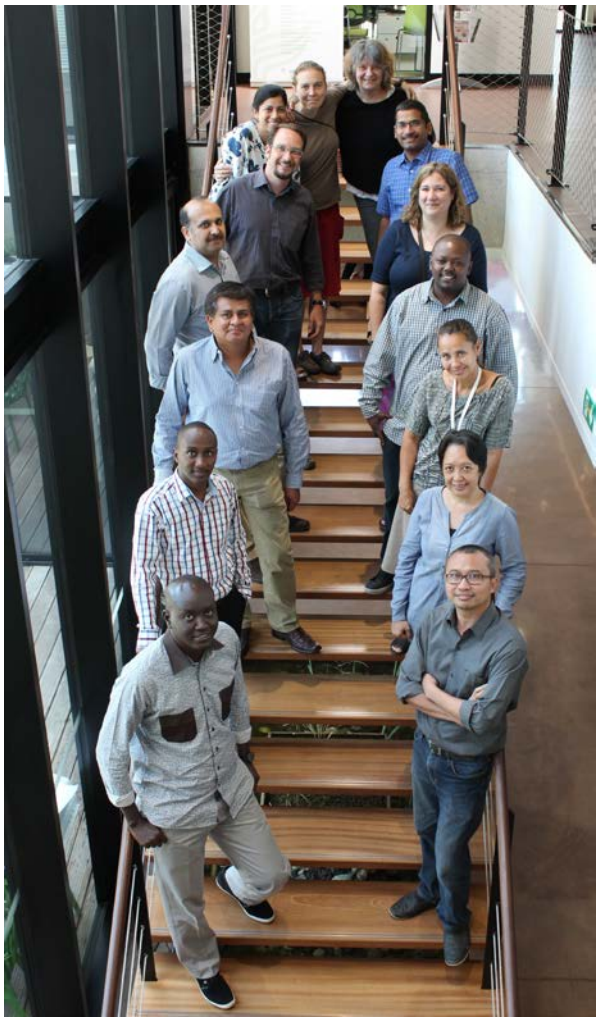

CGIAR Data Management Task Force (DMTF) Annual Meeting Report

June 17, 2016; Montpellier, France

Author(s): CGIAR System Office & DMTF members

CGIAR Data Management Task Force (DMTF)
June 17, 2016 Annual Meeting Report

Contents

Agenda

Friday, June 17, 2016	
9:00 - 9:30	**Introductions, overview, goals**
9:45 – 11:00	**Brainstorming** • In a perfect world, where do you see CGIAR and your Center in the next 3-5 years in terms of open data, and being able to leverage big data capabilities? • What are your top 3-5 challenges in achieving this? • What are some solutions to overcome those challenges – actions from Center, CO, others?
11:00 – 11:15	**Tea/coffee**
11:15 – 12:30	**Center discussion: OA/OD & policy compliance; data capture & management tools (10 min max)** – **not** *formal, more learning/sharing opps; no need for polished slides* • Current status towards full compliance in 2018 (including CG core implementation/mapping etc.) • What's working (including cross-Center, cross-CRP initiatives) • What's not working so well/areas you could use help • QUICK demos of data capture/curation/quality tools and approaches
12:30 – 13:30	**Lunch**
13:30 – 15:00	**Center discussion: OA/OD & policy compliance; data capture & management tools (10 min max)** – **not** *formal, more learning/sharing opps; no need for polished slides* • Current status towards full compliance in 2018 (including CG core implementation/mapping etc.) • What's working (including cross-Center, cross-CRP initiatives) • What's not working so well/areas you could use help • QUICK demos of data capture/curation/quality tools and approaches
15:00 – 15:30	**Tea/coffee**
15:30 – 16:00	**Amazon Web Services** (Eddie Romanzo; Jawoo Koo)
16:00 – 16:30	**CRP mapping project** (Jawoo Koo)
16:30 – 17:15	**Legacy data and data prioritization discussion** (Medha Devare)
17:30 – 18:00	**Next steps; meeting closing**

Participants

	Institution	Name
1	AfricaRice	Ibnou Dieng
2	Bioversity International	Elizabeth Arnaud
3	CCAFS	David Abreu
4	CIAT	Leroy Mwanzia
5	CIFOR	Sufiet Erlita
6	CIFOR	Usman Muchlish
7	CIMMYT	Kate Dreher
8	CIP	Henry Juarez
9	ICRAF/DS CRP	Paul Baraka
10	ICRISAT	Srinivasarao Chukka
11	IITA	Martin Mueller
12	ILRI	Jane Poole
13	IWMI	Salman Siddiqui
14	ICARDA/DS CRP	Enrico Bonaiuti
15	CGIAR System Office	Medha Devare
16	CGIAR System Office	Michelle Fotsy

Session 1: Brainstorming

Group work focusing on key questions in three areas: 1) aspirations, 2) challenges, and 3) solutions. Post-its with individual points (bullets in tables below) were clustered by the group into broad themes, which form the headings of columns in each table.

1) Aspirations

Where do you see CGIAR and your Center in the next 3-5 years in terms of open data, and also in terms of being able to leverage big data capabilities? What can we realistically accomplish in 3 years? The caveat is that the OADM policy goes into full effect in 2018; the intent of the policy is to try and move us in the same direction, recognizing that not everybody will be in the same place at the same time.

	Usability /access	Interoperability	Discoverability	Culture/Incentives
	• Great quality of reporting from project leaders • Shortened data lifecycle (curation/cleaning/storage, etc.) • High quality public data, tagged with appropriate caveats or qualifiers • Big data: providing broader analytical capabilities • Ontologies that can be easily updated by the community (with moderation) and fed back into systems that use them • GPS tagged data collected in fields around the world (by the farmers, collaborators, etc.) -- which can be fed back into CG data systems and used to make	• Platform launched and used by centers and CRPs because consensus is reached to use similar standards and tools/toolboxes • Public datasets are tagged with ontology terms based on manual or automated annotation • Seamless interaction between the three platforms • Compliance with linked open data • Climate data collected at micro & macro level that can be easily passed into breeding decision systems, modeling tools, etc.	• Central workplace • Effective data harvesting system between center repositories • One stop-shop access to data across CRPs/Centers • All publicly available CG datasets can be found through a CG & other aggregating search portals based on metadata • A platform capable of harvesting OD from all centers (metadata only) • Data are captured in internal systems & pushed automatically into public warehouses	• Having OA/OD culture first implemented in the teams →projects →programs → center → CGIAR → CG stakeholders • Centers giving clear guidance and applying robust ethical rules for sharing confidential data + clear link to IP & research ethics committees • Scientists being recognized for publishing data & tools as much as they are for other information products (like papers) • Data management & analyses will be cost effective

	Usability /access	Interoperability	Discoverability	Culture/Incentives
	key breeding, targeting, forecasting decisions, etc. • Data collection using mobile technology becoming the norm (including direct beneficiary input) • Data stored cheaply and securely and connected to cheap high power computing facilities • Data can be accessed and understood by people in all languages	• Availability of data as web services or linked open data • All platforms interoperable at data level • Data is interoperable, harvestable (machine readable) according to I.O. standards • A strong community for documenting different types of datasets (controlled vocabularies, ontologies) • Data can pass easily between a suite of common systems & tools used by the CG and pushed into a wide array of analytical pipelines • Big data leveraged to provide data integration possibilities • Substantial progress in semantics: Agrovoc; ontologies; subject terms; automatic tagging; profiling people and objects; linking (research) data models with data distribution /management • All (most with exceptions) data is open within policy timeframe, documented with metadata at all levels	• Sharing services (Amazon, DOIs, etc.) and harmonization (minimum standards) • Data playground to leverage big data • Ability to ask questions to the big data platform and find the right datasets to answer those	

2) Challenges

What are your top 3-5 challenges in achieving this? What are some stumbling blocks?

	Infrastructure, standards, Data Quality & Prioritization	Cost/Resources	Culture and Capacity	Impact
	• Prioritizing legacy data [for "open"] • Improving quality of reporting • Common standards -- as tools, people, donors change, and 'international' standards also change (e.g. country naming!) • Lack of ontologies in common • Data generated within center and by partners are not standardized and are of variable quality • Even within a center, lack of communication leads to some data siloing & lack of interoperability • Different priorities challenge collaboration across centers (sometimes) • Some key primary databases are not ready – and the whole system of sharing data with the public will draw on these	• Convincing donors to support OD/OA • Cost of making data open: legacy data; resources for curation • Resource mobilization for sharing services • Time/people vs money • Not adequate staff time: training scientists; testing systems; migrating data; curating data; writing policies; scoping new tools, etc. • Resource cost sharing by CG for common activity; e.g. data manager salary can be shared by center (70%) as well as CG (30%) [i.e., need for sharing some key services]	• Lack of expertise • Big data/interoperability • Capacity and time from our researchers devoted to data management • Frequently changing/ advancing tech domain • Capacity building & sustainability • Practices during the project (scientist behavior) • Motivation of researchers to share data	• Tracking OD usage & impact →not just downloads (e.g. citations, new tool using data) • Tracking impact of the data & the platform • Measuring the impact of sharing data •

• Core data management systems in centers must have the data push feature and be compliant to CG metadata (e.g; phenotype data from BMS) • Big Data computing infrastructure required • Inflexibility of hosted open data platforms, platforms; e;g; hosted DataVerse does not support CGCore. No SLA agreements with hosted OD providers e.g. Harvard.			

3) Solutions

What are some solutions to overcome those challenges – including possible actions from Center, SO, others?

	Infrastructure, standards, Data Quality & Prioritization	Cost/Resources	Culture and Capacity	Shared Services

	• Publish only metadata of the legacy, non-curated, data → discoverable [then assess which data sets to make open based on usage/hits]
	• Legacy data:
	  o Priority setting protocols, incl. how far we "go back" -- should be led by key priority research question
	  o Have metadata on the status of the data
	  o Accessible upon request
	  o Crowd-source cleaning of metadata for legacy data
	• Enabling feedback by data user (fitness & quality)
		• Centers write proposals for data curation and infrastructure
		• SO to help mobilize resources for data curation and infrastructure
		• More time + resources to:
		  o Change culture (budgeting for DM, cognition of data as product)
		  o Provide incentives
		  o Collaborate
		  o Provide support (tech + non-tech)
			• Quality of reporting as part of project evaluation
			• SO to take lead in building capacity of centers to try to bring them at same level
			• SO to take lead on capacity building to target technical advances
			• Capacity building workshops on OA/OD, data management good practices
			• Training for data managers on platforms and tools
			• Motivation: adopting incentives based data sharing policy, e.g. performance evaluation
				• Enabling centers have control of OD platforms, e.g.:
				  o CGIAR support financially center hosted platforms, or
				  o CG hosted solutions: CGKAN & CGverse
				• CG-wide cloud computing facility supporting processing and storage reducing transaction time & cost
				• CG-wide (shared services) access to data tracking / usage; e.g. WOS (data citation index) to track the impact
				• Pool/shared consultancy available with different expertise areas

Session 2: Center/CRP updates

- CIMMYT: Hoping to comply with 2018 data release deadline; working on data prioritization to figure out which are the highest priority ones to make available; OA/OD impl plan being worked on.
- CCAFS: Resource-based management system that tracks all outputs from planning to reporting and beyond to monitor OA compliance; this generates a set of qualifiers; though there may be some flaws in reporting accuracy and quality, it is more like a subsite to interact with the community to get those numbers more accurate and go forward; OA/OD implementation plan has been revised a couple of times, and is now being done together with CIAT; comments being finalized before submitting to SO.
- ICARDA: Impl. plan submitted and revised based on SO comments; plan passed on to Executive Committee and needs to be endorsed by Board of Trustees. Being the last center/CRP to have a metadata schema, we adopted CG core. Pulled out of CGspace but agreed that if CG core is in place and a governance structure for CGspace and Dataverse, ICARDA would like to participate in some decisions and again share some of the budget. Having issues due to different installations of DataVerse in programs; planning to have only one installation for all. The recovery of lost data gave back 10000 datasets to register and upload; progress at approximately 100 items/month, and keep an inventory; given the lack of resources, it will likely take a couple of years.
- IITA: Efforts have been focused on infrastructure; in July CKAN will be in place. IITA is restructuring and the outcome is an opportunity to give more power to OA/OD steering; e.g. data processing, data managers, statisticians are put together and closer to ICT. IITA uses SharePoint as a collaboration platform, with project sites, repositories and will be putting workflows in place during the second half of the year. OA/OD impl plan was submitted in August 2015 and will be reviewed and updated.
- ILRI: Impl. plan was in final draft by August last year but still needs to be submitted. CKAN is up-to-date with CGCore; CGspace has had some technical issues. ILRI set up an institute committee (headed by Peter Ballantyne) established around compliance and monitoring of OA. Data portal has about 60 datasets; project datasets are in different places, and ILRI has plans to manage those.
- CIP: Impl. plan was submitted in September 2015; CIP is OA-compliant; has a data management plan working. CIP has a dedicated OA site on cgiar.org with CIP's OA & DM Policy, guidelines and procedures, as well as training modules; support is good at the DG level (e.g., yesterday their DDG research sent a memo to researchers to make open datasets a requirement). Dataverse newly implemented – not many records yet, but expect to have made much progress by 2017 end.
- IFPRI: Impl plan still in development; hoping to have ready by summer's end. IFPRI Dataverse has about 400 datasets; also have sub-Dataverses for CRPs. IFPRI is developing 3 ontologies: agriculture and nutrition, technology and value chain.

- CIFOR: Impl plan submitted June 2015. Have a Dataverse data system called 'PMC', and a Dspace publication system called 'My Cifor'; have been implementing CG Core in DSpace.
- ICRAF: Impl plan still under discussion and being written; currently with management. Have been focusing on defining the workflows. Project start up and project closure meetings are in place to ensure that all the project data is captured. HR clearance is also being implemented, with links to Dataverse for all datasets as a requirement. An ICRAF performance dashboard is also being developed as a data reporting management tool which captures all the different tools at ICRAF. Library catalog has changed over 3 years; now implementing KOHA system (same as IRRI); as Dataverse is using DOIs linking both systems is possible. Landscape portal stores spatial data. CG core has not yet been implemented; crosswalk done, and will see what other centers have done.
- Bioversity: Impl plan submitted to the SO. 2015 external audit conducted on research management and publishing processes, looking at workflows, bottlenecks, how scientists enter data, etc.; this enabled the creation of a resource data management plan; first version of guidelines being discussed with scientists. Bioversity going through portfolio restructuring, and internal reorganization with changes that will affect these processes. Champion users from different units and projects identified; datasets underlying published papers being uploaded to Dataverse and Dspace; champions' feedback on Dataverse is being compiled. Bioversity more or less in compliance with CG Core. Using the ESRI tool but lack a solid strategy for geospatial data. Re: ontologies, Bioversity works with breeders (mainly CRP breeding clusters) on the Crop Ontology and is working on an agronomy ontology; both ontologies have similar concepts, so can link them; have started collaboration with SDG ontology.
- A group discussion was engaged re: risks of Harvard Dataverse; in case of crash, even though there are back-ups, there is a risk of not having data readily available; additionally, there is no legal agreement between a Center and Harvard Dataverse (as opposed to service-level agreements with other providers like Amazon).
- IWMI: Impl plan done and submitted; now under review by management. Modified IWMI OA/OD policy based on CGIAR OADM Policy; have data management workflow in place. CG Core implemented; besides adopting CG Core, IWMI also follows ISO standards (e.g., Water Data Portal includes ISO and CG Core standards). A few OA seminars were held and plan to do more of those.
- ICRISAT: Currently making a shorter version of OA data management policy. Landscape analysis on current and legacy data is being done; scientists, project and team leads have been asked to give their top five datasets, as well as datasets linked to past research publications; these will be uploaded to data portal. Currently trying to convert metadata to CG Core standard. ICRISAT has not adopted a data implementation plan within projects yet, but this is in discussion with the leadership team. ICRISAT has had recent structural changes, and the data management unit is now with biometrics and bioinformatics enabling checking of products before making them public. Discussions with scientists have been initiated to try to understand more about datasets associated with published data and interlink the OA and OD repositories.

- CIAT: A common impl plan for CIAT and CCAFS submitted to SO; a final draft will be submitted to their team. Harvard Dataverse is data platform. Still working on CGspace and waiting to see what ILRI will be doing to conform to CG Core. There is now one team for data information and knowledge; library and data managers are working together to support OA/OD; also creating a CoP for data managers within projects to support their work; core team wants to provide support through a data quality management system; reaching out to institutions like Reading to see how this can be done; CIAT is open to suggestions and any other center interested in this initiative is welcome to join. Datasets related to all the publications are being requested. Working with the project management team to follow up on the project status; for flexibility reasons versions of the management plan is adapted to different project stakeholders; final product is uploaded to Dataverse as open. A new Strategic IT Committee is being created which will also look at OA implementation.
- AfricaRice: Impl plan was submitted to SO. Because many people are involved in managing information products (KM, Comms, DM, legal, editorial and publication committee), need to discuss among them as well as with DDG on activities related to OA/OD; periodically a memo is sent as reminder that all information products generated by AfricaRice staff is property of AfricaRice. Data management plan has not yet been established. A wiki is in place, with an external and an internal part for project management. Publication repository was developed internally, and discussion is ongoing whether to move to DSpace. For data, moving from Dataverse to CKAN; close to 500 datasets already in CKAN; have both AfricaRice metadata and the CG Core in place.

Session 3: Amazon Web Services

Amazon 17June2016 CGIAR Overview.pdf Presentation by Eddie Romanzo, Nonprofit Account Manager & Dustin Sell, Senior Solution Architect

Session 4: CRP mapping project

Discussion led by Jawoo Koo, around a global system like the WorldBank and USAID have, very transparently showing where people work, on what project, and what they generate. For CGIAR, there is no such global system yet; there were/are several efforts, e.g. ILRI had the regional mapping, CIMMYT has Maize Atlas & Wheat Atlas, the Gender Network has the gender map, etc. It would be good to have a better idea of those efforts and be able to search across them from a central place. There is the IATI (International Aid Transparency Initiative) with CGIAR as official member; some time ago, IFPRI was asked to develop a system to map all CRP projects. Now there is funding for this through PIM. The objective would be to have a dynamic map with indicators, of all CRPs' activities.
For centers to support the project, the needs, the audience must be defined. Jawoo Koo will draft something and share with the group.

Session 5: Next Steps & Meeting Closing

1) Proposed institutional capacity building (face to face, online tutorials, webinars etc)

- Tools / specific platforms – e.g. DV, CKAN, etc. (e.g. Open Knowledge Foundation)
- Data management (best practices etc.)
- Technologies:
  - LOD & ontologies
  - New tech / agility – Big Data
  - Interoperability protocols at data level
- Development of a consultancy pool to help with these

2) Priorities / to-do

	Task	Who takes lead
1	Harvesting across repositories	Medha +CIAT, IFPRI, ICARDA
2	DV consulting – talk to CoP	Kate + Leroy + Medha
3	CGverse exploration	Kate + Leroy + Medha
4	Data-level interoperability; CG core mapping	David + Medha (with consultant)
5	DMP tool for CGIAR	Leroy + Jane (+ Medha in loop)
6	Impact of open data o Data citation index o (Altmetrics)?	Chukka (+ Medha in loop)
7	Ontology working group	Elizabeth + Kate + Martin + Enrico + Chukka + Henry + Leroy + Sufiet + Medha
8	DMTF meetings: o Virtual: 1 x 3 months (Sept, Dec, Mar) o Face-to-face: 1 x 12 months –@ 3 days coupled with capacity building/professional development event (target: RDA 9th plenary meeting, April 5-7, 2017 in Barcelona)	Michelle + Medha
9	Webinars	Rodrigo, Medha