WHITE PAPER

Consortium

CGIAR

Shifting the goalposts—from high impact journals to high impact data

Anja Gassner[1], Luz Marina Alvare[2], Zoumana Bamba[3], Douglas Beare[4], Marichu Bernardo[5], Chandrashekhar Biradar[6], Martin van Brakel[7], Robert Chapman[8], Guntuku Dileepkumar[9], Ibnou Dieng[10], Sufiet Erlita[11], Richard Fulss[12], Jane Poole[13], Mrigesh Kshatriya[11], Guvener Selim[14] Reinhard Simon[14], Kai Sonder[12], Nilam Prasai[2], Maria Garruccio[8], Simone Staiger Rivas[14], Maya Rajasekharan[14] , Chukka Srinivasa Rao[9]

[1] World Agroforestry Centre (ICRAF), United Nations Avenue, Gigiri, PO Box 30677 Nairobi 00100 Kenya
[2] International Food Policy Research Institute (IFPRI), 2033 K St, NW, Washington, DC 20006-1002 USA
[3] International Institute of Tropical Agriculture (IITA), HQ-PMB 5320, Ibadan, Oyo State Nigeria
[4] WorldFish, Jalan Batu Maung, Batu Maung, 11960 Bayan Lepas, Penang, Malaysia, PO Box 500 GPO, 10670 Penang, Malaysia
[5] International Rice Research Institute (IRRI), DAPO Box 7777 Metro Manila 1301, Los Baños, Philippines
[6] International Center for Agricultural Research in the Dry Areas (ICARDA, "Dalia Building 2nd Floor, Bashir El Kassar Street, Verdun, Beirut, Lebanon 1108-2010 P.O. Box 114/5055 Beirut, Lebanon"
[7] International Water Management Institute (IWMI), P. O. Box 2075, Colombo, Sri Lanka, 127, Sunil Mawatha, Pelawatte, Battaramulla, Sri Lanka
[8] Bioversity International, "HQ- Via dei Tre Denari 472/a 00057 Maccarese (Fiumicino) Rome, Italy"
[9] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324 Andhra Pradesh, India
[10] Africa Rice Center (AfricaRice) 01 B.P. 2031, Cotonou, Benin
[11] Center for International Forestry Research (CIFOR), HQ- Jalan CIFOR, Situ Gede
Bogor (Barat) 16115, Indonesia, Mailing-P.O. Box 0113 BOCBD Bogor 16000, Indonesia"
[12] International Maize and Wheat Improvement Center (CIMMYT), Km. 45, Carretera, México-Veracruz, El Batán, Texcoco CP 56130, Edo. de México, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico
[13] International Livestock Research Institute (ILRI), HQ-PO Box 30709, Nairobi, 00100 , Old Naivasha Road, , Nairobi, Kenya
[14] International Potato Center (CIP), Avenida La Molina 1895, La Molina, Apartado Postal 1558, Lima, Peru
[15]International Center for Tropical Agriculture (CIAT), Km 17, Recta Cali-Palmira
Apartado Aéreo 6713, Cali, Colombia

Contents

Introduction

CGIAR is a global agriculture research partnership. Its science is carried out by the 15 research centers who are members of the CGIAR Consortium in collaboration with hundreds of partner organizations. Each member of the CGIAR Consortium has the mandate to contribute to the eradication of hunger and poverty at the global level by advancing research in development. Being funded mainly through public funds there is an obligation to extract maximum public good value from the research data the individual centers are collecting. With the exception of a few designated projects, in the past, data itself was merely seen as a means to an end, synthesized to produce selected knowledge products such as publications, technical manuals or policy briefs.

Changing financial realities and the need to reach our aid targets with less available resources is leading to a shift in thinking for new ways to make scientific research, and research data, more available, reusable and reproducible. Sufficiently preserved and replicable data are more absolute than contemporaneously drawn conclusions and, if they are collected to address one scientific question, can later be applied for the solution of entirely different problems (Altman et al., 2006).

A growing number of donors now require data sharing requirements in their standard grant contracts and open access to be applied to the research data of the work they fund. DFID and the European Commission are among those who are now adopting this policy. As one of the first custodians of development related data, the World Bank, launched an open data initiative in 2012. CGIAR itself followed this example and approved in March 2012 a policy that clarifies CGIAR research results, including data, are openly available and accessible by default – and that any deviation has to be justified. Open data is not a new concept for CGIAR, but in the past has largely remained linked to specific projects or core programs that are designed to collect and share data widely.

In keeping with the changes to the amount, management and analyses of data, the role of scientists working in development, such as those at the CGIAR Centers, is changing. Previously, grant awards, and performance evaluations focused strongly on publication in high-impact journals. Fortunately, there is a rediscovery that the real currency of research and scientific knowledge –the data methods and ideas- are also important instruments, in their own right, to accelerate the impact. Our scientists are increasingly using their skills and their strategic advantage of being well integrated in farming communities to create well designed datasets for others to use. While there is no doubt that progress on development related research questions would accelerate with increased access to well documented and readily accessible data, the move also puts more pressure on the gatherers, compilers, analysts and keepers of the data. Because of previous emphasis on scientific publications only, research projects usually do not budget for the extra costs that occur for publishing and long-term storage of data. While projects adequately account for the costs of conducting research, including the collection and analysis of research data, they seldom include preparing of data and metadata and curation as part of the costs of the research process led alone long-term costs for data storage and preservation beyond the immediate lifetime of a project. Even more importantly the research nature of CGIAR makes it mandatory to ensure that the contributions of those individuals and organizations are recognized, whose reputations and careers depend on that recognition. Data generators have little opportunities to gain recognition from publishing high value datasets under a performance evaluation system that strongly emphases scientific publications.

The purpose of this white paper is to provide an overview of the ongoing initiatives at center level to respond to changing public expectations and to the challenge of improving the conduct of science by making research data widely available. We also attempt to provide a framework for implementing open access for research data to maximize CGIAR's impact on development. The remainder of this paper proceeds as follows; firstly a summary of the diversity of research data produced by the centers is given, followed by an overview of the existing infrastructure for data management for each Center. Secondly, some of the limitations and barriers faced by the centers in their process to mainstream research data publishing are addressed. The paper concludes with recommendations for how these limitations and barriers can be tackled.

Types of research data

In close collaboration with national research institutions almost 10,000 scientists, researchers, technicians, are collecting, analyzing and synthesizing data on smallholder agricultural systems in Asia, Africa and Latin America. While the overall mandate of the CGIAR Centers is the same, each Center has their own research emphasis resulting in a vast variety of different kinds of research data. The following section does not attempt to give an exhaustive description of the data we produce, but rather to provide an overview of the main types and their characteristics.

Long term trials

Centers that are focusing on breeding of food crops and livestock need multiple location trials over many years before a new genotype can be fully evaluated, so do agroforestry and forestry trials that are working with slow growing perennials. Centers working more generally on combined systems research, natural resource management and climate change also have research covering decades in order to characterize and influence changes in these systems. These kinds of data are often collected in various consecutive projects and technical and scientific staff that collect and analyze the data can change. From a scientific point of view it is not desirable or useful to make the data available before meaningful intermediate results have been collected.

One-off data collections

Some data are project specific and aimed at answering specific research and/or development related questions, often analyzed and published as part of the project deliverables. Publication should be written within a reasonable timeframe after which the data can be publically released together with the publication. What is a reasonable timeframe does not only vary from researcher to researcher, but by discipline. Economics publishing moves much slower than other fields. It is not uncommon for a paper to reach publication 5-10 years after the original data was collected.

Baseline data

Baseline data, either household surveys or biophysical surveys used either for basic characterization of a new project site, or for an impact evaluation of the project. The data are collected as part of the deliverables of the project, but budget constraints seldom allow sufficient sample size or rigid sampling designs to allow the use of the data for peer- reviewed publications. These data usually get analyzed and results presented in donor reports. For these kinds of data the curation costs are actually very high as they are not accompanied by a scientific publication that provide the necessary metadata and methodologies are stored in various versions on private computers. The data has limited usability for the project itself, but baseline data sets combined from multiple projects potentially form high information assets for the institutes and the public. There is a tradeoff between releasing the data to do comparative analysis or to wait until the follow up study is completed.

Genomic data

Some Centers are involved in genome sequencing projects, specifically in generating the NGS (next generation sequencing) data for gene-phenotype association studies. These studies can generate multi-terabases of sequencing data. One of the key challenges is to devise scalable and robust data management and data sharing solutions. High-performance computing and storage are required to efficiently process data generated by NGS. Bioinformatics support is integral to address data management systems dealing with efficient storage, retrieval, data mining, data analysis and making data available to the public at the appropriate time.

Data collected as part of a research thesis

Various projects have a specific capacity building requirement, whereby postgraduate students collect a substantial part of the data. The data is to be used in peer-reviewed publications as partial requirement for their degrees. Data can only be publically released after the student has published their papers or their thesis.

Value- added secondary datasets

A large proportion of the work of a CGIAR scientist is to review and analyze pre-existing data that was not gathered or collected by the authors of the current research project. Usually it has been collected by another organization or source or data collected from government publications. Typical secondary datasets that are used are meteorological datasets, remotely sensed data often in the form of satellite images or aerial photographs, panel data sets for rural households. The secondary datasets are shared with CGIAR scientists under specific user agreements or licenses and cannot be publically shared by the scientists. Data products derived from these data can be shared however, without the raw data.

Spatial data

While most data generated in the CGIAR research activities has a spatial component (with exception of pure lab analysis not related to specific locations) the GIS units of the individual centers collect, transform and generate a large amount of spatial data. These fall into three categories:

a. Spatial data (in form of polygons and raster, satellite imagery) obtained from other parties like NASA, JRC, National Geographical Institutes, Universities and other research institutions, NGOs, private companies etc. This is used for analysis, mapping, targeting within the work of the units. Often this original data is curated and improved and represents a new international public good and falls under **value-added secondary datasets**.
b. Geo referenced data generated by projects and the GIS units within the centers. This can be any other form of data collected such as socio economic surveys, soil samples, germplasm collections etc. It is then often combined with other spatial data sets as mentioned under a) and utilized for further analysis.
c. Statistical and other related data (climate from stations) that is collected from national and other institutions such as subnational crop production data or poverty statistics. This is georeferenced, often extrapolated and if necessary further disaggregated and converted into common GIS data formats and made available to the public and other centers again as **value-added secondary datasets**.

Data collected in a private public partnership project

When working with private companies some of the data and information is highly sensitive and usually the centers sign confidentiality agreements that state explicitly what the data can be used for and which data products can be made publically available.

R & D Datasets

CGIAR has programs that are designed and funded purposely to create public databases on key agricultural indicators. The databases are dynamic and are updated on regular intervals. Technical and scientific staff that collect and analyze the data can change. Here only aggregated data are made available together with the methods.

Partner data

Most of our projects are done together with partners that sometimes contribute their own non-CGIAR funded data. Here it is important that the partner decides what the data should be used for and who should have access to the data. Often the partner has only agreed to joint copyright of data products, but not the actual data sets.

Research data infrastructure across the centers

While publishing research data as a research output is nothing new for the CGIAR Centers, in the past it was confined to flagship programs and projects. Data sharing agreements and data management policies were developed and data managers and curators hired based on the individual needs of these projects. Centers did not have the infrastructure, capacity and incentive to make more of their research data widely available. In early 2008 a first attempt was made by the Alliance Deputy Executive (ADE) to review data management across the centers and explore mechanisms for strengthening collective action among centers (Anon, 2008). A lot has changed since then and Center wide research data management policies are becoming the norm rather than the exception. Increasing investment in research support units and or staff and the implementation of data archives reflect the change of mind set within center management as well scientists. Most centers have already or are in the process of setting-up 'OAI-compliant'[1] data repositories, which allow the easy harvest of metadata from one repository to another. Table 1 provides a brief summary of the current research data infrastructure across centers. Details for each Center are given below.

Table 1: Overview of existing infrastructure for research data management & bioinformatics across the different institutes.

Centre	Research data Management Policy	Data Management Unit	Geoinformatics Unit	Biometrics Unit	Centralized Data Archiving & sharing
Africa Rice	YES	YES	YES	YES	Since July 2012
Bioversity	In process	YES			Since Sept. 2013
CIAT	YES	YES	YES		In process
CIFOR	YES				In process
CIMMYT	YES	Recruiting	YES	YES	In process
CIP	YES	YES	YES	YES	YES
ICARDA	In process	YES	YES	YES	In process
ICRAF	YES	YES	YES	YES	Since 2011
ICRISAT	In process	YES	YES	YES	YES
IFPRI	YES	Recruiting	No		Since 2005
IITA	In process	In process	YES	YES	In process
ILRI	YES	YES	YES	YES	Partial (shared servers, data portal in development)
IRRI	YES (Currently being updated)	YES	YES	YES	In process
IWMI	YES		YES		YES
World Fish	YES	YES	YES		YES

[1] The Open Archives Initiative (OAI) develops and promotes interoperability standards (Protocol for Metadata Harvesting) that aim to facilitate the efficient dissemination of content.

Africa Rice

AfricaRice research support unit, the Data Integration and Biometrics Unit (DIB Unit) is responsible for data management and to assist AfricaRice research staff (and partners) to enhance the efficiency and efficacy of data management processes: i) data acquisition, quality control and storage: backstopping of AfricaRice research staff in the management of their experimental data, including automating data gathering procedures and quality control; ii) data integration, analysis and visualization: biometrics' advice and support to AfricaRice research staff on experimental designs, GxE analyses, mapping quantitative trait loci (QTLs), analyzing genetic diversity, development of decision support tools, etc. In addition, the DIB unit provides institute-wide data standards, including a common vocabulary and the use of standardized formats for primary and metadata. It enables AfricaRice and partners to more easily integrate, synchronize and consolidate data from different programs, exchange data with other organizations in a common format, and communicate effectively through shared terms and reporting formats.

The Unit consists of two statisticians (one IRS and one Support Staff), one data manager (Support Staff) and one Consultant (Support Staff). Africa Rice is using Dataverse as its research data archive. It was released on July 2012 and consists of 23 studies as of today. AfricaRice developed a Data Management and Sharing Policy in 1999. Revised in 2013, the document is in the process for approbation by the AfricaRice Board of Trustees.

Bioversity

Bioversity has been engaged in developing the CGIAR Open Access policy and guidelines together with other CGIAR Centers. Once approved, these documents will form the basis for Bioversity's 'Open Access and Data Management Plan'. We have already drafted this plan based on the current version of the OA policy, and we hope to have the policy approved in early 2014.

The responsibility for the collection and management of research data is with the Research Planning and Monitoring Unit at Bioversity. The Library staff within this unit are responsible for 1) collecting the datasets; 2) ensuring completeness and accuracy of the associated metadata; 3) finding/linking the datasets with associated publications and tools, and 4) inputting/releasing it on the Dataverse network. No extra personnel have been recruited for these tasks so far. Whilst Bioversity does not have a GIS unit, there are 3 GIS specialists based in regional offices and HQ.

Bioversity recently established an open access dataset repository on the Dataverse (Harvard) Platform. At present we have datasets available there from research carried out in 2012, and we are currently collecting other datasets from previous years. At present, the metadata of each dataset is available immediately but the actual datasets are only released once the scientist gives their approval. In order to collect datasets systematically Bioversity has in place focal points in its 5 research programs that assist the Library staff primarily with collecting the datasets from their scientific staff. In collaboration with partners Bioversity also manages a number of specialized databases and datasets relating to agricultural and forest biodiversity such as:

- Collecting Mission database (http://bioversity.github.io/geosite/)
- Musa Germplasm Information System (MGIS) (http://www.crop-diversity.org/banana/)
- EURISCO (http://eurisco.ecpgr.org/nc/home_page.html)
- New World Fruit Database (http://nwfdb.bioversityinternational.org/)

Bioversity, always in collaboration with its partners, also develops data standards for documentation and protocols to enable information sharing.

The Descriptor Lists publications that are published by Bioversity, assist researchers and genebank curators to improve their capacity to describe, store, manage and share information about plant resources, whether stored in genebanks or growing in their natural environments (e.g. FAO/Bioversity International, 2012).

Center for International Forestry Research (CIFOR)

CIFOR's Director General approved and promulgated a *Research Data Management Policy* and *Guidelines and Procedures* in July 2013. These identify project managers as having primary responsibility for management, including archiving, of research data; require preparation and implementation of Research Data Management Plans; and identify the Center's Data and Information Services Unit as having responsibility for ensuring storage of and access to archived data, with support from the GIS Unit for spatial data. CIFOR is currently recruiting a Data Librarian to play a primary role in supporting project managers and staff implementing the *Policy*.

International Potato Center (CIP)

Around September 2002, CIPs director of research implemented a service unit charged with building and maintaining the institutional 'memory' of research data (todays rough equivalent of KM on scientific data). To this end it provides both centralized archiving, documentation and databasing services as well as support to researchers in data analysis and visualization. This includes also the de-novo production of databases and tools as appropriate versus re-using existing software. The latter aspect was seen as intrinsic to the concept of a 'memory' since no memory 'makes sense' without 'data' and 'data processing' tools. Since the beginning, the unit has been involved in developing community data documentation standards as well as in the application of open-source software principles and the promotion of open access. The unit was called at the time 'research informatics unit (RIU)' and recently renamed to 'Integrated IT and Computational Research. The unit provides services in documentation, database design and development, PC-based and mobile application development, GIS and molecular bioinformatics. The unit has been part of former CGIAR KM-related activities like: the IPD (the Intergenebank Potato Database, a database to cross-reference potato collections and share passport and evaluation data, since ~1995), the SINGER community as a repository of long-term data on germplasm (since end of 1990ies; now moving towards Genesys-2 software from the Global Trust), the GenerationCP on breeding materials including the 'composite genotype sets', and the ICT-KM initiative.

International Center for Agricultural Research in the Dry Areas (ICARDA)

ICRADA has a Geoinformatics Unit for all the geospatial database and related products in coordination with Biometrics, GRS, CODIS etc. At present data is stored in several individual archives and the institute is in the process of setting up centralized data archiving facility for collecting, streamlining, archiving and sharing, supported by a research data management policy. Less than 10% of research data is currently accessible through the public domain.

International Center for Tropical Agriculture (CIAT)

CIAT strongly believes that open access to data produced by Staff and partners strategically supports our mission and ensures transparency and equity in exploitation of the opportunities created. Even before implementing a Data Management policy in February 2012, CIAT led sharing of geo-spatial data together with associated information, at different scale in an organized, standard and consistent way. With the implementation of the Data management policy, CIAT is in the process of putting 40 years of data from the institutional memory (includes phenotypic, genotypic, climatic, spatial and socio-economic data) in the public domain. The work involves compilation of data from the institutional memory, cleaning and re-formatting, respecting current data ontologies and standards, and publication through appropriate mechanisms.

While CIAT encourage the data sharing culture, data management function is currently spread across research areas and different units. The information technology unit is in charge of data storage. Starting 2014, we expect to consolidate some of the ongoing activities under Library and information management. In addition, CIAT also hosts the CCAFS data manager.

Most of the field data are now collected using a hand held PDA, which facilitates easy backup and consolidation. We are beginning to use a CIAT Dataverse, to facilitate data sharing associated with research publications. Effort is ongoing to share trial data from commodity programs via Agtrials. Given the very large size of bioinformatics and genomic data sets, further discussion is ongoing to see how/ and if we share terra bites of raw data and mechanisms for doing that in terms of server capacity etc. Quite a lot of data is already made available through GCP https://www.integratedbreeding.net/.

International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)

Recognizing the value of research data aggregation, analysis and availability, ICRISAT established a Data Management Unit in August 2010 to mainstream, support and manage research data for its preservation and publication across the institute and to the extent possible, to make the data publicly available for all users to use. The unit consists of a full time Senior Data Manager supported by consultants, research fellows and Biometric unit of ICRISAT. Biometric unit at ICRISAT work on data quality activities with a full time biometrician along with five statisticians and research scholars.

The DMU of ICRISAT have come up with various data management platforms for better pedigree management, breeding practice analysis, survey management, climate prediction activities, etc. ICRISAT has also introduced several innovative platforms to its scientists that includes open as well as commercial such as Agrobase, aWhere (cloud based data management tool), VDSA (data warehouse), IBP etc. and also developed several applications that helps better visualization &

sharing of research data with GIS analysis capabilities. The DMU also defined and developed workflows and protocols for managing research data that being produced by various research programs. Training programs and workshops have been organized to the scientists and partners working on ICRISAT's research programs on use of new tools and data management platforms. ICRISAT is using Dataverse as a platform to publish datasets. Nearly about 400 datasets in 5 formats in 42 study areas have been uploaded in to ICRISAT-Dataverse application. ICRISAT has adopted Open Access Policy and launched repository in May 2011 to provide an easy interface for researchers, practitioners, or web-connected farmers to use, build on and share research conducted at ICRISAT.

DMU is working with a strategy to manage the research data that being produced across the institute, covering all the locations and thematic/program areas; and organize it efficiently in a "central data repository" for allowing Global scientific community "access to the data to get successful results" for addressing pressing Global issues. Efforts have also been made to bring a cultural change at scientist level by continuous interactions and by educating them on the advantages of the data sharing. This is a win-win situation for the institutes as well as the scientific community. For this, ICRISAT has come up with Data Management Policy and has been put forward for the research committee approvals.

International Food Policy Research Institute (IFPRI)

IFPRI's Communication and Knowledge Management Division, has a Knowledge Management Unit considered a support unit for research. The KM unit provides support for preparing the data documentation, data curation, creates the connection with associated publications and tools, monitors it uses and also serves as a user support and Q&A center. At present the unit has 1 full time data curator (with research background), a half time knowledge manager, that helps with the taxonomies and metadata and citation statistics.

The Communications and Knowledge Management Division is pursuing to hire a Data Manager, who will provide support during the research cycle, development of data plans when the project is conceived, conceptual frameworks, models, questionnaires and analysis.

IFPRI has a Dataset Policy existing since 2000, was updated in 2010, making it mandatory for researchers to provide open access to the relevant data, while safeguarding the privacy of participants and protecting confidential and proprietary information. IFPRI will make all primary and value-added secondary datasets collected after January 1, 1999, publicly available two (2) years after all data collection ceases or, before two years, at the time of a major publication by the lead data collector. The datasets are released after the lead researcher and Division Director has approved it.

In 2012, 11 datasets were prepared for posting in the open repository at IFPRI Dataverse, and now we have a total of 97 datasets. These datasets have been cited 120 times (according to ISI data citation index) and have been downloaded 19,289 times. Whenever possible we linked the publication to dataset and vice-versa so that the discoverability of dataset and related publications becomes easier. Preparing a dataset for the open repository, takes between 2.5-6 days depending on how the raw data is provided to the unit by the research leader.

International Institute of Tropical Agriculture (IITA)

An in-house survey in early 2013 revealed that most research data is stored on individual scientists and units computers in various different software and statistical packages, including Excel, SPSS, Stata, Access, and SQL. Data on research projects/outputs is collected at an ad hoc basis (through questionnaires, emails, letters, interviews, group discussions, reviews of official reports). The data is produced and used in a variety of formats, including digital, print and physical. Most of research data are still very fragmented. IITA Biometrics Unit conducts regular training and backstopping on data collection and analysis and supports genome sequencing projects, specifically in generating the NGS (next generation sequencing) data for gene-phenotype association studies. IITA's GIS unit manages a geospatial database and related products. The survey also revealed that researchers were highly supportive of having a data and information framework as they felt that it would reduce duplication of data collection and survey fatigue among NARS and other partners.

IITA has taken steps to improve data management. One of the first successes is a relational database on Cassava, which provides cassava breeders and researchers access to data and tools in a centralized, user-friendly and reliable database. IITA is looking at developing other crop data databases using the Cassavabase as a model. The formulation of the functional requirements for this comprehensive Crop Breeding Data Management Platform based on breeders' needs is underway. This platform would include the management of crop information and the development of applications to facilitate breeding processes and agronomic field trials data, soil science data, plant health information (pathology and entomology) and nutrition/post-harvest characteristics.

IITA is also envisioning implementing an E-Research infrastructure that will support research and that enables researchers to undertake excellent research and deliver innovation outcomes, provide the means to manipulate, manage, share, integrate and reuse research data, and enables research teams to share resources and work together more effectively.

International Maize and Wheat Improvement Center (CIMMYT)

CIMMYT is dealing with a broad variety of data ranging from maize and wheat germplasm, crop data at the field, farm, community, country, to regional and global level, socioeconomic data. Data are handled by several units within or across the programs, institution and partners. CIMMYT has been a key developer of germplasm related data management platforms including Fieldbook, IMIS/IWIS, and the IBP. Data volumes have increased rapidly and with that the need to develop tools that allow the manipulation and use of such data (eg genomics applications; decision support tools for sustainable intensification approaches). In 2012 a new data management policy that came into effect that is to accelerate data interchange internally and externally. Six data coordinators were hired to assist in: a) establishing data standards, documentation and data curation processes, b) coordinate receipt, storage, manipulation and quality control of field and germplasm related, c) participate in the design, development and population of versatile institutional databases/repositories, interfaces and output tools, d) introduce new informatics tools to staff and collaborators, provide on-the-job training, and report back user requirements to CIMMYT software engineers, e) manage and document the dissemination of data (raw and analyzed).

There are currently three data coordinators in the Genetic Resources Program (GRP) who work for the Global Maize Program (GMP) and the Global Wheat Program (GWP) handling both molecular data and breeder's trial and nursery data (phenotypic). The breeder's nursery and trial data mainly consists of phenotypic data collected in the field during the crop cycle and can be complemented by quality and nutritional traits (such as nutritional value, forage traits, baking quality or other food processing relevant traits) or molecular information at variable density (few markers) to several genotype-by-sequencing information across significant parts of the genome. The socio economics program (SEP) and the global conservation agriculture program (GCAP) share three data coordinators who work in the regions (Africa, Asia, Latin America as well as globally) on both socio economic data and agronomic trial data, their standardizations and platforms.

The GIS unit collects spatial and meteorological data on global, national and sub national scales as well as statistical data related to maize and wheat production in all countries producing these two commodities which is then converted to spatial data and as geo referenced data collected and generated in specific projects.

CIMMYT knowledge management is co-chairing the Wheat Data Interoperability working group which aims to coordinate worldwide research efforts in the fields of wheat genetics, genomics, physiology, breeding and agronomy. This part of the Research Data Alliance initiative.

International Rice Research Institute (IRRI)

IRRI has various research support groups with different roles and responsibilities. The Research Data Management (RDM) group conducts regular training on good practices in managing research data to ensure that data are well documented (metadata) and organized systematically (file repositories). Areas covered include: Research data planning; Research data collection, authentication and storage; Data backup and security; and Data archival and sharing. RDM also assist research staff in implementing these good practices based on RDM policy and IP Policy. The Biometrics and Breeding Informatics group conducts regular training on experimental designs and statistical data analysis as well as the use of plant breeding tools developed in-house. The Bioinformatics group conducts training on Bioinformatics and supports research groups by providing tools for SNP analysis.

All instruments and lab equipment are ISO certified and in-house verification performed regularly, to ensure accuracy and traceability of research data collected. Additional flagship databases that are maintained by research units are ICIS-IRIS for germplasm data, World Rice Statistics and Household survey database for Social Science unit, Long Term Continuous Cropping Experiments, and Climate data for Climate Unit. IRRI also maintains the Rice Knowledge bank, which is the world's leading repository of extension and training materials related to rice production.

International Livestock Research Institute (ILRI)

The ILRI Research Methods Group consists of: 2 statisticians, 5 systems designers / managers, 2 systems administrators (high performance computing systems) and 2 GIS data managers / analysts. In addition we have a large number of GIS analysts sat in research programs and a few database managers and data systems type people. There is also a bioinformatics group which sits in the Directorate of Biosciences (BECA & Animal Biosciences). Management of data is mainly at the project level although for large projects shared SQL servers are used. With the advent of the CGIAR Research Programs (CRPs) ILRI is currently establishing a data portal / platform for both Institute and the CRP on Livestock and Fish data publishing. Initially the data portal will provide cataloguing and access to data and in later development a module for monitoring and evaluation will allow for meta-analysis across projects, for key indicators (e.g. IDO's). The open-source platform (CKAN) will also allow inter-operability and communication with ILRI's knowledge management system (DSpace – Mahider) and other platforms. A revised data management policy aligned to the CGIAR Open-access policy, the CGIAR Management of Intellectual Assets (IA) policy and various CRP policies in development is currently being finalized with detailed a detailed implementation plan providing ILRI staff and partners will options for collecting, managing and sharing their data.

International Water Management Institute (IWMI)

IWMI's Information and Knowledge Group (IKG) provides knowledge management support, connecting IWMI's knowledge outputs through the integrated Institution-wide search facility, "Poodle" that allows for a comprehensive search across peer reviewed and non-peer reviewed IWMI research outputs. Amongst many other responsibilities IKG provides publication support and manages IWMI's internal publications such as the IWMI working paper and IWMI research report series. IWMI currently has no separate Data Management Unit, but its Research data Management Policy stipulates that all data must be archived in IWMI's central repository (WDP) with standard metadata as early as possible after collection and processing with appropriate access rights. Project leaders are held responsible for ensuring implementation of data management policy at project level while IWMI theme leaders ensure that all projects under the themes are complying with the policy. IWMI's Implementation Framework for research data management follows that developed and used by ICRAF (2008) (IWMI 2011).

IWMI has a dedicated Geo-informatics Unit (GRandD) which is responsible for developing data standards, protocol and training to researchers and research assistants to use these standards. The unit provides research support for database design, data organization, data manipulation, metadata preparation and data archiving. The GRandD unit coordinates with projects for uploading data into the central repository, making them available to appropriate users through its Water Data Portal, an integrated portal for consulting and accessing IWMI research data.

World Agroforestry Centre (ICRAF)

ICRAF has a Research Methods Group as part of their global support units. The group provides services all along the research cycle from development of conceptual frameworks consisting of problem definition, hypothesis, models and research questions, well-documented research designs and methods, data management and curation and statistical analysis. At present the group consists of three statisticians and three data base managers at Headquarter. The group is further supported through 3 regional data managers, with further recruitments in the regions line up for 2014.

In 2012 about 65% of the work was related to data management and 35% to research design and data analysis.

Since November 2012 ICRAF has a research data management policy that makes it mandatory for scientists to provide open access to all relevant primary data that are accompanying their scientific publication. Projects are responsible to ensure that research data is described by appropriated Metadata throughout their lifecycle and are required to have all their datasets submitted to the repository upon closure. Metadata are move to the open domain as soon as they have been compiled, the actual data sets are only released once the scientists give their approval.

ICRAF is using Dataverse as its data repository, which was released on October 2011 2011. As of September 2013 we have 96 studies in our Dataverse. The usage is still largely internal, especially for ongoing projects. From our experience in the last 20 month we find that projects are more likely to implement the policy if they have access to a data manager in the same office. To share datasets and metadata with partner organization ICRAFs prime platform is its GeoPortal. It is primarily designed to provide researchers with secure data storage, sharing and visualization options through its Web Mapping Application. Ultimately, the GeoPortal will be a full-fledged online GIS tool with a number of features for visualization, data management and spatial modeling. Being designed using exclusively open source platforms and tools it is being used by an increasing number of CGIAR scientists, as well as scientists from partnering institutions.

WorldFish

At WorldFish, the Research Data Management Project (RDMP), together with three research support people, is adding value to data collected by WF staff across all our offices. The goal of the initiative is to make all data produced by Worldfish available firstly within Worldfish and perhaps more widely given the necessary permissions. Project Leaders are required to submit well organized and well-described data files after their projects terminate. The RDMP team is trying to collate these individual data sets into a relational database that allows global analysis. WFA also just set up a GIS and data helpdesk for staff to use. ‘Clients’ can request specific data, graphs, maps and other data products. Additionally the team also manages ReefBase and The Coral Triangle Atlas which incorporate online geo-referenced data.

Success stories

Some example success stories of CGIAR open data:

AgTrials

AgTrials: http://www.agtrials.org/ is an information portal developed by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) which provides access to a database on the performance of agricultural technologies at sites across the developing world. It builds on decades of evaluation trials, mostly of varieties, but includes any agricultural technology for developing world farmers. This project will standardize data and information to the benefit of climate change analyses, future multi-environment trials and research and development in international agriculture. With the interface you can share data and information on evaluations of agricultural technology; acquire agricultural evaluation data sets for your own research; explore the geographic dimensions of agricultural evaluation.

ASTI

ASTI: Agricultural Science and Technology Indicators www.asti.cgiar.org is a database that provides up-to-date do quantitative and qualitative data on investment, capacity, and institutional trends in agricultural research and development. The dataset are collected and updated annually. ASTI data are used in informing policy formulation and decision making in many countries of Africa and South Asia.

Ethiopia Rural Household Surveys

Ethiopia Rural Household Surveys (Hoddinott and Yohannes (2011). This dataset has resulted in many publications since its release in 2011. This dataset has been cited at least 10 times. Furthermore, the publications produced with the use of this dataset have also informed policy making in Ethiopia. We receive a lot of personal request to provide this dataset as well in addition to web downloads.

Chronic Poverty and Long Term Impact Study in Bangladesh

Chronic Poverty and Long Term Impact Study in Bangladesh (Quisumbing and Baulch, 2010): According to Thomson Reuter's Data Citation Index, this particular dataset has been cited more than 24 times since its release in 2010 and downloaded 3479 times.

Land Degradation Surveillance Framework

Land Degradation Surveillance Framework (LDSF), http://gsl.worldagroforestry.org/?q=node/239, is a sampling protocol designed around a spatially stratified, randomized sampling design, to provide a biophysical baseline at landscape level. The LDSF was developed at the World Agroforestry Centre for landscape level assessments and studies of carbon dynamics, vegetation changes, soil functional properties and soil hydrological properties. The LDSF has been implemented in more than 20 countries in Africa to date, including the CIAT-led Africa Soil Information Service (AfSIS) project. The methodology has been shown to be appropriate for studies of land health and land degradation risk, as well as for assessing soil organic carbon dynamics in rangeland systems.

Poverty Environmental Network

Poverty Environmental Network PEN. Launched in 2004, PEN, http://www.cifor.org/pen, is the largest and most comprehensive global analysis of tropical forests and poverty. Its database contains survey data on 8000+ households in 40+ study sites in 25 developing countries. At the core of PEN is comparative, detailed socio-economic data that was collected quarterly at the household and village level by 50+ research partners using standardized definitions, questionnaires and methods.

Reefbase, and the CT-Atlas

Reefbase, and the CT-Atlas (http://www.reefbase.org/main.aspx; http://ctatlas.reefbase.org/) are online GIS database systems developed by WorldFish and partners. This year they have been recognized by the Thematic Working Groups of the Coral Triangle Initiative as their official data storage and retrieval tools. These databases are improving the regional coordination of conservation and management activities in the Coral Triangle region. The CT-Atlas will be maintained in future with funding from CCAFS, ADB and the IAEA (International Atomic Energy Agency) and will start to include more socio-economic datasets, and be directed more towards the examination of food security issues.

Longitudinal Village Level Studies

The longitudinal Village Level Studies (VDSA), http://vdsa.icrisat.ac.in/vdsa-vls.htm, of ICRISAT have for over three decades provided profound insights into the social and economic changes in the village and household economies in the semi-arid tropics of Asia and Africa. Over 150 research papers and more than 40 doctoral dissertations have been based on empirical analysis of VLS data in the semi-arid tropics of India and West Africa. A recent search in Google scholar shows that this body of work has generated over 10,000 citations.

Cassavabase

Cassavabase: The Next Generation Cassava Breeding (NEXTGEN Cassava) project, implemented in collaboration with Cornell University, has developed a database which gives The database (www.cassavabase.org) contains Genomic Selection algorithms and analysis capacity, a cassava genome browser, cassava ontology tools, phenotyping tools, and social networking. Tools are developed on Cassavabase that improve partner breeding program information tracking, streamline management of genotypic and phenotypic data, and pipeline that data through Genomic Selection prediction analyses. By the project end, Cassavabase will be fully hosted at IITA, providing a "one-stop shop" for cassava researchers and breeders worldwide.

CIAT Geonetwork

CIAT Geonetwork (http://gisweb.ciat.cgiar.org:8080/geonetwork/srv/en/about), the SRTM digital elevation data (http://srtm.csi.cgiar.org/) and Worldclim (http://www.worldclim.org) are great examples. Some of these datasets and associated information has been downloaded by several thousand users. The research paper published in 2005 (CIAT and Bioversity) by Hijmans et al (2005) very high resolution interpolated climate surfaces for global land areas (Worldclim)'' is one of the most cited article in CGIAR history.

Intergenebank Potato Database

The Intergenebank Potato Database (IPD), a database to cross-reference potato collections and share passport and evaluation data, since early 1995.

SINGER

(Systems-Wide Information Network for Genetic Resources): is a genetic resources information exchange network that provides access to information on the collections of genetic resources. Established in the 1990ies the collection now comprise over half a million samples of crop, forage and tree germplasm of major importance for food and agriculture.

Barriers to mainstreaming data sharing

Sensitive data & data confidentiality

Data sets often have personal information about households or informants that they do not expect to be made public and that would be unethical to make public, given the trust and rapport between researcher and 'informant'. Sensitive research data include data on illegal activities, corruption, land tenure and conflicts, controversial governmental policies, the location of wild and domesticated plant genetic resources with particularly valuable traits. Several countries have some form of data protection laws to regulate the processing of information relating to individuals and their traditional knowledge, including the obtaining, holding, use or disclosure of such information. Very few CGIAR Centers have ethical review boards or committees to ensure that data used for human or behavioral research complies with ethical review standards and that research subjects or participants are protected. While there are a number of standard procedures to anonymise data sets, such as removing names, addresses, and contact information or encryption of location data (GPS) it creates a secondary problem of managing different data sets: one complete version for internal use and the second autonomous data set for public use.

Diverse data sets and backlog

The interdisciplinary nature of the research conducted within each of the Centers results in a vast variety of data formats and types. Automated workflows for data verification, cleaning and aggregation need to be customized for each project, resulting in high demand on research support staff time. In addition global analyses across data sets from different projects requires relational databases which are difficult to realize if data collections have not been standardized. While across the Centers there are various affords to identify "key" indicators that should be collected by each project or to introduce standardized sampling designs and survey modules scientists feel that scientific creativity should not be pressed into rigid protocols. In addition, each centre has a legacy of high value data sets that have not been fully curated. To find sufficient budget to update these datasets and to make them available within and outside the centre is a problem.

Data ownership and recognition of data authors

Data cannot easily be protected against copying that work, or reproducing it without authorization or attribution. Copyright applies not to the facts or the information itself, but to the particular way the facts or information are presented in the dataset or database. As such a database can be protected by copyright, but only the database model and the data entry and output not the actual numbers or names in the database. Data sharing websites such as DataCite or Dataverse assign a persistent authorship identifier (URI), such as a digital object identifier (DOI) or handle to data sets and have specific user agreements. However, while these might be legally binding in some countries in it not clear how scientist or centers can take legal actions against misuse. While the ownership of the data and the right to reproduce the work usually belongs to the centers, scientists are given authorship rights. Unlike scientific publication there are no standard guidelines regarding data authorship. Centers and project managers need to ensure that both technical as well as scientific staff are given the deserved credit for their work, thus all people that have substantially contributed to the creation of the datasets should be data authors.

For projects that consist of multiple datasets produced by different teams of scientists, decisions need to be made about assigning the authorship. Assigning the same authorship (the same persistent identifier) to all project related data ensures the coherence of the datasets, but it does not allow differentiating between the different contributions of scientists, which is problematic with respect to accountability. The same issues arise for dynamic datasets such as R&D databases and trial data that data. Other organizations with a stronger mandate to produce global data sets as public goods such as the OECD and FAO attributing institution/program as the data authors for dynamic datasets, individual contribution are recognized within list of contributors or mentioned in acknowledgement. Assigning authorship for value –added secondary dataset is more complicated. If a person or an organization provided data, should they be included as co-author? Based on substantial contribution definition of data authorship, the input (secondary dataset) is enough for consideration for the data authorship. Several centers have policies and guidance on authorship although some of the details above are not always adequately covered in the guidance.

Institutional culture

Under the previous structure of CGIAR, the former Science Council conducted annual evaluations of two performance indicators for CGIAR research: outcomes and ex-post impact. Output indicators strongly focused on a quantitative publication matrix and were directly linked to allocation of funding. Until now annual performance evaluation of scientists focus strongly on publication rates and the Hirsch-index[2]is used as a standard indicator of scientific performance in recruitment procedures.

When evaluating research a clear distinction should be made between research 'quality' (i.e. the relative excellence of academic outputs intended for academic consumption, e.g. journal papers and books) and research 'impact' (i.e. the benefits that research outcomes produce for wider society). Unfortunately this division is often confused, a prime example being when journal citation ('quality') metrics are incorrectly presented as measures of 'impact' (Donovan, 2011). Even when journal citations are used correctly as a measure of quality CGIAR Centers need to be critical about what is to be measured. Publications are

[2] The ***h*-index** is an index that attempts to measure both the productivity and impact of the published work of a scientist or scholar.

usually reviewed based on their content, their originality and the way analysis and interpretation of the data or information is presented. Publications are seldom evaluated based on the technical rigor of the data collection procedures, the completeness of the data and its description, and alignment with existing community standards. To translate conceptual frameworks into empirical sampling designs takes significant research experience. Thus producing a high value data set that forms the basis of a high quality scientific publication requires a high level of scientific sophistication, whereas writing the paper itself requires a good grasp of language, some understanding of the science you're writing about, and an ability to "translate" technical information into plain English and write about it compellingly (Costandi, 2013). An institutional culture that simplifies research quality to counts of publications and number of technologies released does not nurture the sharing of data, but cultivates protectionism were data is viewed as the intellectual property of individual researchers.

Exclusion of data preparation and publication from the research project lifecycle

Work at the centers is still very much project driven despite the reform that aimed at strengthened and coordinated funding mechanisms linked to the System agenda and priorities. As such project lifecycles are shaped by grant requirements, which do not usually see data as part of the project deliverables. While projects adequately account for the costs of conducting research, including the collection and analysis of research data, they seldom include preparing of data and metadata and curation as part of the costs of the research process led alone long-term costs for data storage and preservation beyond the immediate lifetime of a project. Projects are usually considered closed when all grants requirements have been met, leaving preparing of data and metadata and curation unaccounted and unbudgeted for. Metadata collection and proper documentation, especially of multi-country data sets is difficult to outsource and overstretched scientists are often unable to allocate the necessary time. As data collection, if not specifically stated as project output, is often considered as a means to an end, especially in projects with a strong development focus, the necessary scientific scrutiny for sampling design and theoretical framework is often missing, limiting the value and re-use of the data. A clear recognition of the value of data in developmental work and a clear mandate to make the data available will ensure that projects will use their scientific expertise to create high value data sets.

Disconnect between libraries and data management units

Research libraries are not only places to keep collections, but their real strength and power lies in organizing, preserving, and making knowledge accessible. Unfortunately, libraries and data units in the Centres are positioned at opposite ends of the research lifecycle: data management units have been established to help researchers collect and process their data, and libraries to support the publications that result from research projects. Libraries also help arrange the search for publications as the basis for new research. While both are considered to be research support units, libraries are often anchored with the communication units and data management with the research units; each with its own directorates, reporting lines and working culture. With increasing numbers of journals requesting authors to make the replication data available and the rise of data papers the functional boundaries of librarian and data managers is becoming blurred. With the convergence of and interdependency between both data and publications, the distinctive, but complimentary skill sets of both units are needed to safeguard data availability, discoverability, interpretability and re-useability.

Relative benefits, as opposed to costs, of data publishing not clear

CGIAR has the mandate to conduct and facilitate research in development with a clear target to reduce poverty and hunger. While there are many institutions with a similar mandate, very few have a similar global outreach and the capacity to collect information and data on a variety of food production systems at the household or farm level. While it seems intuitively that making these information and data available would lead to more rapid advances in developmental questions, it is less clear what the impact pathways of data sets would be. Publishing research data comes with an increase cost, both at the project level as well as at the center management level. Both centers and donors need to be assured of the benefits of money spend on data publishing. Rather than simply assuming that open access is beneficial to the work of CGIAR a clear theory of change with an appropriate indicator matrix is needed to allow for a cost benefit analysis on the contribution of data sets to CGIAR's overall impact along the research in development continuum.

Recommendations

Clear mandate to include data as research outputs

The mandate to publish research data as research outputs needs to be explicitly stated in the Strategy and Results Framework (SRF), which sets common goals, strategic objectives and results to be jointly achieved by CGIAR and its partners. Centers and donors have to recognize data sets as important information assets and project deliverables. Only if data sets are included in the performance and impact evaluation of CGIAR, will the centers be in a position to provide incentives for scientists to allocate sufficient time to data quality at all stages of the research cycle; from development of conceptual frameworks consisting of problem definition, hypothesis, models and research questions, well-documented research designs and methods, data management and curation to statistical analysis.

CGIAR Fund allocation to support cross-center collaboration

While research data are produced at centre level and shaped by the research emphasis and character of each centre there are overarching data publishing issues relevant to all research data and all Centers. Specific CGIAR Fund allocations need to be made available to facilitate cross-center collaboration and information sharing with respect to data management and publishing. Research activities in CGIAR are funded through the CGIAR Fund, a new multi-donor, multi-year funding mechanism. It finances research aligned with the Strategy and Results Framework developed by the CGIAR Consortium and endorsed by the Funders Forum to establish common goals, objectives and results for the

CGIAR partnership. The Fund is already facilitating cross-center collaboration in addressing key research questions (i.e. gender, capacity development and communications) via the strategic research programs, but cross cutting issues like data sharing do not have a home yet.

Data Policies at Centre level

Each Center should have a research data management policy that is aligned with the CGIAR Principles on the Management of Intellectual Assets, the CGIAR Open-Access Policy and consistent with the aims of public funded research. The Policies need to lay out the basic principles of research data management at center level, address what data can be made available to the public when, how and by whom. Roles and responsibilities of individual scientists, projects and research support units towards this afford need to be stated. Legal clarity on licenses, data ownership and authorships should be addressed. Policies should be accompanied by flexible implementation guidelines that ensure that data is used by the Center and its partners in the most efficient way while safeguarding that scientists are given sufficient time to produce the scientific publications they set out to do.

Ethical committee to be established in all Centers

Each Center should have an ethical committee that is responsible for developing the centers research ethic guidelines and policies addressing sensitive data and data confidentiality, as well as appropriate handling of research data. Projects should be reviewed based on their adherence to the accepted ethical standards of a genuine research study.

Clear guidelines on authorship attribution

To ensure that both technical as well as scientific staff is given the deserved credit for their work all people that have substantially contributed to the creation of the datasets should be data authors. This includes all people that played a key role in the following:

1) Conceiving and designing the field work in response to questions of recognized scientific importance and/or relevance for developmental impact and policy change.
2) Development and implementation of research designs, choice of methods, quality control on data collection.
3) Database design, data cleaning, validation and verification processes

Zero tolerance of scientific fraud

Each Centre needs to implement zero tolerance policies with respect to data manipulation and should have explicit standards regarding the appropriate handling of research data.

Adoption of OAI-compliant data repositories across CGIAR

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. To enable the broadest level of interoperability, OAI-PMH mandates that metadata should be exposed as Dublin Core*. Data Providers* are repositories that expose structured metadata via

OAI-PMH. *Service Providers* then make OAI-PMH service requests to harvest that metadata. By using OAI-PMH compliant repositories centers can have their own individual institutional repositories each with their own particular collection policies and administrative systems, but to be linked into one large, a virtual, global repository through the use of the OAI-PMH.

Linking data and publications

To improve scientific publications, consensus with scientific peers and public trust in the quality of our research outputs each Centre should make all necessary raw data public to reproduce or replicate every scientific publication that is based on research data.

Building libraries capacity for data curation

The Centers libraries need to embrace the challenges that come with publishing research data. Libraries need to support data units in their efforts to publish research data by providing persistent identification/citation of datasets, guidance on authorships and solutions for data description, documentation and retrieval, which together facilitate findability (Reilly, 2012). They must also ensure long-term data archiving, including data curation and preservation as a condition for data interpretability and re-usability. The Centers need to invest in developing the staff skills required for achieving the data curation role in the libraries and need to of recruit library staff with experience in research disciplines.

Specific funds to publish legacy data

Legacy data sets that have high potential to contribute towards achieving the four system level outcomes should be archived and published. Most of these data sets are fully documented, however documentation and data formats need to be brought in line with today's requirements and standards. To ensure that these data sets can be made available to CGIAR scientists and partners, working in the CGIAR Research Programs, specific financial incentives need to be provided.

Changing institutional culture

Performance evaluations both at individual scientist level as well as center level need to shift from using simplistic indicators metrics such as numbers of papers, positions in lists of authors, and journals' impact factors towards assessing the quality of research itself. Centers and science managers need to put performance indicators in place that not only reward the excellent scientific writers the system has, but also the scientific and technical excellence that leads to the creation of the data, methods and ideas that are supposed to be communicated in the papers. Internal project reviews should take into account the technical rigor of the data collection procedures, the completeness of the data and its description, and alignment with existing community standards. Scientists and their field teams should be encouraged to produce peer-reviewed data papers.

Key references

AfricaRice Dataverse http://africarice.org/warda/dataverse.asp

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA (2011) Public Availability of Published Research Data in High-Impact Journals. PLoS ONE 6(9): e24357. doi:10.1371/journal.pone.0024357

Altman DG, Furberg CG, Grimshaw GM, Rothwell PM (2006). Lead editorial: Trials – using the opportunities of electronic publishing to improve the reporting of randomized trials. Trials, 7:6. doi: 10.1186/1745-6215-7-6. [PMC free article] [PubMed] [Cross Ref]

Anon (2008) Improving Research Data Management and Sharing in the Alliance of CGIAR Centres; A Working Paper for Consideration of the CGIAR Alliance, online http://cropwiki.irri.org/gcp/images/4/40/ResearchDataManagement_CG-ADE.pdf

Bioversity International Dataverse (2013), http://thedata.harvard.edu/dvn/dv/Bioversity

CIAT Dataverse, http://dvn.iq.harvard.edu/dvn/dv/CIAT

Costandi M (2013) A good story conveys wonderment, The Guardian, Monday 22 April 2013, online http://www.theguardian.com/science/2013/apr/22/mo-costandi-science-writing

Donovan C (2011) Impact is a strong weapon for making an evidence-based case for enhanced research support but a state-of-the-art approach to measurement is needed. Citations, REF 2014, Research funding, online http://blogs.lse.ac.uk/impactofsocialsciences/2011/08/22/impact-strong-weaponevidence-based-case-for-enhanced-research-support-but-a-state-of-the-art-approach-to-measurement-is-needed

FAO/Bioversity International (2012) FAO/Bioversity Multi-Crop Passport Descriptors V.2 [MCPD V.2]. 11p. CGIAR Principles on the Management of Intellectual Assets (2012) http://www.cgiarfund.org/sites/cgiarfund.org/files/Documents/PDF/cgiar_principles_management_intellectual_assets_7march_2012.pdf

Graf C, Wager E, Bowman A, Fiack S, Scott-Lichter D, Robinson A (2007) Best practice guidelines on publication ethics: a publisher's perspective. *International journal of clinical practice* 61(152): 1-26.

Hartter J, Ryan SJ, MacKenzie CA, Parker JN, Strasser CA (2013) Spatially Explicit Data: Stewardship and Ethical Challenges in Science. PLoS Biol 11(9): e1001634. doi:10.1371/journal.pbio.1001634

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International journal of climatology 25(15): 1965-1978.

Hoddinott J, Yohannes Y (2011) Ethiopian Rural Household Surveys (ERHS), http://hdl.handle.net/1902.1/15646 UNF:5:k2eYxsY6t/jVXblm/UAkRg== International Food Policy Research Institute [Distributor] V7 [Version]

ICRISAT Dataverse http://dataverse.icrisat.org/dvn

Reilly, S (2012) The Role of Libraries in Supporting Data Exchange http://conference.ifla.org/past/2012/116-reilly-en.pdf.

IFPRI Dataverse http://thedata.harvard.edu/dvn/dv/IFPRI

IMWI Water Data Portal http://waterdata.iwmi.org/

International Water Management Institute (2011) Research Data Management Policy And Implementation Guideline

Muraya P, Coe R (2001) Research Discussion Paper 2: Looking after our investments: Improving Research Data Management in ICRAF, https://sites.google.com/a/cggmail.org/cgiar-data-management-meeting/resource-documents

ODE Report on Integration of Data and Publications (2011) http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Quisumbing A, Baulch B (2010) Chronic Poverty and Long Term Impact Study in Bangladesh http://hdl.handle.net/1902.1/17045 UNF:5:8MUn92HhwQhRKF69wSTwaA== International Food Policy Research Institute [Distributor] V5 [Version]

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Wu L, Read E, Manoff M, and Frame M (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101

World Agroforestry Centre - ICRAF Dataverse (2011), http://thedata.harvard.edu/dvn/dv/icraf

World Agroforestry Centre - ICRAF Geoportal (2012), http://geoportal.worldagroforestry.org/