Open Access (OA) Workshop

Qaribu Inn Nairobi, Kenya
September 28th – 29th, 2018

Correct citation: Okuku. H, Juarez. H, Hidalgo. G and Wanjohi. L 2018. F. Njung'e (Editor). Proceedings of the Open Access (OA) Workshop, held 28-29 September 2018, Nairobi, Kenya, p 14, International Potato Center, Nairobi.

Table of Contents_Toc527553487

Preamble

The Open Access (OA) workshop was held in Nairobi, Kenya on September 28th – 29th, 2018. The workshop was organized and implemented through the leadership of the Research Informatics Unit (RIU), Knowledge Resources Center (KRC) and the Sweetpotato Action for Security and Health in Africa (SASHA). The workshop's main target group were key research staff across all CIP-SSA programs.

The primary goal of the workshop was to assist participants understand the key practices and procedures for OA and Research Data Management. The intended outcome was to have curated datasets from the last five years ready to be published in Dataverse. The two-day workshop was practical oriented meeting with the morning sessions giving theoretical framework while afternoon sessions exclusively dedicated to assisting the participants work on their data to make the OA compliant.

Participant list.

The meeting attracted a total of 25 participants from mainly Sub-Saharan Africa (SSA) projects. A full list of participants is found in Appendix I. The attendees were a mixed group with project leaders and monitoring and evaluation officers being the dominant groups.

1. Welcome and Opening remarks

Jan Low

The participants were warmly welcomed to the Nairobi meeting and reminded that this was the 2nd meeting in OA, the last meeting was held in 2016. The participants were also reminded of importance of OA as it is now a CIP and donor requirement that all data collected and / or published need to be made OA within 12 months after data collection. CIP, as one of the implementing CGIAR centers was given a 2-year grace period to fully implement OA, which ended in September 2018. Being compliant with OA is a good image to the organization.

2. Overview of Progress in OA at CIP

Henry Juarez

CIP has been implementing OA for over 3 years now.

- 257 Journal Articles (67.3% OA) [2015-2018]
- 190 Datasets [2015-2018]

CIP launched the Second Open Access Competition [March 2018].

- The goal is to have around 100 datasets made OA at the end of 2018.
- The competition is focusing mainly on data underpinning publications.

With the grace period over, participants were reminded of the deposits timelines of various publications and data:

- CIP-published journals, books, reports, etc. – Immediately
- Peer-reviewed versions of articles – Ideally at the time of publication and / or latest 6 months after publication.
- Book / book chapters – Within 6 months after publication
- Data and datasets – Within 12 months of completion of data collection.

Some of the success stories so far:

- CIP reached the goal for the first Data Sprint in 2017 with 125 datasets with good metadata description and vocabularies published.
- OA/OD much more accepted now at CIP.
- Data Sprint considered as a good practice to incentive OA and its being replicated in other centers.

So, what are the challenges and opportunities so far:

- There is still a cultural change for scientists and researchers to share their data publicly. There's a need on more training to understand and implement OA especially in the regions.
- CGSpace and Dataverse repositories are constantly being updated, more publications are now linked to their datasets.

- Managing datasets and their quality require more time and resources to put datasets into CIP's standards. Scientists take longer to deliver complete datasets.
- Human Resources will include OA in Global Onboarding Project in the induction programs.
- Other infrastructures like MEL, Dataverse and CGSpace are now inter-connected and so reduce time spent in making data OA.

2.1 Reactions, questions, and concerns from the presentation

- Many breeders felt that it takes more cycles of data collection to completely collect breeding data where different data are collected at different locations and time. For effective use and eventual publication, these datasets need to be put together and cannot be used individually. So, they required clarification on the timelines for making data OA, particularly that condition that requires data to be made OA after 12 months of data collection. The same situation was for projects that collect baseline, midline and endline data.
- **Does CIP have funds to cater for OA publication fees for projects that have ended and did not have budget lines on the same?**
    - Currently No, but those willing to publish need to inquire the possibilities of getting funding via Knowledge Resources Center.
- There is need for the Dataverse to collect more information of the persons downloading and using the data as secondary users. Information like name, contact and associated institution is necessary.
- Since OA is an important policy change in CIP, there is need to integrate it as part of the induction for the new tools and platforms e.g. MEL, Big Data Platform, etc.
- Data from project partners: important to consider partner's policies on data sharing
- GDBX Digital Global, Henry will share with participants the type of information available in GDBX Digital Global and GIS services that RIU can provide
- **When Data repositories were presented, participants asked about Social Sciences qualitative data; how to handle confidentiality vs anonymization?**
  Even if we delete personal information, qualitative questionnaires responses could be easily guessed by users, depending on the knowledge of the region

3. OA & Journal Articles

Gabriela Hidalgo

Since the transition period is over, strict adherence to OA is now required for all publications. The office of the Knowledge Resources Center (KRC) need to be consulted well in advance before publication is made OA. Progress of OA in terms of publications for the period 2015 – 2018 are:

- 2015: A total of 70 publications were made OA; 32 had limited access; 38 Open Access.
- 2016: 56 publications: 24 limited access; 32 open access
- 2017 91 publications: 19 limited access; 72 open access
- 2018: 40 publications: 10 limited access; 30 open access

The participants were taken through the legal issues particularly surrounding on copyright and license.

- All CIP produced and published articles, documentations, books, etc. belong to CIP and are freely downloaded.
- Publications in the context of projects must consider the funder's branding guidelines
- Materials published externally depend on the Publisher's copyright
- Care needs to be taken when sharing publications on social media particularly Research Gate so as not to violate copyright of the publishers.
- Researchers need to take note of "predatory Journals" who endanger the quality of scientific publications and compromise the reputation of authors and editors of legitimate journals
- The participants were encourage to register with ORCID ; a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors. In addition to identification purposes, ORCID Integrates with different online tools and profiles and Offers more visibility for researchers

3.1 Reactions, questions, and concerns from the presentation

- Is there a preferred list of journals with their grading in terms of impact factors?
  - There is a list that will be shared with those interested

- The issues of copyright and ownership arose when it comes to partnerships with students and other partners
- What about issues surrounding presentations and photos during international meetings and conferences; does CIP have a policy governing the sharing and making OA? Are there repositories for such?
    - Not currently but this will be reviewed
- Sharing of publications on a one-to-one basis is perfectly okay and allowed even if the publication is copyrighted. Sharing via a public channel like social media is what is prohibited.
- Open Access fees continues to be a big challenge, because of different circumstances: time period for approval from journals, projects end before an article can be finished, fees can be expensive depending on the journals, etc.

4. Data management and organizing files

Luka Wanjohi

There is need to adapt proper file naming systems and conventions to make it easier to navigate and get documents and files easily in the personal computers. Some of the common issues that are always associated with dis-organized and poorly designed filing systems:

- User cannot easily find information again, and spend time looking for it. 80% of users fail the "show me" test
- File and folder names inconsistent and lack clear structure
- Different types of data (private, public, shared, non-shared) is mixed, and difficult to share data
- Different versions are mixed up, and difficult to find final/last version
- Data is insufficiently documented, and other people cannot understand it (and after a while neither can the original researcher)

Some of the best practices recommended include but not limited to:

- Define a system and stick to the defined system. Avoid mixing up systems on the same computer.

- Immediately save new file in proper place in structure (not in email, desktop, my documents)
- For each separately funded project, one may want to keep everything related to that two folders (1 professional, 1 personal/contractual)
- Use the knowledge structure of your professional work on a high level as possible, and try to define non-overlapping stable categories
- Avoid if possible to use organizations and people as organizing principles as the tend to change e.g. CIP is restructuring, but research topics stay the same
- Use key words that will help in a search of the file or document e.g. subject area, geographical region, organization, type of document (Concept note, Budget, Agenda etc)
- Don't name a file by the recipient or sender.
- If certain file types that are used repeatedly consider a formal naming convention e.g. Series-Year-Season-Experiment (OFUG10A05)
- For non-standard files use longer names that indicates content and key words
- Indicate version and date for docs with revisions

4.1 Reactions, questions, and concerns from the presentation

- When copying files and documents from different media, care needs to be taken to avoid the "255 character" rule in Windows. For example, a file may not open in copied in long nested folder structures.
- When scanning a document and sending the scanned document, participants were encouraged to rename the file accordingly instead of keeping the default names generated by the scanning machines
- The traditional issues of inserting user initials at the end of reviewed document seems common but care need to be taken not to have a long file name. The use of cloud computing to edit and track changes to documents was suggested.
- There is need to avoid using repeated names in file names that already exist if the folder structure. For example, if a file is under "SASHA" folder, there is no need to include SASHA in file names under the folder.
- Many users particularly in SSA were disappointed with use of One-Drive for backup purposes

- Poor training and support particularly for users outside ILRI campus.
- Recovery of files not synchronized and a pain.

5. Documentation: Metadata and data dictionaries

Henry Juarez

As a way of introducing the practical session, the participants were taken through the documentations required to make data and datasets OA

- Introduced to what is data and datasets.
- What is metadata
- Data dictionary and code books

CIP has adopted the CG Core Metadata and were shared with participants via e-mail. All participants were taken over individuals entries in the metadata and data dictionary templates

5.1 Reactions, questions, and concerns from the presentation

- Can financial data be uploaded for Open Access.
  - Yes, by basically removing any identifying information.
  - Discuss with partners of the implication and get them to understand the reason for making it OA.
- There was a suggestion to include research methods in the metadata in addition to the already item of sampling procedure

6. Practical session

Participants listed at least 1 dataset to practically start working with to make it ready for OA. Then they completed the metadata related to the above data

7. Data Management Plan (DMP)

Henry Juarez

The Data Management Plan is the starting point in the Data Life Cycle. However, the plan should be revisited often throughout the project to ensure proper data documentation and management.

a. It outlines what one will do with data during and after you complete of any research

b. In addition to DMP, participants were introduced to common repositories that CIP uses together with their features
    i. BioMart: Institutional Warehouse for Research Data.
    ii. Dataverse: an open source web application to share, preserve, cite, explore and analyze research data. Data authors, and affiliated institutions receive academic credit and web visibility.
    iii. CGSpace: a joint repository if CG centers to archive, curate, disseminate and permanently preserve research outputs and information products

7.1 Reactions, questions, and concerns from the presentation

- **Can DMP be applied retrospectively for projects without it?**
  No, it's just for new projects
- **Is it possible to edit data after publications into Dataverse?**
  Yes, but it will be a new version. The old data will still be there but with a different version
- **How long should a dataset or replication data be on "draft mode"?**
  Not specific but the RUI can request for further information to publish the data
- **Is the DOI (Digital Object Identifier) automatically generated by Dataverse**
  Yes.
- **What happens when a dataset that was initially published as "Dataset for" eventually gets published and so "replication for"?**
  A new entry will be made in Dataverse and will include new data with journal description
- **Does the CGSpace handle videos?**
  No
- **There was a discussion on the quality of the data that we are uploading: is there any review process to finally approve datasets? Where can we establish the main responsibility: the authors, the PIs? Who's reputation is more at risk: CIP's? researchers'?**
- Time to clean data was also an issue; researchers find hard to find enough time to make sure all datasets are cleaned and checked (rather than only the portion they need to publish)

8. Example of documentation of Social Science Data with STATA

Haile Selassie Okuku

OA & Open Data (OD) policies make more and more social science data available for secondary analysis. In secondary data analysis, documentation plays a critical role in transferring knowledge about data from data producers to secondary users. Documentation (aka metadata) usually includes codebooks & data dictionaries, related bibliographies and data collection instruments. Metadata serves 3 main purposes: resource discovery; preservation; & administration. Documentation for social science data is mainly used for resource discovery (searching and judging the relevancy of the data) and secondary analysis. Stata documentation is best done using "a do-file": text editor that saves commands and comments.

- Participants taken through best recommended practices in variable naming
- Labeling of the data, variables and values were taught.
- Use of Stata wildcards to make documentation short were
- General structure of Stata do-file were discussed that included header information, setting environment and the main parts of the do-file

9. Example on documentation of Breeding data with HIDAP and SweetPotatoBase

Luka Wanjohi

Workflow for data management of breeding information were shared

10. Hands-on: Participants continued preparing and documenting individual datasets to upload into Dataverse

- The participants continued to finalize the datasets and related documentation
- The finalized datasets were collected and saved in USB drive
- Henry Juarez will be in charge of uploading the datasets into dataverse
- Haile Okuku will make a follow-up with participants on other datasets that need to be made OA

11. Wrap up Sessions

Jan Low

The training was a big success with at least 15 datasets prepared and will be uploaded into dataverse. It was reiterated the need for better planning of data collection from designing protocols and data collection tools, to the planning of collection and the cleaning process. There is need to target on publishing of every data collected. Participants were encouraged to include this in their talent management in collaboration with the supervisors.

Appendix I: Participant List

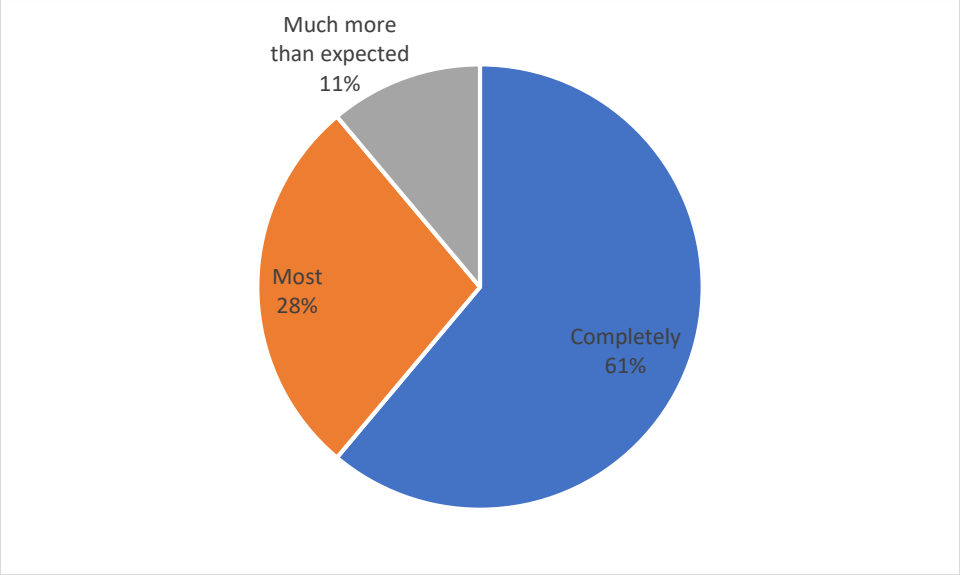
NO.	First Name	Last Name	Gender	Title	Country	Mobile number	Email
1	Abdul	Naico	M	Project Manager	Mozambique	+ 258 829849584	A.Naico@cgiar.org; naico@yahoo.com.fr
2	Bernice	Wairimu	F	Program Specialist	Kenya	+254 723697795	B.Wairimu@cgiar.org
3	Birhanu	Biazin	M	Value Chain Specialist	Ethiopia	+251 916829891	B.Temesgen@cgiar.org
4	Carol	Kamau	F	Data Intern	Kenya		kamaucarol.cwk@gmail.com
5	Eliya	Kapalasa	M	Market chain development officer	Malawi		e.kapalasa@cgiar.org
6	Eric	Magembe	M	Associate scientist	Kenya	254 722379645	e.magembe@cgiar.org
7	Frezer	Asfaw	M	Data processing assistant	Ethiopia	+251 921 402424	F.Asfaw@cgiar.org
8	Frederick	Grant	M	Project manager	Tanzania	+255 759184824	F.Grant@cgiar.org
9	Gerald	Kyalo	M	Senior Research Associate	Uganda		gerald.kyalo@cgiar.org
10	Godwill	Makunde	M	Sweetpotato specialist	Mozambique	+258 825135177	G.Makunde@cgiar.org
11	Haile	Okuku	M	Consultant	Kenya		H.Okuku@cgiar.org
12	Hilda	Munyua	F	Project Manager-BNFB	Kenya	+254 720 297464	H.Munyua@cgiar.org
13	Joyce	Maru	F	Capacity Development and Communication Specialist	Kenya	+254 707 627645	J.Maru@cgiar.org
14	Julius	Okello	M	Impact Assessment	Uganda	+256 756024761	J.Okello@cgiar.org
15	Kwame	Ogero	M	Regional Research Associate	Tanzania	+255 689 457461	K.Ogero@cgiar.org
16	Luka	Wanjohi	M	Data manager	Kenya	+254 722 302 271	L.Wanjohi@cgiar.org
17	Mihiretu	Cherinet	M	Research Associate	Ethiopia	+251935923781	M.Cherinet@cgiar.org
18	Reuben	Ssali	M	Plant Breeder Associate- Post Doc	Uganda		r.ssali@cgiar.org
19	Rose	Chesoli		Research Assistant	Kenya	254 729791934	r.chesoli@cgiar.org
20	Srini	Rajendran	M	Agricultural Economist	Kenya	+254-739 104 556	srini.rajendran@cgiar.org
21	Temesgen	Bocher	M	M&E	Mozambique	+258 846375065	T.Bocher@cgiar.org
22	Thomas Alexander	Van Mourik	M	Project Manager	Ghana	233 265 347339	T.VanMourik@cgiar.org
23	Valentine	Uwase	F	Monitoring and Evaluation Assistant	Rwanda		V.uwase@cgiar.org
24	Henry	Juarez	M		Peru		h.juarez@cgiar.org
25	Gabriela	Hidalgo	F		Peru		g.hidalgo@cgiar.org

Appendix II: Post Workshop Response

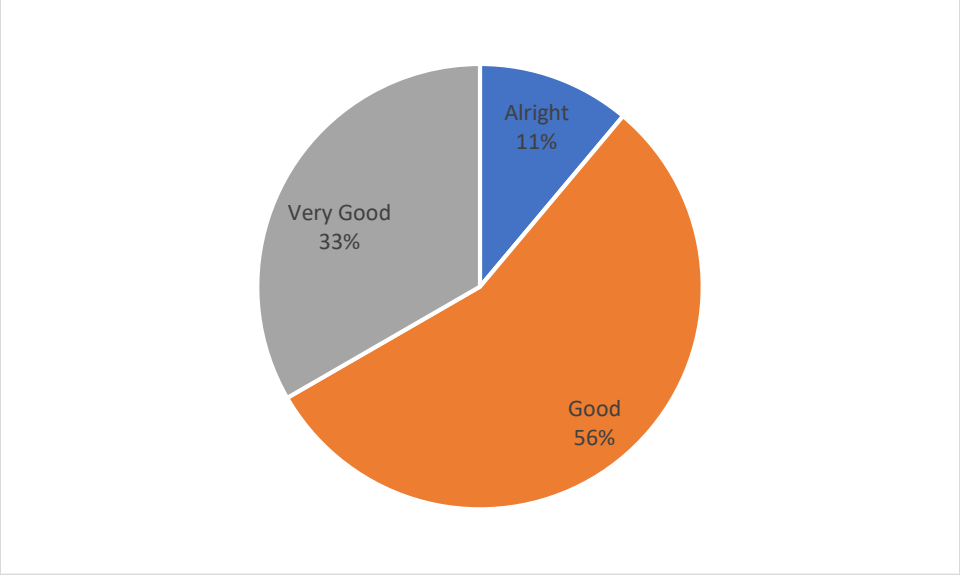
A total of 18 (out of 25 participants) took part in the post workshop survey. Seventy-eight percent of the participants were male (14/18) with their mean age of 37 years. More than half of the participants were directly involved in M&E in their respective projects.

Below are the responses got in relation to various levels of workshop participation:

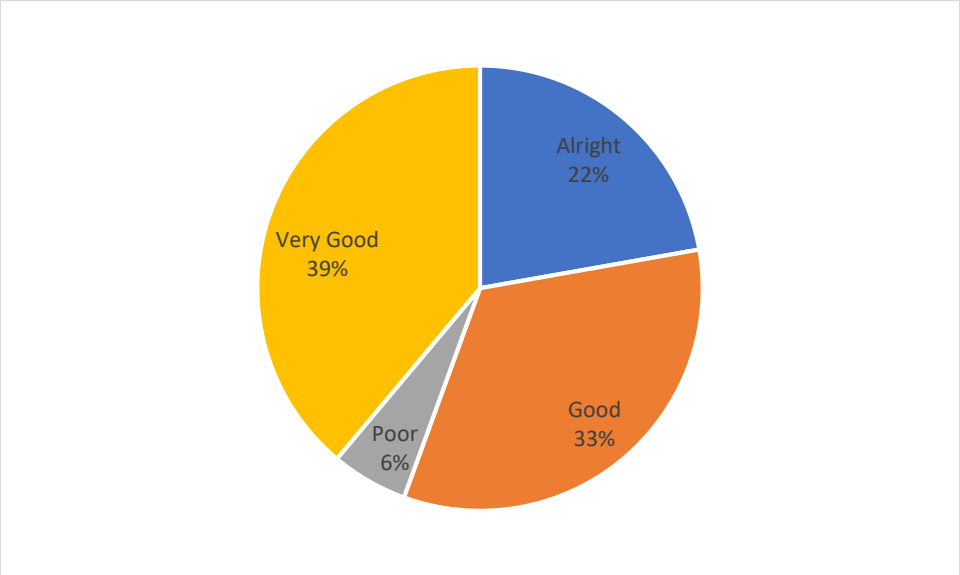
1. Did the 2-day meeting match your expectations?



2. How would you rate the quality of the presentations in terms of content?



3. How would you rate the meeting in terms of organization (logistics, communication?



4. Topics that the participants found least useful include:
    a. OCID dataset
    b. Publishing data
    c. Documentation using Stata
    d. Data management and organizing files
    e. Documentation on breeding data
    f. OA competition
    g. Open data access publishing time period
    h. Poorly defined sweetpotato entology
    i. Using published data sets.
    j. Overview of open access at CIP

5. Some of the suggestions from the group included but not limited to
    a. Good to keep people working on the data and publish them.
    b. Critical discussions on getting clean data published especially in social sciences, where protocols are not standardized.
    c. The meeting has been an eye opener as such we need to have another similar one to focusing on specific areas like social science or just breeding data management
    d. Follow up to ensure we publish from the data and make the rest open access.
    e. Might need to know more on CIP policy and procedure for editing process for the journal papers before submitting to the journal.
    f. We should have more researchers attend the training in the future. The session on publication workflows was extremely helpful and a lot of researchers could benefit from such.