CGIAR Data Summit, Nov 2013. Meeting Report

26 – 28th November, 2013, Rome

Report prepared by: K. Chapman, E. Crothall, M.Marus

Table of Contents

Introduction

The Data Summit was convened by the CGIAR Consortium Office, in partnership with FAO, to discuss the data management landscape with experts in various data domains – within and beyond CGIAR – and to:

- Collaboratively develop a practical roadmap for adoption of open data management standards as CGIAR moves forward;
- Discuss and obtain feedback on Draft CGIAR Open Access and Data Management Implementation Guidelines, with a focus on data management – especially Annexes 1 and 2;
- Agree on fundamental/general metadata standards and vocabulary that can apply to all data domains;
- Encourage each data domain to move towards agreement on domain specific standards (lists);
- Discuss an approach to consolidation of leading data management efforts (e.g. IATI, CCAFS, IBP);
- Formalize a CGIAR DM/data standards working group and an Open Access and Open Data working group and agree Terms of Reference.

The issues in data management are well known and were thoroughly explored during the summit. CGIAR and other partner organizations/initiatives (e.g. FAO, CIARD, IATI) will need to work in close partnership to ensure that data standards are agreed and implemented. There are excellent examples of Open Data within CGIAR (e.g. CCAFS, IBP) and CGIAR stakeholders should work closely to share lessons learned, and to upscale existing tools and processes.

While with hindsight, the original aim of developing a practical roadmap was somewhat optimistic, the meeting did make good progress and the following key themes and next steps were identified:

Key themes

1. While we work across different data domains, our challenges remain common (data collection, curation, sharing and interoperability; appropriate resources; governance; communication; supporting scientists).
2. While we all recognize the need for greater skills/resources at the level of data collection, curation and quality assurance, we also recognize the value of a certain level of centralization.
3. Open Access and Open Data offer new opportunities to promote our science and our scientists, and this needs to be better communicated and explained to a broad range of stakeholders. Our efforts should be directed towards supporting scientists; we must create tools, processes and a culture that support them.
4. Strong governance, monitoring/auditing and reporting will be critical; strength in these areas will also help to prioritize data management resourcing.

Next steps

Building upon existing efforts within and beyond CGIAR to:

1. Develop a second draft of the CGIAR Open Access and Data Management Implementation Guidelines by December 9; with input from across – and outside – CGIAR, finalize Guidelines v1.0 by February 2014, for approval by March by the Consortium CEO.
2. Secure funding to start implementation of Open Access and Data Management (OA & DM) in CGIAR (January 2014).
3. Establish Data Standards Taskforce to identify/ manage standards over time (January 2014).
4. Establish Open Access and Data Management Governing Council (January 2014).
5. Develop communications/advocacy campaign (March 2014).
6. Actively engage with/contribute to AGROVOC and build additional CGIAR language into the system.
7. Identify other opportunities for greater engagement beyond CGIAR, including CIARD, IATI, etc.
8. Define CRP proposal components for pilots by March 2014.
9. Identify monitoring and evaluation metrics, conduct baseline and define process to collect lessons learned.
10. Fund F2F Taskforce & Council meetings (March 2014).
11. Release pilot RFPs (April 2014)
12. Finalize where/how to store, share and access data – use of Dataverse, existing genomic, spatial and other data repositories to set up the destinations for our data pipeline.

Useful links

Link to all background docs and presentations

List of participants

Agenda

Some highlights from the presentations

Setting the stage. Piers Bocock, CGIAR Consortium Office

Piers Bocock, Director of Knowledge Management and Communication for the CGIAR Consortium, provided a short presentation to set the stage for the workshop. He reminded participants of the progress towards Open Access to date – from the establishment of the CGIAR Principles on Intellectual Assets in April 2012, to the identification of the need for an Open Access policy, the development of that policy, and the drafting of the initial Implementation Guidelines. The major challenge for the workshop was outlined as follows: now that all 15 CGIAR Research Centers have signed up for the Open Access and Data Management policy – making it mandatory – how will that policy actually be implemented?

Link to presentation

OA & DM Policy and Implementation Guidelines. Ed Crothall, CGIAR Consortium Office

One of the key long-term goals of the initiative is to have a portal such as 'open.cgiar.org', where all CGIAR related data is searchable and retrievable via one intuitive user portal. We have a clear mandate from CGIAR Members on OA&DM, with all 15 Research Centers having approved the OA&DM policy. The challenge now is to design the processes, tools and incentives to share data in a methodical and intelligent way.

The Guidelines document will include principles of how to manage data. It cannot be all things to all people, but will outline some standards, so that as an organization, CGIAR can speak with 'one data voice'. The plan is to adopt the guidelines by March 2014.

Crothall also highlighted the existence of a white paper Shifting the goalposts — from high impact journals to high impact data. This outlines many examples of data sharing already in progress in CGIAR.

Link to presentation

Data standards and linking current practices - Importance of standard vocabularies. Johannes Keizer, FAO

This presentation explored the role of AGROVOC and OpenAgris in enabling open data and applying standard data formats, to make data more easily readable.

AGROVOC is a controlled vocabulary covering various areas of interest including food, nutrition, agriculture, fisheries, forestry and the environment. Currently, AGROVOC contains more than 32,000 concepts organized in a hierarchy; each concept may have labels in up to 22 languages: Arabic, Chinese, Czech, English, French, German, Hindi, Hungarian, Italian, Japanese, Korean, Lao, Persian, Polish, Portuguese, Russian, Slovak, Spanish, Thai and Turkish. Four more language versions are under development (Malaysian, Moldavian, Telugu and Ukrainian).

Keizer highlighted some of the key features of AGROVOC. He stressed the importance of semantic alignment of data, rather than its physical centralization. Making data accessible through a common language helps to make data available. He urged the group to bring data out of the silos and make it fully interoperable.

For more information, visit: http://aims.fao.org/standards/agrovoc/

Bioversity, Crop Ontology, LOD. Richard Bruskiewich, Bioversity

Crop Ontology case studies were presented, highlighting the increasing need for interoperability of breeding data. Bruskiewich stressed there was an urgent need for a common terminology in breeding data. The vision of the crop ontology initiative is to overcome some of these issues and enable data mining.

For more information, visit: www.cropontology.org
Link to presentation

FAO approach to data exchange and dissemination. Josef Schmidhuber, FAO stats division

Schmidhuber explained that FAO deals with market data, socio-economic data and country level data. He outlined some of the challenges:

1. Lack of data and metadata description/harmonization;
2. Unclear data policies;
3. Uncoordinated data lifecycle management;
4. Poor governance, ineffective institutional frameworks;
5. Limited interoperability;
6. Limited user orientation;
7. Poor data dissemination systems, limited communication and user awareness.

The FAO approach is based on the idea that data and tools are a public good, with no licensing constraints, and full redistribution rights. FAO aims to move open data into accessible data. It is building a statistical governance system, external and internal. This means it is working towards supporting multi-disciplinary technical working groups, linking all data initiatives and collaborating between agencies.

Link to presentation

A unified crop data management system for CGIAR. Graham McLaren, GCP

McLaren highlighted the importance of governance for successful data management. Data can be described as being of sufficient quality if it is 'fit for purpose'. That purpose is rarely 'publication and sharing', and researchers do not need much metadata because of the ways they use their data. But to shape data into a form whereby it can be useful to others requires annotation with standards. There is general agreement that it is valuable to annotate crop information in this way, so as to make it consistent over time and across groups. But how can we get our plant breeders to invest in the necessary annotation? McLaren said he believes the only way is to build the annotation into the data collection process, as the information is readily available at the time of planning experiments. Once the project is over, this information may well have been forgotten and the process becomes more challenging. Tools to assist the process need to make collection easy, so that researchers will actually use them. Link to presentation

CCAFS Knowledge and Data Sharing. David Abreu, CIAT-CCAFS

The strategy consists of the following:

1. Establishing a process: data management strategy; open access policy; stakeholder agreements;
2. Facilitating the systems: platforms; data flow;
3. Enabling a data culture; there are support materials and processes; data management support pack;

4. Implementation (strategic/operational): data flow infrastructure; technical infrastructure.

A guiding principle is not to invent anything that someone else has done, and build on what is already there. But where required, a system should be built and put in place.

Link to presentation

Spatial Commons. Philip Thornton. ILRI-CCAFS

Standards are a subject of much discussion. Some aspects, such as file exchange format, are relatively easy to agree upon. But it is much harder to reach consensus on other areas. These include development platform, sharing platform, and can we go on without metadata? Lessons learned include recognition that we are all busy, and we all hate metadata/documentation! At the end of the project, people just want to move on, rather than go back and spend extra time on metadata. Can we appoint someone devoted to metadata to ensure that this gets done?

Link to presentation

Mapping agricultural investments and technologies. Melanie Bacou, HarvestChoice

Investment mapping is embedded into a number of overlapping CGIAR-led initiatives and multi-partner alliances, with potential duplication of efforts, such as: CGIAR reform; CAADP; ASTI; G8 New Alliance; Dublin Process; IATI.

Link to presentation

Amazon Web Services

The CGIAR Consortium has been talking to Amazon Web Services (AWS) and has been given storage space in the AWS public data sets. This represents a valuable opportunity for overcoming barriers by storing large amounts of well curated, ready-for-the-public CGIAR datasets, as well as promoting the use of cloud computing, so as to take advantage of data that CGIAR and partners store in the cloud. AWS support collaborative application sharing and fast resource deployment of analysis. http://aws.amazon.com/publicdatasets/

Data Streams, access portals and tools for integration and analysis. Budhendra Bhaduri, Oak Ridge National Lab

A case study from the world's most powerful open scientific computing facility.
https://www.bioenergykdf.net/

One of the challenges has been to get people to talk about what data they are going to release in six months' time, and not just about the data that is already available. There is a need to create a community of practice. Indentifying the people behind the data sets is important and was a goal from the beginning. Identifying impacts that can be measured to gauge success was difficult, but crucial.

Potential savings to the program from billion-ton study interface on KDF was approximately US$1,000,000. This was in time saved (see slide 15 on the KDF: Return on Investment).

Link to presentation

Bioversity standards. A common language. Adriana Alercia, Bioversity

A presentation was given on the work that Bioversity has done on GR descriptors and derived standards since its inception. All crop descriptors have been developed in consultation with other CGIAR Centers. The descriptors are the key to a long shelf-life.

Link to presentation

Transparency, accountability, joined-up data. Bill Anderson, International Aid Transparency Initiative

Currently, data on development spending evelopment spent data is not joined up. There is huge potential in linking development money to impact. Everyone in development should be linking up. IATI is a reporting channel, not an information system. The potential exists to provide a common electronic format for forward looking reporting of all activities by all participants in the delivery of development cooperation and humanitarian aid.

Link to presentation

Virtual Lab for Plant Breeding (VLPB). Rob Dirks

This was a private sector case study. As plant breeding developed, it became clear there would be a need for staff to support the bioinformatics. The ultimate wish is to be able to browse through a digital genebank and identify the accessions that contain desirable alleles. VLPB aim to develop tools that will be made available to both industry and academiawith both academic and industry partners.

Link to presentation

Open Agricultural Knowledge for Development; CIARD. Stephen Rudgard, FAO

The Coherence in Information for Agricultural Research for Development (CIARD) movement = open agricultural knowledge for development. It was founded by 15 partners (including CGIAR) and there is now a community of 400+ working to ensure that information becomes more accessible to those who need it.

CIARD resources include:

- checklist of good practices
- set of pathways - descriptions of how to achieve items in the checklist
- advocacy toolkit

For more information and resources, visit: http://www.ciard.net/

Link to presentation

CGIAR Consortium Science Team perspective. Philippe Ellul, CGIAR Consortium Office

This presentation discussed harmonized project management in CGIAR. There is a great deal of different language referring to different components throughout the system. The Consortium Office is responsible for reporting and monitoring, including: CRP annual reports and IA reporting. The Consortium Office Science Team is currently working with project teams on the second call for proposals. In terms of monitoring, it has become necessary to have some harmonization of terms used.

Link to presentation

ISPC perspective. Rachid Serraj, ISPC

There is a need for data metrics and an indicator framework, so as to be able to measure changes in agricultural productivity across scales and monitor the associated impact. A management tool is required to inform CGIAR and donors on progress. We now have progress in our planning, but how to track it? Questions remain about what, where, when and how to measure, and by whom? Consistency is required across the CRPs.

Ref: Strategic Study on Metrics, Benchmarking and Monitoring for CRP Evaluation and Impact Assessment. Ref: ISPC White Paper (June 2013) to look at SLOs to identify the routes through which agricultural research can reach them.

Link to presentation

Working groups

On Day Two, the participants were divided into four groups to concentrate on specific issues: (1) Genetic resources and germplasm data; (2) Socio-economic, geospatial, environmental data; (3) Harmonized research management across CGIAR; and (4) Incentives, capacity, culture, governance for OA&DM in CGIAR.

Outputs from working groups

The following domains were defined for which IT and semantic standards should be recommended or for which specific follow-up actions are required:

1. Genetic resources, crop breeding and genomics (collectively 'germplasm');
2. Socio-economics, geospatial and environmental;
3. Harmonized research and development;
4. Governance, incentives, capacity building, culture.

It was decided that a taskforce should be formed to manage/coordinate all standards on all the above domains. The taskforce should consist of CRP focal points and CGIAR Research Centers, plus experts for each domain. Further definition will be provided. The purpose of the taskforce was outlined as identifying 'best practices for managing and publishing data through open access'. It was decided that it should deal with the following subjects:

- Metadata
- Interoperability (use cases, standards for all domains)
- Technologies (data collection to deposit), which should include the identification of suitable datasets to be hosted on AWS Public Datasets
- Quality assurance
- Capacity building (training, tools)
- Collaboration

A second governance committee will need to be established to consider:

1. Establishing clear lines of accountability – we must show Center/CRP/scientist accountability;
2. Ensure workflows are in place to put context to data;
3. Establish OA&DM functions /skills /capacity/throughout CGIAR;
4. Decide how and where to invest financial resources and develop M&E indicators;
5. Define a process to minimize data loss ('fugitive loss');
6. Plan a clear communication awareness/capacity/campaign for scientists;
7. Work with the data standards group to oversee metadata documentation;
8. Agree a data quality auditing /certification process;
9. Plan for changes to CGIAR culture – making the connection between scientists and overall goals, mission of organization – collective responsibility/contribution;
10. Oversee implementation budgeting – appropriate allocation of resources to valuation data research collection storage processes.

Annex I: List of participants

Name	Organization	Specialist data domain(s)
Abreu, David	CIAT-CCAFS	Spatial
Alercia, Adriana	Bioversity	Genetic resources
Anderson, Bill Elliot	Development Initiatives	Socio-economic
Bacou, Melanie	HarvestChoice	Socio-economic, spatial
Bamba, Zoumana	IITA	
Barahona, Carlos	University of Reading	Socio-economic
Bhaduri, Budhendra	Oak Ridge National Lab	Spatial
Bocock, Piers	CGIAR Consortium Office	Spatial
Bruskiewich, Richard	Bioversity	Genetic resources, crops & breeding
Caracciolo, Carolina	FAO	
Chapman, Kay	CGIAR Consortium Office	
Chukka, Srinivasarao	ICRISAT	Socio-economic, spatial
Crothall, Edward	CGIAR Consortium Office	
Dieng, Ibnou	AfricaRice	Crops & breeding
Dileepkumar, Guntuku	ICRISAT	Genetic resources, crops & breeding, genomics, spatial, socio-economic
Dirks, Rob	Virtual Lab for Plant Breeding	Crops & breeding, genomics
Ellul, Philippe	CGIAR Consortium Office	Genetic resources, crops & breeding
Erlita, Sufiet	CIFOR	Socio-economic
Fotsy, Michelle	CGIAR Consortium Office	
Gardiner, Peter	ISPC	
Garrucio, Maria	Bioversity	
Genari, Pietro	FAO	Socio-economic
Keizer, Johannes	FAO	
Marus, Michael	CGIAR Consortium Office	
Matteis, Luca	Bioversity	Genetic resources, crops & breeding, spatial
McLaren, Graham	GCP	Genetic resources, crops & breeding
McNally, Kenneth	IRRI	Genetic resources, genomics
Obreza, Matija	Global Crop Trust	Genomics
Parr, Martin	CABI	
Porcari, Enrica	CGIAR Consortium Office	
Quiros, Carlos	ILRI	Socio-economic, spatial
Rudgard, Stephen	FAO	
Schmidhuber, Josef	FAO	
Serraj, Rachid	ISPC	
Subirats, Imma	FAO	
Thornton, Philip	ILRI-CCAFS	Socio-economic
Van Den Berg, Marco	IRRI	