Building the CIARD framework for data and information sharing

Background Note for the e-Consultation April 4 - 15

The international CIARD initiative (http://www.ciard.net) is working to make agricultural research information publicly available and accessible to all. Among its actions are advocating for and promoting open access, improving applicability and enabling effective use of data and information in agricultural research and innovation.

Challenges and opportunities

Knowledge generated from agricultural research should be easily accessible and so easily taken up by agricultural researchers and the agricultural community at large. This issue is always more vital in the face of rapidly shifting challenges such as climate change, trans-boundary pests and diseases, effectively use of water for agriculture, combating desertification and degradation of fragile soils and managing agricultural biodiversity.

Conventional pathways for communicating research outputs, through scientific publications (journals etc) and face to face events, have been enriched by new digital formats including the newer social media such as wikis, informal blogs and online communities of practice. The volume of outputs of agricultural research such as scholarly or informal publications, blogs, discussion forums, institutional directories and very importantly raw data, is rapidly increasing.

Whereas an increasing amount of information becoming available represents an enormous opportunity, a challenge exists in the scattered way in which this information is being made available. The sources of information have become more heterogeneous and so coordination and coherence has become more problematic, making it more difficult to share data and information efficiently within and between scientific communities and more importantly with the wider range of stakeholders involved in agricultural innovation systems.

New approaches have to be adopted to facilitate sharing without resorting to tight coordination and centralization, which, for political and logistical reasons, are not foreseeable in the current spontaneous proliferation of data sets and services in agriculture.

This is where the current trends in information technology and web practices represent a unique opportunity to develop a coherent framework for sharing agricultural information. Approaches like Web 2.0, Linked Data, the OAI-PMH protocol, etc., are all means of improving the interoperability of distributed information and datasets without requiring strict coordination or the omologation of the software and data environment.

The CIARD partners are organizing a series of international consultations (through electronic and face to face events) to build a framework for action in making outputs of agricultural research truly accessible through coherent management, sharing and exchange of knowledge, information, and data. This background note is the basis for a discussion on the e-Agriculture platform from April 4 to 15. The e-consultation will be followed by an expert meeting in Beijing in June 2011 with the same title. These two events will be an exercise in
describing the current status and analyzing the needs for tools, standards and infrastructures, leading on to defining future actions.

**Background Figures**

Scientific publication and data production is growing at a much faster rate than ever before. Figure 1 charts the increase in numbers of articles indexed by MEDLINE from 1950 to 2010. The rate of increase in publication has clearly risen, especially since the year 2000, but the production of scientific papers is only the tip of the iceberg.

![MEDLINE-indexed articles published per year](image)

Figure 1. MEDLINE-indexed articles published per year

Behind growing numbers of scholarly publications is a growing amount of scientific data. Furthermore, scholarly papers are no longer the only way in which scientific information is exchanged. Researchers are using more social platforms such as blogs to discuss results before they are published in scholarly journals or after they have been published.

Data on publication rates in agriculture are not readily available, but it is quite clear that the trend will be similar, although perhaps less dramatic, to the other life sciences which are monitored by MEDLINE.

---

The second graph shows that the steeper increases in publication rates since 2000 is mainly due to the entrance of a few large countries with so-called “emerging economies”,

Figure 2. Thomson Reuters. Web of Science Database

especially China, Brazil and India, into the scientific mainstream. Taking 1990 as a base, Brazil has increased its scientific production by 800% and China by 1200%. The growth of these key new players in the scholarly communication arena has made interoperability a more important global issue, with special regard to the handling of languages other than English.

**Objectives of the e-Consultation**
The CIARD partners aim to provide an opportunity to discuss some specific questions that will provide resource material for the expert consultation that will follows in Beijing. The e-consultation has the following objectives:

- To increase awareness about opportunities and technical options for creating an infrastructure for data interoperability;
- To create a list of discussion points about the current state of the art of sharing and interoperability in the agricultural research context as basis for the discussions of the expert consultation in Beijing;
- To provide suggestions for actions that will enable increased sharing and exchange of data in agricultural research and innovation.

**The Topics**
The e-consultation will consider the following topics/questions over two weeks, focusing on two questions in each week:

1. What are we sharing and what needs to be shared?
2. What are the prospects for interoperability in the future?
3. What are the emerging tools, standards and infrastructures?
4. What actions should now be facilitated by the CIARD Task Forces?
Baseline Material for the Topics of the e-Consultation

1. What are we sharing and what needs to be shared?
The landscape of information and data flows and repositories is multifaceted. Peer reviewed journals and scientific conferences are still the basis of scholarly communication, but science blogs and social community platforms become increasingly important. Research data are now increasingly managed using advanced technologies and sharing of raw data has become an important issue. This topic thread will address and discuss details about the types of information that need to be shared in our domain, e.g.:

- Information residing in communications between individuals, such as in blogs and community platforms supported by sources such as directories of people and institutions;
- Formal scientific data collections as published data sets and their associated metadata and quality indicators, peer-reviewed scholarly journals or document repositories;
- Knowledge ‘derivatives’ such as collections of descriptions of agricultural technologies, learning object repositories, expertise databases, etc.;
- And surely more....

Figure 3. Schema of data repositories and flows in agricultural research and extension. Data flows exist between distributed repositories of the same type, and also between repositories of different types.

There are several interesting examples of successful data exchange between distributed datasets, and some of them in the area of agricultural research and innovation. There are
also ambitious attempts that still have to live up to expectations. A common characteristic of most examples is that they are based on specific ad-hoc solutions more than on a general principle or architecture, thus requiring coordination between “tightly coupled” components and limiting the possibilities of re-using the datasets anywhere and of replicating the experiment.

In some areas there are global platforms for sharing and interoperability. Some of these address the need to access scholarly publications, mostly those organized by the publishers, and others address the interfacing of open archives. With regard to standards and services in support of interoperability, there are several very successful initiatives, each dealing with different data domains. Among document repositories, the most successful initiative is surely the Open Archives Initiative (OAI) Protocol for Metadata Harvesting used by a global network of open archives. The strength of this movement is changing the face of scholarly publishing. Geospatial and remote sensing data have strong communities that have developed a number of wildly successful standards such as OGC that have in turn spurred important open source projects such as GeoServer. Finally, in relation to statistics from surveys, censuses and time-series, there has been considerable global cooperation among international organizations leading to initiatives such as SDMX and DDI, embraced by the World Bank, IMF, UNSD, OECD and others.

Singer System\(^3\), GeoNetwork\(^4\), and GeneOntology Consortium\(^5\) are examples of successful initiatives to create mechanisms for data exchange within scientific communities. The SDMX\(^6\) initiative aims to create a global exchange standard for statistical data.

There are more examples, but these advanced systems cannot have a strong impact on the average (smaller, less capacitated) agricultural information systems, because overall there are no easy mechanisms and tools for information systems developers to access, collect and mash up data from distributed sources. An infrastructure of standards, web sevices and tools needs to be created.

2. What are the prospects for interoperability\(^7\) in the future?

“Interoperability” is a feature both of data sets and of information services that gives access to data sets. When a data set or a service is interoperable it means that data coming from it can be easily “operated” also by other systems. The easier it is for other systems to retrieve, process, re-use and re-package data from a source, and the less coordination and tweaking of tools is required to achieve this, the more interoperable that source is.

Interoperability ensures that distributed data can be exchanged and re-used by and between partners without the need to centralize data or standardise software.

Some examples of scenarios where data sets need to be interoperable:

\(^3\) Singer System \(\text{http://singer.cgiar.org/}\) Last accessed March 2011
\(^6\) SDMX \(\text{http://sdmx.org/}\) Last accessed March 2011
\(^7\) \(\text{http://en.wikipedia.org/wiki/Interoperability}\)
- transfer data from one repository to another;
- harmonize different data and metadata sets;
- aggregate different data and metadata sets;
- virtual research environments;
- creating documents from distributed data sets;
- reasoning on distributed datasets;
- creating new information services using distributed data sets.

There are current examples of how an interesting degree of internal interoperability is achieved through centralized systems. Facebook and Google are the largest examples of centralized systems that allow easy sharing of data and a very good level of inter-operation within their own services. This is due to the use of uniform environments (software and database schemas) that can easily make physically distributed information repositories interoperable, but only within the limits of that environment. What is interesting is that centralized services like Google, Facebook and all social networks are adopting interoperable technologies in order to expose part of their data to other applications, because the huge range of social platforms is distributed and has to meet the needs of users in terms of easier access to information across different platforms.

Since there are social, political and practical reasons why centralization of repositories or omologation of software and working tools will not happen, a higher degree of standardization and generalization (“abstraction”) is needed to make data sets interoperable across systems.

The alternative to centralization of data or omologation of working environments is the development of a set of standards, protocols and tools that make distributed data sets interoperable and sharing possible among heterogeneous and un-coordinated systems (“loose coupling”).

This has been addressed by the W3C with the concept of the “semantic web”. The semantic web heralds the goal of global interoperability of data on the WWW. The concept was proposed more than 10 years ago. Since then the W3C has developed a range of standards to achieve this goal, specifically semantic description languages (RDF, OWL), which should get data out of isolated database silos and structure text that was born unstructured. Interoperability is achieved when machines understand the meaning of distributed data and therefore are able to process them in the correct way.
3. What are the emerging tools, standards and infrastructures?

The new paradigm for interoperability on the web and for building the basic layer for a semantic web is the concept of Linked Open Data\(^8\) (LOD).

Instead of pursuing ad hoc solutions for the exchange of specific data sets, the concept of linked open data establishes the possibility to express structured data in a way that it can be linked to other data sets that are following the same principle. Examples of an extensive use of “linked open data” technologies are the NYT or the BBC news service. Some governments too are pressing heavily to publish administrative information as LOD.

![Figure 4. The Linking Open Data cloud diagram](image)

The technology of LOD is based on W3C standards such as the “Resource Description Framework\(^9\)” (RDF), which facilitates the exchange of structured information regardless of the specific structure in which they are expressed at the source level. Any database can easily be expressed using the RDF, but also structured textual information from content management systems can be expressed in RDF. The presentation of data in RDF makes them understandable and processable by machines, which are able to mash up data from different sites. There are now mainstream open source data management tools like Drupal or Fedora commons which already include RDF as the way to present data.

Within the area of agricultural research for development an infrastructure to facilitate the production of linked open data is needed. The four key elements to make this possible are:

- a registry of services and data sets (CIARD RING, http://www.ring.ciard.net);
- common vocabularies to facilitate automatic data linking (thesauri, authority files, value vocabularies);
- technology (content management systems, RDF wrappers for legacy systems);
- training and capacity development.

---


\(^9\)Resource Description Framework [http://www.w3.org/RDF/](http://www.w3.org/RDF/) Last accessed March 2011
4. What actions should now be facilitated by the CIARD Task Forces?

There are immediate actions and strategic interventions.

Immediate actions:

- Registering services at the CIARD RING: the CIARD RING is a platform on which information systems and data sets can be registered and technical details about them can be provided together with instructions on how to use these sources, making their “interfaces” (parameters, formats etc.) transparent for others.
- Extensively using shared vocabularies and frameworks: well known common vocabularies are already available for the description of data. Generic vocabularies like “Dublin Core” or “FOAF” and specific vocabularies like AGROVOC and the Library of Congress Subject Headings are accessible openly on the web. The use of concepts from those common vocabularies will enormously facilitate the future production of linked data. Statistical data finds a coalition of international partners in the SDMX initiative. GIS standards such as OGC continue to lead the pack in sheer production quantity of interoperable data points.
- Creating document repositories using existing data exchange protocols such as OAI-PMH.
- Documenting and reporting successful examples of interoperability.

Strategic interventions:

- A blueprint is needed for a global infrastructure for data exchange in agricultural research for development. This blueprint should be also the basis for mobilizing financial resources.
- Ad hoc working groups could be established for specific areas.
- A series of events could be organized for advocacy and capacity development.

Participation in the e-Consultation

Contributions can be posted on the e-Agriculture Community platform at http://www.e-agriculture.org/ by registered members of the community. Contributions should ideally be provided in English, although French and Spanish will also be accepted and will be summarized in English. At the end of each week a summary of discussion of each question will be produced and distributed for additional inputs and comments. The final report of the consultation will acknowledge all contributors, and will be published electronically on the Community’s platform. The report will serve as a background document for the expert consultation in Beijing to be held in June 2011.

If you are interested to contribute to this e-consultation, please register on the e-Agriculture portal and look out for announcements on the site. Please feel free to contact us at info@e-agriculture.org should you need any additional information or assistance in registering.

Looking forward to your active participation,

Johannes Keizer (FAO)
Ajit Maru (GFAR)
Valeria Pesce (GFAR)
Stephen Rudgard (FAO)
## Acronyms

- BBC - British Broadcasting Corporation
- CIARD - Coherence in Information for Agricultural Research for Development
- CIARD RING - CIARD Routemap to Information Nodes and Gateways
- DDI - Dynamic Data Integration
- FOAF - Friend of a Friend
- GIS - Geographic Information System
- IMF - International Monetary Fund
- LOD - Linked Open Data
- MEDLINE - Medical Literature Analysis and Retrieval System Online
- NYT - New York Times
- OAI - Open Archives Initiative
- OECD - Organisation for Economic Co-operation and Development
- OGC - Open Geospatial Consortium
- OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
- OWL - Web Ontology Language
- RDF - Resource Description Framework
- SDMX - Statistical Data and Metadata Exchange
- UNSD - United Nations Statistics Division
- W3C - World Wide Web Consortium
- WWW - World Wide Web