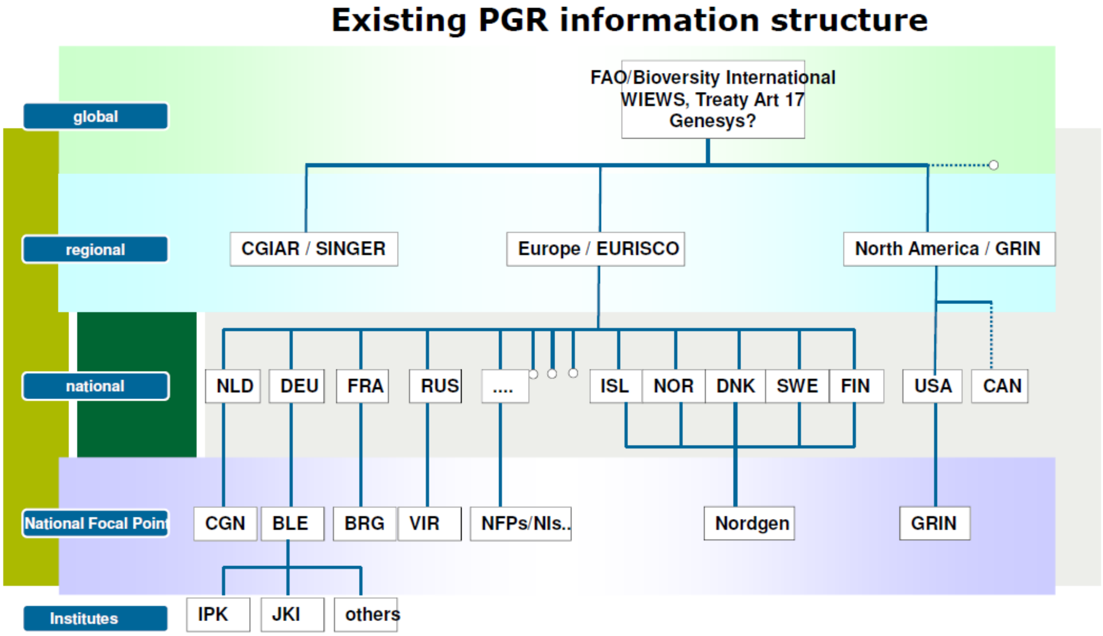


Figure 8.1: Begemann, Frank (2011) Documentation and information of genetic resources in Europe. Experiences from EURISCO. Questions for EFABIS?

8.5 Outlook for a supportive infrastructure

A critical article about the present state of the germplasm collection catalogues (TJL van Hintum, 2010) mentions a number of problems in the current catalogues that can be amended with investments in information technology:

- User interfaces are outdated (“clumsy”)
- Data is inconsistent, and there are no effective vocabularies or ontologies to control data about organisations, or traits for characterization and evaluation.
- The catalogues are lacking effective ways to select accessions with specific traits (which is probably more relevant for the users than their provenance).

Work is under way to develop effective RDF vocabularies and ontologies for germplasm collections (Endresen & Knüpffer, 2012). The Darwin core vocabulary, originating from the world of natural history museums and further developed in the GBIF network, is now developing a germplasm extension as an SKOS / RDF vocabulary. Trait descriptions include elements from the Crop ontology (Shrestha *et al.*, 2010) that builds on older ontologies like the plant ontology and the phenotypic quality ontology.

However there is a striking contrast between these efforts to develop semantic standards and the picture of the network as a whole. If the data is not collected, and not collected in a consistent way it cannot be effectively expressed and linked open data. The best support that the infrastructure can offer is to provide tools that can be used at the institutional level to maintain good catalogues that can produce semantically rich data.

9 Overview of characteristics of the six type of information, workflows and communities covered

The table below summarises some characteristics of the six type of information, workflows and communities that are covered in this deliverable:

	Bibliographic resources	**Open Educational Resources**	**Geospatial information systems**	**Genomics and bioinformatics**	**Plant Genetic Resources (germplasm)**	**Agricultural Economics**
Relation of the data domain to agINFRA	Cover all subject domains	Cover all subject domains	Cover all subject domains	Closely related to agricultural domain	Agricultural domain	Agricultural domain
Community building	Strong library community managed by information management specialists Communities of researchers not enough integrated with those of developers	Some distance between technical expertise and educational communities, intended to be overcome in more recent initiatives	Large and global initiatives in geospatial information, e.g. OGC, GSDI, UN Spatial Data Infrastructure, EU INSPIREinitiative Strong core community of geospatial services developers Small domain-focused communities unused to management of	Highly advanced services have been put in place by international consortia of core institutes in the field of genomics	Established community of germplasm collections that follow traditional practices Union catalogues have been established, but little cooperation with technical expertise centers	Two different communities: Micro-level (small-scale) economic and social sciences surveys; Macro-level model-based economic analysis that uses data of statistical offices

			large data volumes			
Standardisation	Widely accepted and used standards for the creation of metadata	Lack of central foci of activity because of many different needs Not many agriculture-specific repositories of OER Many resources in silos and not interoperable	Widely accepted and used ISO and OGC standards	Core standards and services developed by the International Nucleotide Database Collaboration (INSDC) Genomic Standards Consortium (GSC), established in 2005	Bioversity international standards and guidance	Micro-level: DDI Alliance, Council of European Social Science Data Archives and others Macro-level: coordinated effort of the major statistical offices and agencies
Main (meta)data standards	Dublin Core and others (MODS, TEI) according to digital library focus	Dublin Core and LOM (and APs thereof)	ISO/TC 211 OGC standards	INSDC Feature Table Definition EMBL, FASTA/FASTQ, GenBank file formats Many metadata specification initiatives (minimum information guidelines) for experimental and other research data (cf. MIBBI	Multi-Crop Passport Descriptors (MCPD) Darwin Core germplasm extension	Micro-level: Data Documentation Initiative (DDI) Macro-level: Statistical Data and Metadata Exchange (SDMX)

				portal)		
Aggregation	OAI-PMH harvesting	OAI-PMH harvesting, e.g. GLOBE, LACLO, Organic.Edunet use the ARIADNE harvester	Maps of geo-referenced places, objects, etc. UNSDI and INSPIRE will provide integrated access to resources	Submission by researchers to domain data archives / services	Different ways of transferring data from collections to union catalogues	Submission of micro-level survey data to archives and portals such as CESSDA, UK Data Archive, DANS and others Databases and metadata portals of statistical offices and other agencies (e.g. EUROSTAT, FAOSTAT, OECD, World Bank)

Examples of workflows covered	AGRIS Network	ARIADNE;
		Organic.Edunet;
		AgLR-TF (Green Learning Network) intended workflow
			agtrials.org (CGIAR CCAFS programme)
			Other examples GeoNetwork and UNSDI
			Kepler workflow of project REAP
				Brassica Genome and Sus scrofa (pig) genome projects	EURISCO	Micro-level: IFPRI@Dataverse
						Macro-level: AGMEMOD

Identified issues	Different types of content, including data, becoming openly accessible	Many resources maintained in 'silos' – not interoperable	Quality assurance in global geospatial initiatives	Among the core challenges of the field are the rapidly growing	Collection databases vary widely what concerns data	Data of micro-level surveys are often not shared openly

	Mostly related to researcher and institutional behavior, i.e. lack of open access sharing of content Temporary division between the publication/creation of resources and the description of bibliographic resources	Semantic interoperability in global aggregations IPR/licensing of multimedia learning objects	Uneven distribution of expertise in geospatial data services among domains of research Limited availability of openly licensed remote sensing data (e.g. satellite images) or data of on-farm surveys	volume of data and data integration that allows for comparative and other analysis of available data Different approaches are used such as data warehousing, web services and semantic integration of resources	completeness and quality Lack of effective vocabularies or ontologies Inconsistent use of plant nomenclature Often lack of data that is relevant for researchers, e.g. plant traits (C&E data)	Micro-level researchers use local computers and spreadsheets, statistical software and databases Lack of complete and consistent data for model-based macro-level analysis
Implications for agINFRA	Automated discovery of links in documents Automatic indexing services for bibliographic resources Association between the indexes of bibliographic databases and other sources Document versioning Licensing and	Collect/link metadata of relevant educational resources from global federations such as GLOBE Support the harmonization of vocabularies Bring educators and users into closer alignment with technical development	Effective integration and processing of knowledge of different domains of research (e.g. environment, ecology, agriculture) that are related to geospatial data	Need to identify the specific requirements for including genomics data in agINFRA How to link genomic data to agricultural information, e.g. about useful plant traits?	Collect/link germplasm accession data available in the major catalogues (EURISCO, GENESYS) Include plant nomenclature and relevant ontologies (e.g. plant traits ontology) in the agINFRA semantic layer Offer tools that	agINFRA should focus only on semantic integration of metadata and data discovery services

	ownership of content Integrated authoring environment for CMSs (particularly for metadata) Level of user control and choices of content-related services Provision of widely used taxonomies and thesaurus				allow institutional collections producing semantically rich data	

10 From lifecycles and practices to agINFRA-supported workflows

The aim of the agINFRA task T5.1 is the analysis of the existing and anticipated new policies & practices in content/data management and sharing, particularly of the potential user communities of the agINFRA Integrated Services and Components. It is one of the building bricks for T5.2, the task of which is deliver a generic “management and curation workflow that may be used for the cross-community integration of the agINFRA data sources/repositories using the components of the agINFRA infrastructure”. Apart from this report T5.2 can also build on the results of T2.3. The expectation is that value-added services can be built on top of the agINFRA “virtual data layer” the latter is using Linked Open Data methodologies. However, it is likely that there will be a gap between what is desirable from a standard setting perspective and the realities in terms of of policies and practices of content/data communities “on the ground”.

10.1 From a checklist to a staircase towards full access & reuse

Data infrastructures such as agINFRA aim at providing enhanced services on top of institutional and subject-based/domain data collections, such as repositories. Therefore the situation at the local/institutional and domain levels is important for reaching certain levels of “openness” of data, interlinking of resources in the infrastructure, and providing enhanced services based on them. However, some of the interventions to give more semantic meaning to this data (starting with harmonization) can be done at a higher aggregation level. For agINFRA a staircase model has been suggested that we will use in the subsequent discussion.

10.2 Four layers of workflows and the involvement of agINFRA

A “workflow layer concept” developed in the second agINFRA project meeting in Alcalá de Hernares, Spain (24-26 April 2012) is used to reflect what the implications are of the analyses of the different areas for possible agINFRA interventions. There are at least four layers of workflows agINFRA will need to consider:

10.2.1 Level 1 Workflows of researchers

These workflows of researchers range from data creation to final research results (publications, reusable datasets) which may be deposited in an institutional or subject-based repository or aggregated in another way, e.g. as contribution of a simulation model. (= Level 2 below) Note that some workflows of researchers may involve interactions with systems at higher aggregation levels. For example: researchers in genomics will contribute the sequences that they produce to one of the central systems (GenBank, EBI) for their area, but they will also retrieve data from those services to give their findings meaning, e.g. “ dig out genes”.

This report did not concentrate on such primary data collection workflows. In the area of plant germplasm collections issues were identified with regard to the consistency of primary input.

Semantic tools can be helpful to get “cleaner” and more consistent input at higher levels. Especially if some of these interventions lead to more meaningful description of traits other areas, such as genomics, may benefit.

10.2.2 Level 2 Workflows of repository managers

These workflows of repository management range from content/data ingest to metadata exposure and provision of access based on local systems/tools. They may also include enhanced local workflows, for example, if a repository provides additional services to researchers beyond depositing research material or if it is interlinked with internal, administrative systems.

Some interventions at this level are already foreseen for agINFRA. In the area of bibliographic resoures and Open Educational resources a tool like Agrotagger can contribute to more meaningful and semantically richer metadata. In the area of micro-economics – and the same is probably true for other primary data collections from social sciences – the realization of repositories can be promoted by providing semantic tools. Discovery across repositories of datasets is an issue that needs to be addressed and where semantic tools can be of use.

10.2.3 Level 3 Workflows of data/metadata aggregators & service providers

These workflows range from data/metadata harvesting to provision of value-added services on top of the aggregated data from several repositories. In this study we came across such services specifically in the area of bibliographic resources and open educational resources such as metadata aggregators in the context of Organic.Edunet and AgLR. The same may apply for related services and service providers, such as the CIARD RING.

10.2.4 Level 4 Workflows of agINFRA Integrated Services and Components

agINFRA Integrated Services are envisaged to carry out or support many workflows, for example, RDF-based integration and visualization of related information resources from many harvested sources (e.g. OpenAgris). agINFRA data-processing components will be provided to allow for many additional workflows, for example, metadata extraction and indexing or visualization of networks among research groups. We have made suggestions in this chapter how some of the areas that were studied could benefit from tools that may be developed for agINFRA. With regard to integrated services especially the areas of micro-economics (and other areas that produce datasets in this manner) as well as geospatial information systems may benefit from better discovery options for relevant datasets.

11 Impact on agINFRA Vision

This study of various fields of direct relevance to agricultural research has elaborated on their current information management patterns and workflows. The agINFRA Vision requires the integration and/or interoperability of data sources which are either a part of, or are directly related to, agricultural research. It is suggested by this task 5.1 that to achieve this vision productively there are some unstated subtexts of agINFRA's aims which can be sketched out as follows:

- To be successful a contribution must be made to the seamlessness and effectiveness of 'user' experiences (whether these users are technical developers or researchers in the field).
- Hand in hand with the above point, how much does agINFRA need to stimulate the development of, or be of relevance to, one or more communities of practice in the field of agricultural research?
- Does the semantic web and related ontologies enhance the achievement of the two points above?
- How can the agINFRA development be carried out so as to enhance the sustainability of the information infrastructure and hence be of long term benefit to agricultural research?

The following figure presents the agINFRA Vision as it will be produced in "D1.3 – agINFRA Scientific Vision White Paper":

agINFRA Vision

""To develop a shared infrastructure and computationally empowered services for agricultural research data that allow for producing and transferring scientific and technological results into effective agricultural practice. A key element will be achieving a higher level of interoperability between agricultural and other data resources.""

agINFRA Vision objective	Impact
Develop shared technology infrastructure	The review of the existing agricultural data management practices, lifecycles and workflows will be used as a basis for the agINFRA project which will build its envisioned infrastructure (including tools and workflows) based on the existing status and identifying the existing needs.
Improve agricultural research data services	Based on the existing status of the agricultural research services, the agINFRA project will identify current issues and will support further improvements in these services

Leverage interoperability of data resources	The outcomes of this deliverable will increase the level of interoperability between agricultural and other data sources, based on the common ground identified in this review.

12 References

12.1 Objectives and methods

Pattern recognition. OCLC 2003 environmental scan: a report to the OCLC membership. http://www.oclc.org/reports/escan/toc.htm

12.2 Bibliographic Resources

Besemer H. et al (2011). ‘Agricultural Research’ pp 19-68. In: Studies on Subject-Specific Requirements for Open Access Infrastructure (OpenAIRE). Universitätsbibliothek, Bielefeld

Edge P.A. et al (2011). ‘Researcher Attitudes and Behaviour Towards the ‘Openness’ of Research Outputs in Agriculture and Related Fields.’ AgInfo Worldwide, Vol 4 No 2, pp. 59-69.

Falagas M.E. et al (2008). ‘PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses’. The FASEB Journal, Vol. 22, No. 2: pp. 338-342

FAO (2010). ‘Survey of Open Access Repositories in the Agricultural Domain.’ http://aims.fao.org/advice/open-access/survey. Last accessed September 2012

Suber, P. (most recent revision, 18 June 2012). ‘Open Access Overview’ http://www.earlham.edu/~peters/fos/overview.htm Last accessed August 2012.

Subirats I. et al (2007) ‘Towards an architecture for open archive networks in Agricultural Sciences and Technology’. In ‘International Conference on Semantic Web & Digital Libraries’, Bangalore (India).

12.3 Open Educational Resources

Ebner H. et al (2009): ‘Learning Object Annotation for Agricultural Learning Repositories’, IEEE International Conference on Advanced Learning Technologies, Riga, Latvia, 15-17 July.

FAO (2007). Metadata Application Profile for FAO’s Learning Resources. ftp://ftp.fao.org/gi/gil/gilws/aims/metadata/docs/learnap.doc. Last accessed August 2012.

Geser G. (2012). ‘Open Educational Practices and Resources’, OLCOS Roadmap 2012. http://www.olcos.org/english/roadmap/. Last accessed August 2012.

Hylen J. (2007). 'OER: Opportunities and Challenges'. UNESCO paper http://www.knowledgeall.net/files/Additional_Readings-Consolidated.pdf. Last accessed August 2012.

Manouselis N et al (2010). 'Metadata interoperability in agricultural learning repositories: An analysis' Computers and Electronics in Agriculture. Volume 70 Issue 2: 302–320. Elsevier Science

Marlino M. et al (2009) 'DLESE: A Case Study in Sustainability Planning'. http://pubs.or08.ecs.soton.ac.uk/25/1/submission_65.pdf Last accessed August 2012.

Nilsson M. (2008): 'Draft Recommended Practice for Expressing IEEE Learning Object Metadata Instances Using the Dublin Core Abstract Model.' Draft IEEE P1484.12.4tm/D1.

Sicilia M.-A. et al (2011). 'Navigating learning resources through linked data: a preliminary report on the re-design of Organic.Edunet'. Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age, co-located with the 8th Extended Semantic Web Conference, ESWC2011, Heraklion, Greece, May 29, 2011.

Stracke C.M. and Hildebrandt B. (2007). 'Quality development and quality standards in e-Learning: adoption, implementation, and adaptation.' In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunication 2007, AACE, Chesapeake, VA, 4158–4165.

Ternier et al (2010). 'The Simple Publishing Interface (SPI)' D-Lib Magazine, Vol.16 No.9/10.

Tzikopoulos A. et al (2009). 'An Overview of Learning Object Repositories' In: Selected Readings On Database Technologies And Applications. Ed. by Terry Halpin. Idea Group Inc.

12.4 Geospatial Information Systems

Borgman C. L. et al (2007). 'Drowning in data: Digital library architecture to support scientific use of embedded sensor networks.' In: Proceedings of the 7th ACM/IEEE joint conference on digital libraries, JCDL 2007: Building and sustaining the digital environment, June 18–23, 2007, Vancouver, BC, pp. 269–277.

Gehagen M. et al (2009). 'Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure.' Computers & Geosciences 35(4), pp. 836-854. Elsevier Science.

Hey T. and Trefethen A. E. (2005). 'Cyberinfrastructure for e-Science'. Science, 308(5723), pp. 817–821.

ISO (2010). Text final ISO/CD 19144-2 ‘Geographic information — Classification systems — Part 2: Land Cover Meta Language (LCML)’, as sent to the ISO Central Secretariat for issuing as Draft International Standard.

Marlino et al (2009). ‘DLESE: A Case Study in Sustainability Planning’. http://pubs.or08.ecs.soton.ac.uk/25/1/submission_65.pdf. Last accessed August 2012.

McFerren G. et al (2010). ’FOSS Geospatial Libraries In Scientific Workflow Environments: Experiences and Directions’. FOSS4G Barcelona, September 6th-9th 2010.

Pepe A. et al (2010) ‘From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web’. Journal of the American Society for Information Science and Technology, 61(3), pp. 567-582.

Yang C. et al (2010), ‘Geospatial Cyberinfrastructure: Past, present and future.’ Computers, Environment and Urban Systems, 34, pp. 264–277. Elsevier Science

Zimmerman A. (2007) ‘Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse.’ International Journal on Digital Libraries, Volume 7, Number 1, 5-16. Springer Verlag

12.5 Bioinformatics and genomics

Cochrane, G., Karsch-Mizrachi, I., & Nakamura, Y. (2011). The International Nucleotide Sequence Database Collaboration. Nucleic acids research, 39 (Database issue), D15–8. doi:10.1093/nar/gkq1150

Dunwell, J. M. (2005). Review: intellectual property aspects of plant transformation. Plant biotechnology journal, 3(4), pp. 371–384. doi:10.1111/j.1467-7652.2005.00142.x

D’Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature. doi:10.1038/nature11241

Fan, B., Gorbach, D. M., & Rothschild, M. F. (2011). The pig genome project has plenty to squeal about. Cytogenetic and genome research, 134(1), pp. 9–18. doi:10.1159/000324043

Hu, X., Gao, Y., Feng, C., Liu, Q., Wang, X., Du, Z., Wang, Q., et al. (2009). Advanced technologies for genomic analysis in farm animals and its application for QTL mapping. Genetica, 136(2), pp. 371–386. doi:10.1007/s10709-008-9338-7

Iniguez-Luy, F. L., Lukens, L., Farnham, M. W., Amasino, R. M., & Osborn, T. C. (2009). Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling Brassica rapa and B. oleracea. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 120(1), pp. 31–43. doi:10.1007/s00122-009-1157-4

Moose, S. P., & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. Plant physiology, 147(3), pp. 969–977. doi:10.1104/pp.108.118232

Splendiani, A., Burger, A., Paschke, A., Romano, P., & Marshall, M. S. (2011). Biomedical semantics in the Semantic Web. Journal of biomedical semantics, 2 Suppl 1(Suppl 1), S1. doi:10.1186/2041-1480-2-S1-S1

Zhang, Z., Bajic, V., & Yu, J. (2011). Data Integration in Bioinformatics: Current Efforts and Challenges. Bioinformatics - Trends and Methodologies, (1). Retrieved from http://www.intechopen.com/source/pdfs/22488/InTech-Data_integration_in_bioinformatics_current_efforts_and_challenges.pdf

12.6 Agricultural economics

Bartova, L. (2008). Impact Analysis of CAP Reform on the Main Agricultural Commodities. Report II. AGMEMOD - Member States Results. http://ftp.jrc.es/EURdoc/JRC38152.pdf

Besemer H, Addison C, Pelloni F, Porcari EM, Manning-Thomas N (2011) In: Studies on Subject-Specific Requirements for Open Access Infrastructure. Meier zu Verl C, Horstmann W (Eds.); Bielefeld: Universitätsbibliothek: 19-68. http://pub.uni-bielefeld.de/publication/2458698

Dol, W. (2009). Metabase User manual, (December). Hhtp://www3.lei.wur.nl/gamstools/metabase.doc

SDMX USER GUIDE. (2009)., (January), 1–98. http://sdmx.org/?page_id=38

Salamon, P., Thünen-institut, J. H. V., & Chantreuil, F. (2008). How to deal with the challenges of linking a large number of individual national models: the case of the AGMEMOD Partnership Welches sind die Herausforderungen bei der Verknüpfung einer großen Anzahl von nationalen Modellen: Das Beispiel der AGMEMOD Par. Agrarwirtschaft, 57(8), 373–378.

12.7 Germplasm collections

Begemann, F. (2011). documentation and information of genetic resources in Europe Experiences from EURISCO Questions for EFABIS? Existing PGR information structure. http://www.rfp-europe.org/fileadmin/SITE_ERFP/WG_Docu/WGdocu_Begemann_EFABIS_April2011.pdf

Endresen, Dag T. F. & Knüpffer, H. (2012). The Darwin Core Extension for Genebanks Opens Up New Opportunities for Sharing Germplasm Data Sets, pp. 12–29.

Hintum, T. V., & Begemann, F. (2008). The European ex situ PGR Information Landscape, (July), 1–9. Retrieved from http://www.epgris3.eu/docs/activities/2-06/Paper European PGR Information Landscape V4 0.pdf

Hintum, TJL van. (2010). Innovation in Conservation, How Information Technology Tools Improve the Ex Situ Management of Plant Genetic Resources. ISHS Acta Horticulturae 918: XXVIII International Horticultural Congress on Science and Horticulture for People (IHC2010): III International Symposium on Plant Genetic Resources pp. 29–33. Retrieved from http://www.actahort.org/books/918/918_1.htm

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., Morrison, N., et al. (2010). Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. AoB plants, 2010, plq008. doi:10.1093/aobpla/plq008

van Hintum, Theo, & Knüpffer, H. (2010). Current taxonomic composition of European genebank material documented in EURISCO. Plant Genetic Resources, 8(02), 182–188. doi:10.1017/S1479262110000158