Technical Issues Paper

A Framework for Sharing Data and Information for Global Agricultural Research

October 2011

TABLE OF CONTENTS

BACKGROUND AND CONTEXT

Improved access to data and information is essential if present-day global problems such as climate change and sustainable and more effective use of natural resources and biodiversity are to be addressed. Currently data and information are often not immediately accessible and the benefits that could be derived from their use are restricted. Better access would result in better use and would stimulate effective research and enhanced innovation. Moreover, duplication of effort would be reduced and broader participation would ensue. In turn, this would result in greater equity of access to and use of data and information within and among communities. Ultimately agricultural research can have a greater impact than at present.

CIARD (Coherence in Information for Agricultural Research for Development) is a global movement that was established in 2008 with the specific aim of enhancing access to data and information in the public domain to improve development based on agricultural research results (see **http://www.ciard.net**). A more coordinated approach in enabling and supporting accessibility of data and information in the public domain globally would relieve many smaller organizations of the need to develop and operate their own systems from first principles. Through a series of CIARD-facilitated face-to-face and virtual consultations in 2010 and 2011, broad consensus on needs for sharing data and information at national, regional and global levels was achieved among a diverse group of actors.

CIARD partners organized two specific international consultations in 2011 to understand and analyze the features of data and information sharing problems. The major product of these consultations was the design of a framework and action plan to improve data and information sharing, and enhanced collaboration, worldwide.

SHARING DATA AND INFORMATION

TYPES OF DATA ALREADY SHARED

The types of data already shared by agricultural organizations include:

- bibliographical descriptions of research outputs
  (e.g. **http://agris.fao.org**);
- information about standards, tools, services, datasets and events
  (e.g. **http://aims.fao.org**, **http://ring.ciard.net** and **http://www.agrifeeds.org**);
- data on plant genetic resources
  (e.g. **http://www.sgrp.cgiar.org** and **http://www.genesys-pgr.org**);
- agricultural science and technology indicators
  (e.g. **http:/www.asti.cgiar.org**);
- agricultural factsheets and e-books
  (e.g. **http:/www.cabi.org/cabebooks**);
- locally produced research re-packaged for wide dissemination
  (e.g. **http:/www.gains.org.gh**); and
- soil and land-use maps
  (e.g. **http://www.fertimap.ma**).

SHARING RESEARCH DATA

The web represents an ideal environment for sharing the results of agricultural research for development and innovation in the form of data and information. Such data and information cover a huge range of subjects and are provided in numerous formats. The major problem with respect to such potentially invaluable data and information is that they are often not freely available. Research results are often retained by researchers and their organizations, especially when associated with intellectual property and financial concerns. Others restrict access for fear of data theft, plagiarism and misinterpretation. Even when there is a willingness to share data and information freely, there are problems with formatting and presentation that hinder access, interpretation and further use. Despite this, there is a movement in the research community to make the results of publicly funded research more generally accessible and useable.

SHARING HIDDEN DATA AND INFORMATION

Scientists and researchers increasingly communicate among themselves and with the public through informal channels such as web-based newsletters, social networking sites and blogs. This allows much more data and information to enter the public realm than is possible if research results are solely presented at specialist meetings and published in specialist journals. However, a substantial amount of potentially useful data and information remain hidden and can only be accessed through personal contacts, and knowledge leaves an organization when its staff leave.

SHARING WITH FARMERS

Farmers are the ultimate users of data and information generated through agricultural research. They have interest in numerous areas related to their livelihood systems and require data and information to be made available in a wide range of formats and languages in a useable form. This is a virtuous two-way process as farmers communicate their indigenous knowledge and their feedback to researchers. For development to have an impact, it is essential that there is adequate flow of data and information among the various communities, but particularly between farmers and researchers, especially given that on-farm research is now providing huge quantities of data on crops, weather, soils, pests etc.

SHARING WITH MACHINES

In the present day much of the data and information associated with agricultural research and development is maintained electronically. Linking through networks, the Internet and electronic devices, allow the data and information to be distributed and shared. Such systems also allow data and information from diverse sources and on diverse subjects to be linked. Enhanced data and information delivery to end-users has become possible through using various protocols and formats, such as:

1. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for sharing metadata records. (**http://www.openarchives.org/OAI/openarchivesprotocol.html**)
2. Linked Open Data (LOD) for integrating multi-site information. (**http://www.w3.org/DesignIssues/LinkedData.html**)
3. Rich Site Summary (RSS) for distributing news items on the web. (**http://en.wikipedia.org/wiki/RSS**).
4. Resource Description Framework (RDF) for describing information in an easily integrated form. (**http://en.wikipedia.org/wiki/Resource_Description_Framework**).

REUSABILITY OF SHARED INFORMATION

The value of data and information depends on reliability and on users' needs. Although sharing everything allows users to sift and sort and eventually extract useful data and information, the process is time-consuming and expensive. In order to be of optimal use, the quality of data and information, in terms of correctness and completeness, needs to be assessed. Only reliable data and information can be used to maximum effect.

INTEROPERABILITY - A KEY TARGET

INTEROPERABILITY DEFINED

Interoperability is feature of datasets and the services that provide access to them. Interoperability determines the extent to which data and information can be retrieved, processed, reused and repackaged by other systems. The easier this is, the more interoperable the source. Good interoperability indicates that data and information can be exchanged and used among partners without having to store them centrally or adopt common software. Interoperable data and information can be gathered from numerous sources and put together in Internet portals and interactive virtual research environments, for example, to allow greater insight into the relationships among diverse factors and to work collaboratively on widely disbursed data.

INTEROPERABILITY ON THE WEB

Interoperability is relatively easy within closed systems, including large ones such as Google and Facebook. This requires particular information formats and custom-built software. It is also possible in the highly heterogeneous web environment through use of generic 'semantic web' standards. Concentration of data and information in large centralized repositories is not optimal for numerous social, political and practical reasons. Because data and information formats evolve continuously, systems quickly become obsolete. Therefore, exchange of data and information is best served by standard representations that support sharing among loosely coupled sources. Such an approach embraces diversity rather than tries to eliminate it.

Some of the key elements for interoperability through common standards are:

1. GLOBALLY UNIQUE IDENTIFIERS

The role of URIs (Uniform Resource Identifiers) is central to interoperability. URIs name entities, which allows them to be cited and linked. 'Expandable descriptions' represent new aggregations of data and information that derive from multiple repositories (Box 1).

Box 1: Trending technologies

1994: "World Wide Web" and URIs (Uniform Resource Identifiers): First proposed in 1989, an Internet-based network of documents linked using globally unique URIs, which took off with the spread of graphic Web browsers in 1994.

2000: "Semantic Web": As proposed by Tim Berners-Lee, inventor of the Web (of documents), the notion of a web of structured data meaningfully processable by machines. See: **http://www.w3.org/standards/semanticweb/**.

2001: OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting): A computer protocol for aggregating ("harvesting") metadata records over the Web from multiple repositories. See: **http://www.openarchives.org/pmh/**.

2004: RDF (Resource Description Framework): First introduced in 1999, a key Semantic Web standard for data interchange that achieved widespread use after the release of a major revision in 2004. See: **http://www.w3.org/RDF/**.

2006: RSS (Really Simple Syndication): First introduced in 1999, a format for disseminating news items (or Rich Site Summaries) which took off with support by major Web browsers after 2005. See: **http://en.wikipedia.org/wiki/RSS**.

2008: "Linked Data": First introduced in 2006, the notion of data expressed using RDF and URIs - also known, when published world-readably on the Web, as "Linked Open Data" (LOD). See: **http://linkeddata.org/**.

2. A COMMON GRAMMAR

Resource Description Framework (RDF) (**http://www.w3.org/RDF**) technology allows publication of generically understandable data. Linked Open Data (LOD) (**http://linkeddata.org**) uses URIs to establish browsable links between diverse datasets and tag resources according to precise search concepts. LOD uses entities and metadata, which describe those entities. A broad range of interoperable metadata can be published using RDF.

3. SHARED VOCABULARIES

Interoperability does not depend solely on a shared grammar. Data and information must make use of shared vocabularies.
These include Dublin Core metadata standard (**http://www.dublincore.org**) and RDF-enabled thesauri such as AGROVOC (**http://aims.fao.org/agrovoc**). With the minimum amount of information (e.g. title, date, location and topic) an 'event' becomes interoperable through the Agrifeeds system, and can be disseminated widely. Mapping between vocabularies (alignments) such as AGROVOC and the National Agricultural Library Thesaurus (**http://agclass.nal.usda.gov**) allows interoperability among diverse concept schemes.

PRODUCTS THAT DEPEND ON INTEROPERABILITY

A significant amount of data and information is available on the web and is interoperable. An example is the FAO Global Hunger Map, which relies on data from numerous sources (**http://www.fao.org/hunger/en/**).
Moreover, the BBC uses interoperable data and information to provide its customers with the news they require (**http://www.bbc.co.uk/nature**).

An example of Linked Open Data technology from the agricultural domain is AGRIS (**http://agris.fao.org**).

Data and information that are linked and interoperable improve access to users. Increased sharing of such data and information encourages use and stimulates innovation. The benefits accrue among all stakeholders, including researchers, advisors, policy-makers and ultimately farmers. Linked and interoperable data and information generate new perspectives.

EMERGING TOOLS, STANDARDS AND INFRASTRUCTURES

The Linked Data Approach represents a pathway that leads information providers towards progressively higher levels of interoperability (see Box 2). There is a continuous sequence of choices available for making data and information linked and interoperable, and the degree to which they can be shared depends on making those choices.

Box 2: Five steps to open data and information

1 star	Your content is **available on the Web**, in whatever format, under open licenses.
2 star	Your content is available as **machine-readable structured data** [*i.e. MS Excel table is better than an image of the same*].
3 star	Your content is available in **non-proprietary formats** [*i.e. Comma-Separated-Values (CSV) format in preference to MS Excel*].
4 star	You use **RDF standards** and URLs (URIs) to identify your content so that people can point to it.
5 star	Your content is **linked through RDF** to other people's content to provide context and add value.

EXISTING DATA EXPOSED AS RESOURCE DESCRIPTION FRAMEWORK (RDF)

One of the simplest staring points in data and information sharing is to make an existing database available as linked data by using an "RDF wrapper"[1]. This does not require the database management software to be changed. If the database model does not allow complete mapping to RDF, those elements that can be mapped become the focus.
The AGRIS (**http://agris.fao.org**) database, for example, accepts data that are not specifically formatted but which can be mapped to standard RDF vocabularies.

RDF GENERATED BY CONTENT MANAGEMENT SYSTEMS AND TOOLS

Mainstream open-source platforms such as Drupal (**http://www.drupal.org**) and Fedora are Content Management Systems that support publication of structured data and information in RDF. Drupal can include an OAI-PMH module to harvest content from several providers. An 'AgroTagger' tool developed by the Indian Institute of Technology (Kanpur) describes text content using AGROVOC concepts and natural language processing. AGROVOC VocBench provides an online vocabulary editing and workflow tool for maintaining vocabularies in dispersed environments and multiple languages.

BEST-PRACTICE SERVICES

The consultations highlighted several services, including VIVO (**http://www.vivoweb.org**), which 'facilitates interoperability between people' by providing information on scientists, academic departments, courses, grants and research publications. Another service that was discussed was eScienceNews (**http://www.esciencenews.com**), an aggregator for scientific news and blog postings that uses natural language processing and machine learning to annotate web contents semantically. Such services enhance sharing and use of data and information.

COMPLIANCE WITH STANDARDS

'Compliance with standards' should represent the norm. Much depends on access to qualified staff that are not afraid to venture into new territory and use the tools and systems that are available to maximize interoperability. This is a fast-moving area and it is necessary to stay abreast of developments.

NEW PROSPECTS FOR CLOUD COMPUTING

Applications and storage space are becoming increasingly available in web-based server banks, known as 'the cloud'. This allows organizations with relatively limited capacity and funds to make full use of powerful computing services. Cloud computing could help CIARD-RING (**http://ring.ciard.net**) to serve its broad community better by managing and aggregating the data and metadata of its providers, which could then be used to develop value-added services.

1- A software layer that translates a proprietary data format into a generic data format using common standards.

A FRAMEWORK FOR DATA AND INFORMATION SHARING - ACTION AREAS

The consultations identified the importance of setting up a framework to allow agriculture-related data and information to be shared. The framework comprises three dimensions:

1. Technical issues and technologies
2. Institutional and organizational issues
3. Championing change in policy and practice

The framework should accommodate various types of data and information produced by a range of organizations operating in diverse economic and political environments. A more coordinated approach would relieve many smaller organizations of the need to fund and operate their own systems for sharing data and information, and would help many organizations in developing countries that experience difficulties in recruiting and retaining qualified staff. A framework based on a good coordination mechanism should be able to link activities and organizations on a global scale.

CIARD takes full account of the developments in data and information management and communication, and advocates and facilitates data and information sharing in the public domain. It represents a platform for collaboration upon which the framework can be based. Various areas requiring action have already been identified that should improve global sharing and exchange of data and information related to agricultural research.

TECHNICAL ISSUES AND TECHNOLOGIES

There was an explicit demand for CIARD to provide advice on:

1. When and how to use open or proprietary data formats.
2. Whether to use Content Management Systems.
3. How to describe specific types of information.
4. When to use a traditional library system, an open-archive repository or a tailor-made application.
5. How to use multimedia social reporting for effective information sharing.

It was not expected that CIARD should promote a uniform approach to such issues among organizations, but should be able to detail options and provide guidance when needed.

Action Area 1 - Services, Tools and Infrastructure

CIARD partners need to engage more content providers so that the CIARD-RING becomes more comprehensive and the registries for agricultural information become easier to use. The user community should be able to add information on the quality and usefulness of information. If a 'Tools-Wiki'[2] were to be set up on the web, it would allow less well endowed partners to contribute. The tools would include those most commonly needed by content providers. The CIARD community should also be able to provide plug-ins, customized versions of useful software and prototypes for information-sharing platforms in a variety of languages. The utility and value of cloud services could be assessed through a survey of the CIARD community.

2 - A community Wiki that describes information management tools and collects experiences for evaluation.

Action Area 2 - Standards and Systems Architecture

Standards and systems architecture should be approached collaboratively, building on existing platforms such as AIMS (Agricultural Information Management Standards - **http://aims.fao.org**). Collaboration among the international community should endorse open standards for protocols, ontologies and vocabularies that are used across a wide range of domains. Descriptive templates need to be developed or adapted for high-priority information, and recommendations should be formulated for:

1. Metadata; to describe datasets.
2. Data packaging and documentation formats.
3. An automatic tagging/indexing service.

Non-technical guidelines are also needed for producing LOD with an associated system of mapped vocabularies to enable better classification and organization of data and information. Automatic translation services are also needed. Other key requirements relate to digital conservation of data and information, and quality of data and information provided.

INSTITUTIONAL AND ORGANIZATIONAL ASPECTS

Many organizations have established policies and practices for general data and information management that could be adapted for use with agriculture-related data and information. However, capacity often needs to be built in the technical aspects of interoperability and practical management of data and information. To help in this regard, CIARD could arrange partnerships with organizations that develop capacity, including education programmes in agricultural information and communication management.

Development and introduction of new technologies and processes will place greater responsibilities on information professionals to institute new structures and practices in the workplace. There will become a greater need also for specialists to communicate, transform and translate information that derives from scientists and which should benefit a range of users. Many actors in the agricultural sphere will need to be involved to ensure that this is done effectively. Improved information flow might require new organizations to be formed or, at least, established ones to be adapted to modern needs.

Action Area 3 - Policies, Strategies and Institutional Structures

CIARD partners should continue promoting the CIARD manifesto for increased access to information and greater sharing. A review should be undertaken in support of national initiatives on:

1. Policies on access to public goods and copyrights.
2. Incentives and benefits of sharing data and information.

International agreements and generic guidelines for policy design should be developed. Organizations should be encouraged to enact legislation that encourages open access to information and interoperability, and information sharing should be embedded in organizational systems and processes. In addition to making explicit the benefits from sharing data and information, with due consideration given to Creative Commons licensing, organizational copyright and IPR statements should be developed and made public.

Action Area 4 - Development of Skills and Competencies

Training needs require to be assessed at various levels and training programmes should be designed and implemented to improve better data and information sharing. All modern methods, including e-learning, should be employed in a coordinated effort among CIARD partners. Coordination should take place from the international level to the local level, making sure that the five steps listed in Box 2 are adhered to. Current platforms offering guidance to the CIARD community, including AIMS and IMARK (Information Management Resource Kit), should continue to be supported, and links should be forged among them.

Action Area 5 - Appropriate Organizational Structures and Work Practices

Adequate investment has to be made in computer hardware and software. In addition, skills and content should be developed in order to improve practices in data and information sharing. This is best accomplished by setting out and adhering to a series of norms, standards, rules and regulatory mechanisms within the setting of appropriate organizational structures.

Action Area 6 - Global Improvement of Data and Information Flows

Once the framework is designed, it should be used by the agricultural research and development community to improve data and information flows. This should take place at local, national, regional and global levels, realizing that organizations will be at various stages of development and will consequently have to adapt according to circumstance. There will be no single ideal approach to data and information sharing; flexibility and adaptability will have to be built into any system. Many organizations in the public domain, including the CGIAR (Consultative Group on International Agricultural Research), FAO and GFAR (Global Forum on Agricultural Research), will play a leading role in improving flows of data and information in important areas such as germplasm, agronomy and climate change.

CHAMPIONING CHANGE IN POLICY AND PRACTICE

CIARD partners organized two specific international consultations in 2011 to discuss issues of data and information sharing. It was recognized that a major obstacle to progress was the lack of support available to persuade managers of organizations that change represented advance. It was considered that advocacy would help address the issue.

Advocacy for sharing data and information requires addressing a broad range of stakeholders, who fall into three major categories:

1. Policy makers and research managers who need to be convinced of the benefits that derive from sharing data and information from many disciplines.
2. Information specialists, who are relatively few in the area of agricultural science, also need to become more involved. They need to push for policy and strategy changes that result in greater investment in data and information sharing.
3. The generators and users of data and information. Many already use digital tools related to social media and web-based services that could be adapted for greater sharing of data and information on agricultural research.

In order for advocacy to be effective, organizations need to be able to demonstrate clearly the benefits to be had from greater openness and improved sharing of data and information. This requires specification of concrete outputs and outcomes that result from adherence to a well-defined framework.

CIARD encourages collaboration and has set out its aims in its manifesto. CIARD is in a position to support advocacy at various levels for increased sharing of data and information. The consultations identified several areas of action where CIARD could make an impact.

Action Area 7 - Advocacy and Evidence

Advocacy initiatives at global and regional levels, especially regarding interoperability, need to be documented and a tailored advocacy toolkit should be developed through collaborative stakeholder action. Case-study evidence, including possible cost-benefit analyses of sharing and interoperability, would strengthen the case for increased data and information sharing. It is possible that evidence exists for outcomes that have impacted economic, social or environmental circumstances. These need to be documented and shared. Use should also be made of high-level events, such as GCARD (Global Conference on Agricultural Research for Development) in October 2012, to advocate data exchange and interoperability among senior decision-makers. Traditional donors and other potential investors should also become informed.

A general outline was drawn up for defining and taking forward an advocacy programme for greater openness, which could be adapted to local circumstance. The aim is to attract funding by:

1. Assessing the obstacles to change.
2. Assigning roles and tasks to senior staff champions.
3. Defining and prioritizing advocacy targets.
4. Defining key advocacy messages.
5. Providing evidence to support advocacy.
6. Organizing advocacy opportunities around planned events.
7. Documenting case studies at the organization level.

STRENGTHENING THE CIARD COMMUNITY AND ITS ROLE

The CIARD movement is managed by all its participating organizations and its sponsors. While coherence is sought in data and information communication, it is also necessary among the contributing organizations. This will be achieved through the technical framework, but there remain opportunities to strengthen the sense of community among stakeholders. This will be best achieved through sharing and discussing experiences and ideas among the CIARD partners.

Action Area 8 - Partnerships and Information Managers

CIARD's international and regional partners need to act on a global scale to promote openness and increased sharing of data and information related to agricultural research. In this respect CIARD should become appreciated as a multi-dimensional learning movement that recognizes the value of individual contributions in the field of data and information sharing. To promote collaboration and discussion, it will be necessary to establish a virtual platform for the community built around existing entities such as AIMS and e-Agriculture sites (**www.e-agriculture.org**), and CIARD's own site. This will enable a wealth of data, information and experience to be shared, spanning the range from technologies and policies to case studies and success stories.