Platform for Big Data in Agriculture

ICARDA
Science for Better Livelihoods in Dry Areas

POWB and AR 2017

February 2018

Chandrashekhar Biradar, Enrico Bonaiuti, Valerio Graziano

Guidelines (PPA Annex 6)

"CGIAR Platform for Big Data in Agriculture"

2017 Seed Funding Recommended Allocations

Even if all CGIAR Centers are committed to making final information products and research outputs Open and FAIR (Findable, Accessible, Interoperable, and Reusable) to maximize their usage and uptake, the Centers retain different positions regarding the planning, launching, and scaling-up operations, and particularly in terms of staffing and infrastructure. As part of the Big Data Platform's commitment to achieving Open and FAIR data, metadata, and repositories, the seed funding is allocated to CGIAR Centers to help offset associated costs, including planning, initial staffing, platform development and other works related to FAIR data, metadata, and repositories (see helpful resources below).

USD 135,000 is available to each Center in 2017. This funding is not intended to provide ongoing, long-term financial support; Centers would ideally supplement this investment by co-funding as possible (e.g. via CRP Open Access/Open Data funds and line items in bilateral project budgets, in-kind investment etc.) This funding should be used by Centers to support a range of prioritized efforts towards Open Access/Open Data, with primary focus on Open and FAIR data, using their implementation plans and specific needs and gaps to target efforts. Some examples include:

Planning, advocacy, capacity building:

- Developing data and publications workflows; organizing trainings, hands-on sessions (e.g. data sprints, facilitating ORCID adoption)
- Establishing a Center-wide, cross-functional committee on data management and FAIR data best practices that meets monthly or at regular intervals (e.g. consisting of key focal points including: data manager/curator, information specialist, IT expert, communications specialist, legal specialist supporting this domain etc.)
- Developing institutional data management best practices (which may include data management plans) and/or recommendations for data management for researchers

Focal point time and travel ***(up to $10,000 total)***:

- Enabling the Center/CRP focal point to the Big Data Platform to spend time (1) staying abreast of key developments and requests from the Platform; (2) bringing Center/CRP questions and concerns to the Big Data Platform; and (3) communicating these to leadership, researchers, and any relevant Center-wide cross-functional committees

1. Enabling the Center/CRP focal point to travel to key Big Data Platform events (e.g. Big Data Convention planned for September 18-23 at CIAT in Cali, Colombia)

Staffing:

- Addressing data curation, data quality assurance needs by hiring a data manager, data curator, or repository manager[1]

Infrastructure development:

- Migrating content from a closed website or database to a FAIR-compliant repository
- Enhancing metadata and ensuring full implementation of/mapping to CG Core Metadata Schema for high-value, high-priority content (data or publications)
- Hiring developers to update repositories to meet FAIR criteria
- Migrating to the latest version of Dataverse (4.6 in May 2017)

Enhancing interoperability and the FAIRness of repository infrastructure:

- Hiring or allocating time of appropriately-skilled staff to fully implement/map existing repository metadata schema to the CG Core metadata schema
- Hiring developers to extend the functionality and interoperability of existing repositories (e.g. via tags calling on controlled vocabularies like AGROVOC/GACS and ontologies)
- Hiring developers to add new tools to the repository infrastructure – for instance, installing/launching SWORD[2], setting up the Dataverse Data Deposit API or Data Sharing API[3], setting up the PLOS Article-Level-Metrics or related API[4]

Enhancing the FAIRness of data:

- Hiring appropriately-skilled staff to ensure usable data (e.g. quality metadata within datasets to explain column headings, treatments etc, or accompanying code books/readme files)
- Allocating staff time as needed to identify and prioritize FAIRness and uploading to the institutional repository of high-value, low-risk, low-cost legacy data sets (e.g. using the data prioritization matrix in the "Advocacy and Guidance" section of the CGIAR OA/OD Support Pack, or other framework)
- Allocating legal time (Center staff or consultant) to ensure appropriate licensing to enable reuse (for data, software etc.)

Repository-related subscriptions:

- Testing or subscribing to services such as Plum Analytics[5], Altmetric.com[6], or Impact Story[7]
- Subscribing/implementing services such as ORCID[8] for Center-wide author IDs

Helpful Resources:

Wilkinson, M.D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3:160018. DOI: 10.1038/sdata.2016.18

Olavo Bonino, L. 2017. An Ecosystem to Support FAIR Data.
https://www.slideshare.net/secret/wmta73Yxmock6g

[1] It is recommended that long-term staffing and/or part-time FTE needs be funded through a combination of Big Data, CRP OA/OD funds, and bilateral budget lines.
[2] SWORD: http://swordapp.org/about/
[3] Dataverse API Information: http://thedata.harvard.edu/guides/dataverse-api-main.html
[4] PLOS ALM: http://api.plos.org/alm/using-the-alm-api/
[5] Plum Analytics: http://www.plumanalytics.com/
[6] Almetric: http://www.altmetric.com/
[7] ImpactStory: https://impactstory.org/
[8] ORCID: http://orcid.org/

Activity Plan (POWB)

"CGIAR Platform for Big Data in Agriculture"

Background

The CGIAR Platform for Big Data in Agriculture was created to promote, support, and accelerate the efforts of CGIAR Centers and Research Programs to both leverage and shape the digital transformations happening in global agriculture. The Platform program is organized according to three modules which also serve as strategic pillars: "organize"," convene" and "inspire". "Organize" is related to supporting the efforts of the CGIAR network to share open and/or Findable, Accessible, Interoperable, and Reusable (FAIR) data in order to accelerate research efforts and enable new collaborations. "Convene" is related to building new partnerships to help shape and leverage digital technologies in support of global agricultural development. "Inspire" Centers on using data to push the limits of research and/or transform ways of doing business.

Duration: 1-Jan-17 to 31-Dec-2022

2017 Activities:

Module 1:

ICARDA contribution to the CGIAR Open Access Policy, http://hdl.handle.net/10947/4488 has been initiated in 2015 with the drafting of the first open access / open data implementation plan with the support of Agroknow, a company based in Greece with extensive experience working on international initiatives including e.g. FAO, BMGF and CGIAR Centers. In 2016 the implementation plan has been reviewed in like with the template provided by the CGIAR system office and a special committee was formed to analyze the needs and implement solutions based on resources available. The CRP on Dryland Systems provided initial support. The program partially sponsored the development of Geo-Informatics portal, the formation of a community of practice among CG and external members and the development of online training modules in OA/OD and IP. The CGIAR BIGDATA platform funds for 2017 will be invested in several strategic initiatives in accordance with the policy, the implementation plan and the suggestions provided in annex 6 keeping in consideration the amount provided in order not to create false expectations.

In 2017 the Center will ensure that data and publications are stored on the latest versions of interoperable cloud-based (AWS) repositories (D-SPACE / DATAVERSE) through proper workflow and in accordance with CGCORE metadata schema including usage of AGROVOC and ORCID. The core group responsible for the implementation of the OA/OD is in place and functioning, ensuring that most of past and present content from closed website/database are available in the repository. The team will also study potential analytics to be included in the existing repositories. The group will be responsible for the stimulation of joint trainings among different disciplines of data producers/users in the Center and the exploration of advanced trainings and data management practices already developed by other institutions for 2018. The Center focal point will represent ICARDA at the platform level and provide feedbacks to the OA/OD group.

2018 Activities:

The deployment of the latest advanced tools, such as Breeding Management System (BMS), Open Data Kit (ODK), update of the GeoAgro, GeOC and MEL, requires an updated capacity able to use them at their full potential. The BigData enhancement will be undertaken by collecting the spatial and non-spatial data and development of metadata for all the data gathered to ensure standards to feed into the GeoAgro and MEL portals. The BigData Capacity course to train ICARDA scientists on the use of the latest technologies available to the Center through the BigData Platform, Breeding Management System (BMS), Geoinformatics, Open Data Kit (ODK), Monitoring Evaluation and Learning (MEL) and data quality and standards. The course will treat a) data collection b) dataset reporting and metadata c) Intellectual Property (IP) and copyright essentials for knowledge and technologies d) use of BMS, ODK, GeoAgro and GeOC to capture and disseminate datasets. The course (s) might include the formation of one or more Data Managers, meta-data creator/curator, best practices and standards for spatial, not spatial, field data and other specialised datasets and related publications. The course implementation will make use of the ICARDA e-learning platform and video calls (10,000 USD). The Breeding Management System (BMS) Course will be implemented on the foundation of the BigData CapDev course and be focused on the BMS platform (7,000 USD). The 2017 and 2018 reporting operations will carry on the BigData CapDev objectives, by including within the perspective of data quality assurance also an ICARDA scientists and secretariats reporting training, with special attention to ORCID, copyright and IP essentials (6,000 USD).

The necessity of systems interoperability, data FAIRness and advanced tools shall be further addressed through several activities during 2018: the development and actualisation of the MEL-MARLO interoperability, also for the MEL/Dataverse interoperability with BMS using the Breeding API (BrAPI) and ODK Aggregate server, the implementation of SharePoint/OSCAR and the GeoAgro-MEL enhancement for better usability and performance with learning analytics, the implementation of Analytics for Centers' repositories to be actualised in collaboration with CODIS, the enhancement of the global datasets tool GeoAgro for enhanced quality and interoperability, including metadata review, activation of basic analytics and study of future analysis capabilities.

The BIGDATA focal point together with the OA-OD Implementation Plan team will provide assistance to scientists within the center. It is also expected the participation to related conferences where the work under BigData will be presented with proper branding. The team will also proactively draft project proposals as long as suitable calls are available.

Contacts

- Administrative contact

Name	**Tareq Bremer**	**Ralda Gareg**
Position	Grant management focal point	Finance focal point
Title	Grants Management Officer	Budget Officer
Address	2 Port Said, Victoria Sq Ismail El-Shaaer Building, Maadi Cairo, Egypt	2 Port Said, Victoria Sq Ismail El-Shaaer Building, Maadi Cairo, Egypt
Phone	+20-2-35724358	+20-2-35724358
Email	T.Bremer@cgiar.org	r.gareg@cgiar.org

- Technical contact

Name	**Chandrashekhar Biradar**	**Enrico Bonaiuti**	**Francesca Re-Manning**
Position	Main technical focal point	Monitoring, reporting and repositories focal point	Intellectual Property and Legal focal point
Title	Geoinformatics Unit Head	Monitoring, Evaluation and Learning Specialist	Legal Counsellor
Address	2 Port Said, Victoria Sq Ismail El-Shaaer Building, Maadi Cairo, Egypt	8 Yousef Sukkar St., Hay Rawabi, Wadi Seir Area PO Box 950764, Amman, Jordan	138 Sedgwick street Cambridge, UK CB1 3AL
Phone	+20-2-35724358	+962-6-5920120	+497443330034
Email	C.Biradar@cgiar.org	E.Bonaiuti@cgiar.org	F.ReManning@cgiar.org

Name	Katrin Park
Position	Communication focal point
Title	Head of Communications
Address	2 Port Said, Victoria Sq Ismail El-Shaaer Building, Maadi Cairo, Egypt
Phone	+20-2-35724358
Email	K.Park@cgiar.org

Report

Contractual Process and Budget Allocation

On 25th August the BIGDATA platform Lead Center (CIAT) initiated the PPA process with ICARDA for a 2017 allocation of 135,000 USD. ICARDA PDGMU received additional information (revised annex 6) and the internal team developed an initial activity plan together with budget allocation (OCS and CIAT templates). On October 24th all documents were submitted to CIAT and first transfer amount (68,000 USD) was transferred on December 8th. The budget code (BUS) was opened on October 24th aiming at utilizing at least 50%. However, the optimistic target could not be achieved for several reasons, detailed in the cost analysis below. In two months (Nov-Dec 2017) we were able to utilize 17% of the funds and overall to commit 45%of the total allocation. This is a demonstration of the institutional commitment to the BIGDATA objectives within the current framework, including internal processes to invest the funds received and to actualize the expenditure.

On March 8th the ICARDA finance department informed that the external auditors informed that consultancy expenses should be planned and reported under supplies and services instead of personnel. The BIGDATA team planned consultancies under personnel as for other Donors (GIZ). This change will be notified to the lead center in order to re-adjust the budget allocation requested in October 2017.

The summary below is presented based on the POWB submitted (i.e. Consultancies planned in personnel).

Personnel (incl. Consultancy): 100% utilization of staff time for the coordinator and the manager of data workflow. 35% utilized (50% committed + 15% available) of planned consultancy work. Underutilization is due to the process in developing consultancy contract. CRP funds do not allow the Center to initiate commitments with external sources (suppliers, consultants) before the opening of a BUS. This caused the accomplishment of fewer results, to be reported in the first half of 2018.
Travel (incl. Training): 60% utilization (25% committed + 15% available) of planned travel expenditure. Most of the travel costs were dedicated to ensure ICARDA contribution to the first BigData convention. 0% utilization (0% commitments) for training activities. The amount allocated to trainings was moved to 2018 since the Breeding Management System (BMS) Training has been postponed to February 2018. The BMS is an important component of ICARDA BiGDATA, however these start-up funds should be complemented by the investments of crops CRPs (WHEAT, GLDC, Genebanks) and the allocation from the corporate IT charges (charge-back funds) since the BMS is a corporate data storage system and its maintenance, staff training and server costs (AWS) should be centralized.

Supplies (incl. operational): 26% utilization. (100% committed). Similar to the issue experienced with consultancies. We were only able to commit funds with software developers. However, the suppliers delivered several results presented below. The AWS support has been conceived only to initiate the cloud storage for BMS and the International Nursery. However, we should ensure a sustainable approach from 2018 onwards since we have several corporate tools such as Data & Publications Repositories, Open Data Kit (ODK) Aggregate Server, GeoSpatial Tools (GEOC), GeoSpatial Portal (GeoAGRO), M&E tools (MEL) and the IT charge-back is benefitting only OCS as a financial system.

Fundraising and cost sharing

In 2017 the ICARDA team, while the funding confirmation was still pending, supported the overall goals of the BIGDATA platforms in the areas below with a contribution of 110,000 USD:

1) Ensuring BIGDATA as a cross-cutting theme with indicators for the new ICARDA strategy that will be launched in May 2018 and update and review the ICARDA OA-OD Implementation plan and its committee. (Staff time contribution: 10,000 USD)
2) Developing new project proposals having BIGDATA as central theme. One project proposal submitted to USAID in partnership with US Institutions. If approved it will be mapped to BIGDATA. (Staff time contribution: 8,000 USD)
3) Amazon Cloud Computing Investment (50,000 USD)
4) Communication investment to develop an Application Programming Interface (API) to interrelate the Website with Publication (D-SPACE) and Data Repository (Dataverse). (12,000 USD)
5) Consultancy investment to tag corporate resources using standard metadata schema (CGCORE). (5,000 USD)
6) Participation to CGIAR COPs (DMTF, OAWG, MELCOP) and contribution to documentation and System requirements in relation to BIGDATA (20,000 USD)
7) Preparation and installation of the Breading Management System (BMS) to be launched in 2018 (5,000)

Implemented activities

In addition to the activities implemented using its own funds and resources, the ICARDA team has contributed to the CGIAR BIGDATA Platform in the areas mentioned below with an investment of 22,500 USD (utilized) or 61,365 (utilized + committed) within the overall allocation for 2017/2018 of 135,000.

Planning, advocacy, capacity building

- Development of internal workflows for Publications and Data appointing key staff for their testing and feedbacks.

Focal point time and travel

- Participation to the BIGDATA Platform Convention and contribution to its start-up.
- Contribution to the review of CERES.

Staffing

- Review Terms of Use and Privacy Policy of data storage for staff.
- Preparation of Online User Guide for staff while adding metadata to data and publications.

Infrastructure development:

- Initiate the migration and tagging of historical information from closed repository to open ones as per OA-OD Implementation Plan.
- Review of terms and definition for our controlled listed using internal references as COAR, GODAN, Google, Library of the Congress).
- Update D-SPACE and Install the latest version of Dataverse in partnership with the CRP LIVESTOCK (Strategic Innovation funds).

Enhancing interoperability and the FAIRness of repository infrastructure:

- Activation of three webservices to optimize data and publication tagging within the framework of the draft CGCORE. The three web services are: AGROVOC, ORCID and SHERPA/RoMEO.
- Develop harvesting and depositing processes in collaboration with other CGIAR Partners (CIP&CRP-RTB for Dataverse; ILRI &CRP-Livestock for CGSPACE).
- Developed network analytics among ICARDA scientists to understand trend in publications and datasets.

Recommendation to ICARDA Management

1. Ensure an annual allocation from the research support fund linked to the results of BigDATA in ICARDA's SRF. Such allocation should also ensure the Capacity Development of existing and new staff.
2. Ensure that ICT based infrastructures, maintenance, related subscriptions and cloud servers are funded through the mechanism of the IT Charge Back.
3. Commission the development of Data Management Guidelines to be adopted as early as at the proposal development stage. For this objective it is possible to seek a co-funding from BIGDATA.
4. Validate the 2018 BigData Platform plan based on the 100,000 USD expected contribution. Ensure that the plan does not cover corporate investments but is focused on projecting ICARDA into the BIGDATA arena, while supporting global partnerships with students (IBM Data Science for Social Good Research Fellowship, Stanford University, etc) in collaboration with the CDU.
5. Approve the new ethics guidelines drafted in 2017 by DDGR, the Team Leader-Socio-Economics Component, and the Legal Counsellor with the support of an external consultant.

Recommendation to CGIAR and BIGDATA Platform

In order to ensure maximum accountability, empower researchers and inspire investors, it feels necessary to adopt and follow strict, joint copyright and intellectual property regulations, in addition to invest in information security, in line with 2018 EU updates on personal data. Such measures shall enhance the prestige of all platforms implemented by CGIAR and related Centers

through strong reliability, meaning better guard and management of sensible data which in turn is likely to attract a larger number of investors.

Data safety is a matter that concerns all actors, institutions, organizations and scientists. By adopting high information security standards, the CGIAR will deliver a strong signal of proficiency in data management, making scientists more confident in reporting their products and meeting the needs of actors attentive to sensible data, growing trust and confidence toward the System Office.

Annex 1: Focal point updates (January 2018)

Module One leader Medha Devare and Brian King spent last week in Lima for a workshop with developers of the Highly Interactive Data Analysis Platform (HIDAP), a data collection and management tool developed initially for breeders that has since expanded into agronomy uses as well. Over the course of the week colleagues from CIP, IITA, CIMMYT and the Big Data Platform examined how to harmonize field book efforts and map them to common ontologies. HIDAP presents a nice example of data integrity--interlinked processes for data collection, analysis, storage, and use--that can be built on to great effect to be a force for data quality and unlock new analyses at scale.

In the months since the Big Data Convention I have checked in with multiple efforts to develop digital interventions in food systems as well as the science infrastructures that can enable them to scale. I visited U.S. and European universities, cloud computing providers, three CG Centers, and IBM Research. I learned a few things about the state of the conversation going on across the digital science infrastructure landscape:

- **A majority of actors are still getting their bearings.** They are examining such themes as: What are the bottleneck issues, what alliances must be built, what standards must be developed?

- **Open research data is gaining momentum**, but there is still work to be done developing the *tools, services, and infrastructures of reference* for our sector.

- **There is pent-up demand for the CGIAR** to shape how these digital science infrastructures develop in service of the global public good.

- **We are, for the moment, a half-step ahead of several other efforts:** we have a *prototype of a CGIAR-wide indexer of research products and associated datasets*, a task force *working through the complexity of data ethics guidelines*, and a genuinely *multi-stakeholder agenda for action* in our sector that was ratified at the last Convention.

PLATFORM Business

- **Carry over rate from 2017:** As you know, the Big Data Platform Management Team decided to set a **target carry over rate of 50% for 2017.** The late start in the year made it challenging for some Centers to hit that target. It will serve the longer-term health of the program well if we aim to execute 2017 funds aggressively yet responsibly.

- **Funding expectations for 2018:** The final 2018 numbers have still to be transmitted to the Big Data Platform, though we do have workable-enough estimates to begin to plan. More details under the Module updates below.

- **Welcome to Mathilde Overduin!** Mathilde just joined the Secretariat to help with project management and a little of everything else. *We'll be reaching out to you all individually to set up calls to check in on how things are going with 2017 activities and what you have in mind for 2018.* This will help us maintain a good global overview of how the program is unfolding.

ORGANIZE

- **Guidelines for expending Module 1 funds were included as Annex 6 the PPA** (it is attached for reference). We'll have this in mind when we check in on how your open access/open data investments are progressing.

- The Big Data Platform **prototype data harvester** *CGIAR e-Research (CeRes)* continues to discover datasets and publications across all public CGIAR databases. A new functionality for data upload was recently developed and is being tested. We aim to have the harvester ready in the next 3-4 months as one good-quality, CGIAR-branded solution for discovery of public food security-related data.

- **Organize Funding expectations 2018.** As you know in 2017 grants of *$135,000* were made to Centers to invest in their own implementation of the Open Access/Open data policy (Annex 6 to the PPA outlines recommendations). In 2018 we should be able to do follow-on grants of *$100,000 in 2018,* or more. In addition, we aim to identify some creative ways to support emerging innovations or high-impact investments in data mobilization.

- **Results of the requirements gathering survey so far.** The **requirements gathering survey** has yielded some rich results. You will not be surprised to know that there are varied needs, capabilities, and budgets across Centers related to storage, computation, datasets, and building capacity for 'big data.' For example, genomics researchers and some climate researchers report that they need high performance storage and computing power, and find that they prefer the cost-performance of doing this on-site rather than in the cloud. Some Centers report that some amount of cloud-based data archiving and store-and-compute services would be valuable to them. As a result, we will need to segment Center needs a bit more to identify where shared services can provide value. (*A synthesis of the survey results to date is attached.)*

CONVENE

- **Communities of Practice** are forging ahead on their work plans. We have prepared a draft survey that will be used for a 360 review, in February.

- The **Privacy and Ethics Task Force** should have some draft products next month that we will then aim to turn into guidelines for *responsible data practices* for our sector.

- We have an emerging collaboration with the **IBM Data Science for Social Good Research Fellowship.** Under the program, IBM recruits talented and capable computer science researchers and they take on a project with a social good mission. We notionally identified using machine learning/AI techniques to cut through the noise of large time series-based datasets to help improve decision support, but ***your ideas are welcome for concrete research opportunities for one or more fellows.***

- We also have an **emerging collaboration with Stanford University** whereby we can enlist groups of computer science undergraduates to analyze large data sets (under the supervision of a Computer Science professor with food security domain expertise). ***Let me know if you might have such a finite analytic project we can send their way.***

- Our discussions with Microsoft (both Azure and Research) about how to partner are ongoing, but along the way I happened to pick up a **$20,000 Azure Credit**. ***Please let me know if this will be of use for you (it expires in December 2018).***

- **The Big Data in Agriculture Convention** will be happening in September or October this year. More details soon.

- **Community of Practice funding** should remain at 2017 levels in 2018.

INSPIRE

- The researcher we engaged to develop an **Inspire Challenge synthesis** has produced a near-final draft, but it needs some copy editing. Hopefully this will be ready for distribution within the next couple of weeks.

- **Inspire Challenge Awardees from 2017** are fine-tuning the Impact metrics for their projects.

- **Inspire Challenge 2018.** We plan to run *another Inspire Challenge this year*, culminating at the CGIAR Convention for Big Data in Agriculture. Details to follow.

Fundraising

Should any funding opportunity be available in the program and country environments of your interest, related to open data or "big data" analysis, the Big Data Platform will gladly oversee the opportunity to help you in pursuing and obtaining them for your Center or Program. Recently, we offered to provide a cost-share on high-resolution satellite imagery to all members of the Consortium on Spatial Information who were preparing concept notes for a USAID solicitation seeking innovative approaches to monitoring and evaluation.