"Building the CIARD Framework for Data and Information Sharing": Preliminary results of an International Expert Consultation

About CIARD and the 2011 International Expert Consultation. This discussion paper presents preliminary results from an International Expert Electronic Consultation on developing a CIARD Framework for Data and Information Sharing held online from 4 to 15 April 2011 by Coherence in Information for Agricultural Research for Development (CIARD). CIARD is a movement by major actors and stakeholders in Agricultural Research for Development (ARD) devoted to making Agricultural Research Information (ARI) publicly available and accessible to all by helping member organizations disseminate their knowledge more effectively, especially to innovators addressing key challenges of agricultural development and food security. [1] A workshop from 20 to 23 June 2011 in Beijing will continue the consultation and chart next-step actions. [2]

Question 1: What are we sharing and what needs to be shared?

Types of data already shared. Types of data already shared by CIARD partners include bibliographical descriptions of research outputs (e.g., AGRIS); information about standards, tools, services, datasets, and events (e.g., the CIARD Ring, AIMS Website, and AgriFeeds); data on plant genetic resources (e.g., SINGER, GENESYS); agricultural science and technology indicators (e.g., ASTI); agricultural factsheets and e-books (e.g., CABI); locally produced research re-packaged for wide dissemination (e.g., GAINS); soil and land-use maps (e.g., INRA Morocco); and remote sensing data (e.g., AREA Yemen).

Sharing research data. Sharing other products of research on the Web, including raw datasets and other re-usable results, is seen as essential for enabling innovation on important topics of agricultural research for development and food security such as desertification, managing the spread of pests and diseases, and biodiversity. Such data may include earth observation data and the results of field trials, surveys, or cultivar tests. Potentially breakthrough findings remain hidden within institutes, or even on home computers. While many scientists are in principle willing to share more of their data in the interest of better science, others regard research results as personal or institutional property, especially where there is an emphasis on securing patents. Others fear that their work could be "copied" without credit or that incompletely contextualized data could be misinterpreted. Providing proper documentation for research data requires more effort than busy scientists may be willing to provide. In order to justify the investment, scientists and their institutions need to see potential gains, for example in terms of visibility, reputation, or standing with donors. There is a significant movement in the research community towards making all publicly funded research open and transparent and to enable public access to all research data, though in the area of agricultural research for development, sharing is still quite limited.

Sharing "hidden" knowledge. Communication among scientists increasingly occurs through informal channels such as blogs and community forums, where "tacit" or "hidden" knowledge is made manifest. Other types of information currently "hidden" include planning documents, reviews, meeting minutes, posters, Powerpoint presentations, video clips, preliminary project results and — where academic programs exist — theses and teaching materials. Some institutes already use social media sites such as Flickr and Slideshare to post photographs and slide decks. Hidden knowledge can also be discovered by knowing whom to ask; finding people with specific subject or technical expertise can be supported by databases of experts.

Sharing with the users of research data — farmers. When asked, one organization of farmers expressed interest in information on pesticides (and biopesticides), seeds, hydroponics, crop recommendations, pest alerts, market prices, irrigation forecasts, and crop trends — not just in Web or print form, but with videos. The organization also requested information on research plans, ongoing projects, and research outputs. Where farmers are invited to view on-farm trials, communication between farmers and researchers can become a two-way channel in which potentially valuable

[1] http://www.ciard.net

[2] http://www.ciard.net/events/international-expert-consultation-building-ciard-framework-data-sharing

indigenous knowledge is elicited. Communicating with farmers requires outputs to be translated into local languages, simplified, and published in accessible formats — efforts less likely to be undertaken if the results are rejected as "not scientific" when presented for reward and promotion.

Sharing with machines. Standing between research information and its ultimate beneficiaries are the computer systems which deliver that information. In this sense, "machines" are the front-line consumers with which information is "shared". Modern Web technologies — protocols and formats such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH; for sharing metadata records), Linked Open Data (LOD; for integrating information between many sites), Rich Site Summary (RSS; for distributing news items on the Web) and Resource Description Framework (RDF; for describing information in a form that easy to integrate) — do not so much determine how information is delivered to end users as to encapsulate, structure, or link that information in ways that flexibly support its delivery to users downstream. (See "Trending Technologies" below.)

Reusability of information shared. The potential value of any given information to someone else is not always obvious, so some people recommend erring on the side of sharing "everything". Intelligent aggregators can sift through well-tagged chunks of information, extracting and re-packaging the information for new audiences or purposes (example: the extensionist who merely needs one photo from a long report) or re-packaged for different target audiences. The counter-argument is that sharing too much risks spreading scarce resources too thin while presenting users with too much information — much of which, in the absence of peer review, will be of dubious quality. Information is needed about what is being shared, preferably ranked for completeness and quality for the benefit both of information consumers and of creators of value-added services.

> **For discussion in Beijing.** For what purposes does ARI need to be shared? Priorities for sharing should be informed by purposes related to natural resource management, climate change, innovative agricultural practice, sustainable production, equitable access to markets, or the dissemination of research results. What are the social or cultural barriers to sharing, and how might those barriers be overcome?

Question 2: What are the prospects for interoperability in the future?

Interoperability defined. Interoperability is a feature of datasets — and of information services that give access to datasets — whereby data can easily be retrieved, processed, re-used, and re-packaged ("operated") by other systems. The less pre-coordination required to achieve this, the more "interoperable" the source. Interoperability ensures that distributed data can be exchanged and re-used among partners without needing to centralize data storage or adopt common software. "Mashing up" data from multiple sources can lead to new insights about relationships between factors such as weather, markets, crops, and geographic location. Interoperable data can more easily be pulled together into specialized services, such as crop portals and virtual research environments.

Interoperability on the (inevitably) diverse Web. Interoperability can be achieved within closed systems — in some cases very large closed systems, such as Google and Facebook — by using specialized information formats usable by custom-built software. However, interoperability can also be achieved within the highly heterogeneous environment of the Web on the basis of open and generic "Semantic Web" standards. For social, political, and practical reasons, the concentration of information in big, centralized repositories, using centralized tools, is both unrealistic and counter-productive. Computer applications and data formats will continue to evolve, virtually guaranteeing that any particular system in use today will sooner or later grow obsolete. Where data exchange based on ad-hoc solutions requires pre-coordination between tightly coupled components, exchange based on standard data representations supports sharing among heterogeneous, "loosely coupled" information sources —- an approach that copes with diversity not by trying to eliminate it, but by embracing it as inevitable.

Globally unique names (identifiers) for things. Central to the concept of open interoperability described here is the role of URIs as identifiers for "things" (resources). URIs give names to things, making them citable and "linkable." Metadata allows applications to see information through different lenses, re-packaging it into different aggregations or incorporating new information, as it becomes available, into "expandable descriptions."

Interoperability through a common "grammar" for data. Resource Description Framework (RDF) technology is used to publish data in a form is generically "understandable" by applications. Linked Data builds on the power of RDF by using globally unique identifiers (URIs) to establish browsable links between diverse datasets and tag resources with precise search concepts. In Linked Data, the boundary between "data" ("things") and metadata ("descriptions of things") is blurred. RDF can be used to publish interoperable metadata or, in principle, any other structured dataset, from earth observation data to financial spreadsheets.

Interoperability through the use of shared vocabularies. A shared grammar does not, of itself, ensure interoperability. To be fully interoperable, data must be expressed using shared concepts ("vocabularies") — whether well-known properties such as Dublin Core or topic identifiers from RDF-enabled thesauri such as AGROVOC. To be interoperable with AgriFeeds, for example, an event description must minimally include a title, date, location, and topic. Explicit mappings between vocabularies ("alignments") such as AGROVOC and the National Agricultural Library Thesaurus establish interoperability between entire concept schemes.

> **For discussion in Beijing.** How can new information and communications technologies (ICTs) be harnessed to provide information to the people who most depend on them? What specific policies and structures — open data repositories and trust organizations with requisite standards and norms — are needed at the global, regional, national, instititutional, even individual levels?

Question 3: What are the emerging tools, standards and infrastructures?

A continuum of choices from basic to advanced. Linked Data is not an all-or-nothing proposition but a continuum starting, at the low end, with simple choices. Tim Berners-Lee summarizes the Linked Data approach as a pathway leading information providers towards progressively higher levels of interoperability (paraphrased here):

⋆	**On the Web, open licenses.** Make your stuff available on the Web, in whatever format, under open licenses.
⋆⋆	**Machine-readable data.** Make your stuff available as machine-readable structured data; a table in Excel is better than just an image of the same.
⋆⋆⋆	**Non-proprietary format.** Use plain Comma-Separated-Values format (CSV) in preference to Excel.
⋆⋆⋆⋆	**RDF standards.** Use URLs (URIs) to identify your things so that people can point to them, and describe them using RDF.
⋆⋆⋆⋆⋆	**Linked RDF.** Link your data to other peoples data to provide context and add value.

Existing data exposed as RDF. One of the simplest starting points is to expose an existing database as Linked Data by using an RDF wrapper that does not require changing the underlying database management software. If the entity model of a database does not map cleanly to RDF, the mapping can focus on data elements of particular utility. The AGRIS Application Profile, for example, was intended to provide a target for such mappings while leaving underlying applications untouched.

RDF generated by Content Management Systems and tools. New Content Management Systems support the publication of structured data in RDF, such as the mainstream open-source platforms Drupal and Fedora. Drupal can be extended with an OAI-PMH module for harvesting content from providers. AgroTagger, developed by IIT Kanpur, uses natural-language processing to describe the content of a submitted text with AGROVOC concepts. The AGROVOC VocBench provides an online vocabulary editing and workflow tool for maintaining large vocabularies in highly distributed environments and in multiple languages.

Best-practice services. Services highlighted in the e-consultation include VIVO, a search engine which "facilitates interoperability between people" by providing information about scientists, academic departments, courses, grants, and research publications; and eScienceNews, an aggregator for news and blog postings of scientific interest which uses natural-language processing and machine learning to semantically annotate Web contents for enhanced discovery.

Compliance with standards. "Compliance with standards" frightens managers because it sounds expensive. Ideally, compliance with standards should "just happen" as a by-product of routine workflows and simple tools. Many research institutes lack IT specialists, or if they do have qualified staff, find that they resist exploring new approaches, such as RDF, which lie outside the comfort zone of familiar SQL and XML databases. Once trained, qualified staff often leave for positions that are better paid. To stand a chance of success, an open interoperability strategy must be based on tools that are easy to set up, use, and maintain.

New prospects for Cloud Computing. The increasing availability of applications and storage space in The Cloud (Web-based server banks) may mitigate this problem by allowing institutes with less capacity to implement advanced services without increasing local staff or computing power — a prospect to be explored by FAO in an upcoming EU project, AgINFRA. Cloud computing could help trusted organizations such as CIARD.RING better serve a broad community by managing and aggregating the data and metadata of its partners and using the data to develop value-added services.

> **For discussion in Beijing.** What capacities must be developed, and at what levels, to facilitate the creation of exemplary services in compliance with standards? What kinds of technical services can be externalized to servers in the cloud, and what capacities must be developed locally?

Question 4: What actions should be facilitated by CIARD Task Forces?

Re-think the role of information managers and of "communicators." Addressing the challenges of data and information sharing is not only a question of technology, but of institutions, culture, and processes. The introduction of new technology and processes may imply a more central role for information professionals, possibly in the context of "regional data transformation centers." In addition, there is a need for "communicators" with the knowledge and skill to "translate" between scientists and practitioners, creating ways to present scientific information in practice-oriented advisory services.

Provide advice and support for information management choices. There is a clear demand for advice from CIARD on basic choices such as when and how to use open versus proprietary data formats; whether to manage Web sites with Content Management Systems; how to describe specific types of information interoperably; when to use a traditional library system, OAI repository, or custom-build application; how to augment face-to-face meetings with multimedia social reporting; and which "star level" to target in a given situation given costs, difficulty, connectivity, and required level of ICT skill. The role of CIARD could be to describe available options, making the case for open solutions and providing information on specific tools — but with no expectation that partner organizations should all adopt a uniform approach.

Provide capacity development for adopters. Based on an analysis of capacity gaps, capacity development events could target key people for training both in the technical aspects of data interoperability and in practical methods for migrating data. Mentoring arrangements between institutions at different levels of the "star system" could provide a channel for the transfer of practical knowledge and experience. The CIARD Capacity Development Task Force could form partnerships with other organizations involved in developing capacity, such as new degree programs in agricultural information and communication management.

Package well-tested and popular tools into a CIARD toolkit or CIARD-hosted Website. CIARD could package well-tested and popular tools into a toolkit — a "filled shopping bag" with solutions that work "out of the box" to meet the most common requirements. The CIARD Website could point to tools customized for the agricultural research community such as AgriOcean DSpace and AgriDrupal. The Website might also host content on behalf of CIARD partners with less local capacity or facilitate the creation of data repositories at the national or even global level.

Improve CIARD Ring as a global "signpost" and directory for ARD information. Consistent follow-up with registrants of services on the CIARD Ring could ensure the provision of basic information such as location (for the map), URLs for OAI-PMH providers and RSS feeds, and subject headings. By adding to the current offering of tutorials on OAI-PMH and RSS, the Ring portal could become a one-stop source of practical information on how to implement the CIARD Pathways and improve interoperability.

Work with regional authorities and donors to advocate for targeted funding. CIARD should develop a blueprint for a global agricultural research infrastructure and use that blueprint to mobilize investments in a strategy for information sharing based on technologies for open interoperability and in capacity building of its champions. The blueprint should demonstrate the strengths of a Linked Data approach by highlighting success stories and showing how shared data has benefitted rural communities.

In conclusion, some brainstorming... Define Core Data Sets for topics such as Plant Breeding or Natural Resources Management. Identify "Trust Organizations" for holding specific types of data (as CGIAR, the Global Trust Fund, and the International Treaty on Genetic Resources for Food and Agriculture do for seeds). Ensure the long-term preservation of Data Repositories. Create something like the Open Access movement to energize people to take action, perhaps in the form of a Global Treaty on Information Sharing.

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/.

Prepared by Tom Baker, June 2011