Platform for
Big Data
in Agriculture

CGIAR

CGIAR PLATFORM FOR BIG DATA IN AGRICULTURE
PLAN OF WORK AND BUDGET
2018

CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

CGIAR Platform Plan of Work and Budget (POWB) for 2018

1. Expected Key Results

1.1 Adjustments/ Changes to Your Theories of Change

The Theory of Change (ToC) of the Platform focuses on increasing the capacity of CGIAR and partners to embrace big data and ICT approaches. This is achieved through 1) Enhanced collaboration across Centers, CRPs and partners to leverage appropriate state-of-the-art data standards, analytics and ICTs; enabling unrestricted discoverability of inter-linked data sets to understand and tackle multi-faceted challenges in data and evidence-driven ways; leveraging CGIAR expertise with external partners to broaden and deepen experience and capacity and position CGIAR in the broader big data and ICT sphere to establish CGIAR as an innovative thought leader; and development of initiatives using proven big data innovations to drive agricultural growth in developing countries. No major changes are envisioned in this ToC from the proposal, though the strategy to achieve it has been developed much more deeply and is outlined in this POWB.

1.2 Expected Platform Outputs

2017 was a year of establishment for the Platform, which included investments in terms of the governance system, leadership team and establishment of legal agreements with the centers to promote improved data management across the system. 2018 is the year of aggressive implementation for the Platform on Big Data in Agriculture, having all the foundations now in place.

The primary focus in terms of effort and resources in the early years (including 2018) of the Platform is on getting CGIAR data resources organized and available (FAIR – Findable, Accessible, Interoperable, Reusable compliant). The Platform is working both centrally on establishing the infrastructures and systems to make CGIAR data visible and usable, but also decentrally through all 15 centers to build the capacity of the CGIAR system to effectively manage its valuable data resources. Key outputs under this work will include deepening functionality and subsequent release of the CGIAR flagship data harvester which has been branded as GARDIAN. GARDIAN enables the discovery of publications and datasets from the thirty-odd institutional repositories across all CGIAR Centers to enable value addition and innovation via data reuse. It employs text-mining to enrich the associated metadata to enhance discovery, and will soon test data mining techniques with cleaned, well-annotated datasets to enhance interoperability. 2018 plans for GARDIAN include further demonstration of the value of interoperable data via seamless interactivity of discovered data with key analytical/visualization tools, including a model and maps. The Platform will also invest in data quality and harmonization via standards by promoting ontology-based field books for data collection, providing data cleaning tools and services, and improving data annotation and curation where necessary. It will also be instrumental in aligning CGIAR's open data methodologies, standards, and ontologies with the wider open and FAIR (Findable, Accessible, Interoperable, and Reusable) data community. Center capacity on standards and data management to enhance discoverability, interoperability, and reuse will be significantly enhanced through sub-agreements and seed grants to all 15 centers. Amongst other outputs, these sub-agreements will deliver high quality datasets that are well-annotated using standard ontologies and controlled vocabularies for improved discoverability and reusability via GARDIAN, capacity building materials and events to enable consistent approaches to data privacy, standards such as the CG Core metadata schema,

well-stewarded and mature ontologies including crop and agronomy ontologies and a draft socioeconomic ontology, and an early prototype of an ontology-based agronomy field book.

Mobilizing CGIAR data to accelerate research and spur new data-driven innovations, building collaboration across our organization and with the wider sector, leveraging CGIAR expertise while claiming a leadership voice in digital innovation for agriculture--the Platform aims to effect several digital transformations both inside and outside of CGIAR. In order to achieve this and measure progress, we will need to continually assess the state of digital "readiness" inside CGIAR and examine its links to the data innovation needs of the wider sector. The Platform will undertake the first such expansive strategy assessment this year to produce a "State of the Network" report and multi-stakeholder plan of action.

Significant effort continues in 2018 to build novel partnerships for the CGIAR, following the Theory of Change to position the CGIAR as the go-to-place for expert knowledge and experience on digital agriculture in developing economies. New partnerships for the CGIAR will be established with companies offering cloud computing and storage (negotiation is ongoing with Microsoft Azure and Amazon AWS), providers of high resolution satellite information (e.g. DigitalGlobe), data analytical infrastructures (e.g. University of Minnesota Supercomputing Institute, IBM), machine learning service providers (Google Tensor Flow, IBM Watson), and partners providing new pathways to data analysis or impact at scale (e.g. the Digital Impact Alliance to liaise with the mobile industry, the GSM Association, and Dalberg Data Insights to model new big data analytic methodologies and help integrate them into CGIAR research).

The Platform will continue to support six communities of practice (COPs) across the centres and with external partners. These CoPs will generate greater integration across the CGIAR system and with partners around these themes, and also focus on collective action to produce the following outputs:

- A design of a minimum viable product for digital, on-farm decision support from the Data-Driven Agronomy Community of Practice, and webinars to validate digital tools and technical approaches.
- A high-level ontology and set of suggested standardized questions for harmonizing the collection, organization, and re-use of socioeconomics data from the Socioeconomic Community of Practice.
- A gap analysis of modeling approaches and capabilities within the CGIAR and partners, as well as results from three modeling 'mini-projects' from the Crop Modeling Community of Practice
- An enhanced geospatial data cataloging and analysis tool and its codebase released to the public by the Geospatial Community of Practice via GARDIAN
- Tools developed by the Ontology Community of Practice, such as a metadata quality checking tool for Dataverse, a lookup service for agricultural data annotation, and a matrix of existing standard vocabularies/ontologies with their guidelines/tutorials, mapping solutions between ontologies and current gaps for knowledge domains.
- A compilation of the Data-driven Agronomy CoP's perspectives on the Agronomy Ontology and the fieldbook prototype, by the Ontology Community of Practice
- Improved discoverability of livestock data in GARDIAN, and development of a functional ontology for livestock data.

Emerging data resources under GARDIAN and the emerging new partnerships will be leveraged through a new INSPIRE innovation process. New topics for the INSPIRE call will be identified together with CRPs, and five $100,000 grants will be provided to winning projects.

The Platform-organized "Convention on Big Data in Agriculture" will be held in October in Nairobi, with a focus on Africa, and will be the venue to cement many of the new partnerships, and kickstart the Inspire innovation processes. Over 200 participants are expected, and the Convention will serve a dual role of positioning the Platform as the go-to-place, enhancing CGIAR capacity on digital agriculture related topics, and getting Platform business done through side meetings, training workshops and governance meetings.

1.3 Plans by Platform Modules

1.3.1 Module One ("Organize")

Investments in open and FAIR data will continue in 2018 building on early successes that include an exponential increase in the number of discoverable datasets and publications, strong buy-in to making information resources interoperable (aiming for a score of 4 or higher on the 5-point Data Archiving and Networked Services or DANS scale), explicit focus on Center data sharing and knowledge management such that each Center continues to add 50 or more well-annotated datasets to their repositories, with more staff time dedicated to these, and strategizing to ensure ethical and responsible sharing and use of data across all Centers. GARDIAN already enables CGIAR's information resources to be searched and discovered no matter in which institutional repository they reside, and by the end of 2017 showcased about 50,000 publications and 1800 datasets. In 2018 this harvester will enable discovery of at least 100,000 publications and around 2500 datasets. Its functionalities will be further enhanced to allow data filtering in a variety of ways, and upload by partners without institutional repositories. Enhancements will also allow discovered, contextually-linked datasets to be easily integrated and analyzed or visualized via an analytics and mapping capability. Collaborations with the University of Minnesota, Norwich University, Wageningen University and others will be leveraged to explore state-of-the-art ways to enable secure data storage, appropriate handling of private information, and improved data annotation and curation capabilities to enhance interoperability, and a secure analytic environment meeting the needs of CGIAR researchers. Module 1 will also ensure high-quality agronomic data at the collection stage through an ontology-based field book, to be developed as an extension of CIP's HIDAP breeding platform, with quality checking, statistical analysis capacity, one-click upload of collected data with associated metadata to institutional repositories, and semantic web capabilities to enable easy aggregation, integration, and querying across related datasets. This effort will bring together allied interests and efforts across a number of partners, including CIP, Bioversity, IITA, CIMMYT, CIRAD, the University of Florida, the University of California - Davis, Wageningen University and Research, and others. Most Centers are implementing the CG Core metadata schema as a result of prior efforts; a mature version of the schema will be finalized and released to enable harmonization and consistent metadata across repositories and associated tools, including field books where possible. This module will also prioritize capacity building through a blend of online and in-person efforts focused on enabling standards in data and metadata annotation to maximize interoperability, machine-readable licensing, improved and consistent approaches to data ethics and privacy concerns, and addressing data quality.

1.3.2 Module 2 ("Convene")

In 2018 the Platform will invest in shared services to address common research data needs. In 2017, the platform surveyed centers on their key infrastructure and shared service demands, and identified a number of areas to make system-wide agreements to build the capacity of centers and CRPs to deliver on digital agriculture potential. One area where opportunity was identified was in

high resolution satellite imagery, with many centers demanding this but lacking access. The Big Data Platform will negotiate an agreement with Digital Globe to source their images, and big data related analysis packages and make this available for CGIAR research operations. Common data sharing and analytic environment needs were also identified as priority, where CGIAR-generated data can be combined with other data to facilitate modeling or other analysis. Agreements with cloud computing initiatives will be sought in 2018 to solve this need. Pilot approaches to data curation will also be developed as a shared service for CGIAR researchers.

The Platform will again produce an annual Convention on "Big Data for Agriculture" with an agenda and invite list curated specifically to foster alliances around agriculture development themes. This year the event will take place in Nairobi in October, and will focus on developing new partnerships with upstream and downstream digital agriculture players in Africa. A technical leadership voice for the Big Data Platform is being developed and adopted, as well as for the CGIAR as a whole and the Convention is becoming a key element to that component of the Theory of Change. The Platform will implement a communications campaign in 2018 that coalesces the Convention and efforts from the Communities of Practice to position the Platform and the CGIAR as a leader in the emergent discipline of data for agriculture development.

Finally, the Platform will continue to support open Communities of Practice to advance leading-edge work on technical themes of importance to the sector, and communicate widely about critical issues for data-driven agriculture development, and identify solutions to advance the discipline of applying data for agriculture development. For 2018, one additional Community of Practice is part of the portfolio, with external funding to support it, focused on data for decision making in the livestock sector.

1.3.3 Module 3 ("Inspire")

In 2017, the Platform funded 5 novel innovations in the fields of data driven farming and pest/disease surveillance. The Platform will closely follow the progress of these projects, and extract lessons learnt on the successful deployment of digital solutions to these problems, and provide synthesis and communications materials on these topics.

In 2018, the Platform will again provide up to five (5) one-year grants of US$100k pilot funding in 2018 for innovative data and information and communications technology (ICT) ideas that have the potential to lead to enhance the efficacy of development efforts. Topics for the Inspire call, which will culminate in the selection of ideas in the Convention in October, will be defined in conjunction with CRPs to ensure a close match to the needs of the broader CGIAR research portfolio, ensuring leveraging of the funding and high likelihood of further development of the ideas under the CRP in subsequent years. Initial priorities for the Inspire call include filling the gender-disaggregated data gap in conjunction with the Gender Platform, and possibly re-issuing a call for ideas on under-subscribed elements of the 2017 call on big data for understanding food systems and nutrition, novel approaches to impact assessment, and data driven farming.

In addition, the Platform will award up to $250,000 "Scale-up funds" for one winning Inspire Challenge winners of 2017, to take their projects to the next level. This will be awarded based on robust assessment of the emerging results from the initial pilot, and will be agreed together with the relevant CRP who would be expected to co-invest in the scale-out.

1.4. Cross Cutting Dimensions

1.4.1 Gender, Youth and Capacity Development

The Platform aims to increase the likelihood of CRPs reaching their targets and CGIAR IDOs. Big data and ICT solutions will contribute to gender and youth IDOs and sub-IDO's. In 2018 the Platform will further work to increase the ability of individuals to manage and use big data, and generate greater organizational capacity to embrace big data approaches. It will provide technology that enables women and youth to access and use information that improve farming and livelihoods.

- **Gender**: The Gender Platform assists with gender review and co-design of key activities with the Platform for Big Data in Agriculture, as well as specific collaborations in research, communications, and innovation. 2018 datasets made open access through Module 1 activities will be gender-disaggregated wherever applicable, and annotated with metadata and ontologies documenting the gender sensitive protocol; this will be achieved in close cooperation with the CGIAR Community of Practice on Gender. Data standards developed in Module 1 will also include appropriate metadata for representing the gender dimension. Under Module 3, a topic on filling the gender data gap is considered for an INSPIRE call, and is being actively discussed with the Gender Platform under PIM. Finally, significant communications outreach will focus on encouraging greater participation of women in digital solutions, including testimonies for International Women's Day.

- **Youth**: It is expected that increased access to agricultural data and ICT-based applications will help retain and attract more rural youth to agriculture. Young men and women will be considered a major target user group for the annual data/knowledge consultations and design of communication materials. In 2018 significant youth involvement from within the CGIAR and with partner organizations will be sought in both the Convention and in CoPs to bring new ideas to the table. Similarly to gender, some 2018 Inspire project topics may specifically address youth related impacts. Given the technological nature of big data innovations, it is to be expected that youth are more likely to engage in projects given their additional capacity to manage ICTs. Likewise, impact assessment efforts will look at youth related impacts of different innovations.

- **Capacity Building**: The entire Platform is essentially a capacity building platform for the CGIAR to up its game on digital agriculture. Investments with the 15 centers will improve their capacity to manage valuable data resources at center level, and provide the systems, culture and capacity to leverage CGIAR data resources through GARDIAN. Specifically related to capacity building activities, in 2018 the Platform will:
    - Implement a workshop series with Big Data Platform focal points and Center and CRP Data Managers, leveraging competencies from within CGIAR.
    - Curate or create relevant online learning materials related to open data and 'big data' analytics.
    - Develop a model position description that describes the key capacities needed for data managers or data management consultants, for potential use by Centers who may wish to use the OA/OD grants to recruit for these skills.

1.4.2 Open Data and Intellectual Assets

Supporting implementation of the CGIAR Open Access and Data Management (OADM) Policy is central to the Big Data Platform impact pathway. The Platform will advance this objective concretely in the following ways in 2018:

- Further build and expand the data harvester (GARDIAN) to enable easy, cross-domain search and identification of research products and datasets, and showcase its value by enabling seamless interactivity of the discovered data with analytical and visualization tools.
- Invest in data quality and standards in annotation by promoting ontology-based field books for data collection, providing data cleaning tools, and curating data employing interoperability standards.
- Harmonize open and FAIR data methodologies, standards, and ontologies with the wider CGIAR community while serving as a conduit for learning and new approaches from beyond CGIAR through strategic partnerships. In line with this, all centers will receive Module 1 seed funding in 2018 to make strategic investments that can accelerate progress towards OADM compliance.
- Engage an intellectual property lawyer to collate and develop best practices and guidelines from among ethical and legal frameworks for responsible generation sharing, and use of data.

2. Planning for Platform Effectiveness and Efficiency

2.1 Platform Staffing in 2018

The Big Data Platform proposal calls for creation of a Secretariat of up to 5 people which was progressively established in 2017, and will be finalized in 2018. In addition to the Platform Leader (rebranded as "Coordinator" for external communications) the Platform Secretariat consists of the following positions for 2018:

- **Communications Coordinator:** Position the Big Data Platform and the CGIAR as a leadership voice in digital agriculture.
- **Project Manager:** Assist with communications and with tracking progress and outcomes of over 20 projects per year.
- **Data Scientist/Data Science Advisor:** Serve as technical lead on analytics tasks, advise on capacity building and scale-up of leading edge approaches to data analytics and 'big data' methodologies across the CGIAR.

In addition, the Steering Committee has approved recruitment of a Program Manager for Module One who will be recruited by IFPRI and serve as liaison between CIAT and IFPRI (the Co-leading centers of the Big Data Platform) .

2.2 Financial Plan for 2018, including use of W1/2

Expected uses of W1/W2 funds are generally consistent with those outlined in the original Proposal. See Table D for more details. Significant carry-over from 2017 is planned to be executed in 2018, much of which is already committed with agreements from late in the 2017 financial year. Some bilateral/W3 projects are expected to be aligned in 2018 from the distinct participating centers

2.3 Collaboration and Integration

2.3.1 New Key External Partnerships:

The Platform will continue to develop new partnerships for the CGIAR to access and leverage new big data capabilities, including:

- cloud computing and storage (ongoing negotiation with Microsoft Azure and Amazon AWS),
- access and use of high resolution satellite information (Digital Globe),
- data analytical infrastructures (e.g. University of Minnesota Supercomputing Institute, IBM),
- machine learning service providers (Google, IBM, Microsoft),
- access and use of data from mobile phone networks for food systems research (the Digital Impact Alliance, Dalberg Data Insights, the GSM Association).

In addition, in 2018 the Big Data Platform is developing wide-ranging partnership with IBM that includes:

- Research collaborations on application of Artificial Intelligence in agriculture;
- Sharing of geospatial and weather data;
- Use of IBM data infrastructure to accelerate modeling and geospatial analyses;
- Potential collaboration on piloting blockchain technologies for supply chain traceability.

2.3.2 New Contributions Expected from the Platform

The Big Data Platform will accelerate OADM Policy compliance via improvements to the Global Agricultural Research Data Innovation and Acceleration Network (GARDIAN). GARDIAN will also spur collaborations across all Centers as well as other entities with relevant open platforms to facilitate sharing, discovery, and analysis/visualization of agricultural knowledge resources. New standards and services to enable semantic interoperability and reuse are envisioned, including an agronomy ontology (with Bioversity) that will underlie a field book to enable data harmonization at the collection phase (with Bioversity, CIP, IITA and others), a draft socioeconomic ontology, and efforts to harmonize term choice and submission across ontologies (with Bioversity and the University of Minnesota). An overview of licenses and data privacy/ethics protocols across CGIAR will help guide improvements in practices governing these, along with support, trainings, and how-to materials. In 2018 the Platform will also complete design and prototyping of a common analytics environment where CGIAR data can be encrypted if needed, securely stored, more easily assembled, and analyzed alongside other commonly used data for research purposes, providing some cloud storage and computation power if needed.

The Platform Communities of Practice (CoPs) will also be generating significant contributions in 2018:

- **Crop Modeling CoP** will conduct a CGIAR-wide gap analysis to determine strengths and weaknesses of each CG center in terms of modelling capacity, skills and expertise, and how modeling can support other research and outreach activities at the CG centers. As a result of this analysis, 4 reviews that document CGIAR modeling activities (in the areas of Policy/Socioeconomic, Environment, Crop Management and Breeding) will be submitted to a high-impact international peer-review journal for their publication. Furthermore, the CoP will continue to create and maintain working groups with the wider Crop Modelling community to promote more impactful use of data for agricultural development. Finally, they will try to fund the two remaining mini proposals of 2017 to support activities related to modelling that would achieve a demonstrable boost/impact through injecting a small amount of funding.
- **Data-driven Agronomy CoP** will address several key questions through webinars, blogs and newsletters regarding their main topic for 2018: validating DDA tools. This will lead to an annual report and form the basis of a project regarding the design of standardized protocols for validation of the data-driven tools. Another cross-institution workgroup will be mapping

gaps in technology access for the world's smallholder farmers by bringing subnational data sets (farm distributions, telecoms coverage, technology use, purchasing power) together and undertaking a geospatial analysis to look at major shortfalls in technology access and utilization. This will lead to a key baseline for policy and for private- and public-sector data-driven agronomy services. Finally, this CoP will also engage in research on the design of a minimum viable product for digital, on-farm decision support.

- **Geospatial CoP** will organize activities to provide technical training on the use of programming in R for geospatial analysis, provide two mini-grants for support developing key open geospatial datasets, and support community members' participation in the conferences for showcasing new geospatial data and analytical products. A new working group on UAV will be created to promote the best practices of using UAV-generated geospatial big data in CGIAR's research. The CoP will also coordinate to develop a new process to spur innovation in geospatial science contribution to the real-time mapping of crop distribution. Additionally, the CoP plans the following products for 2018: 1) ten flagship geospatial datasets fully documented, published, and made available through GARDIAN, 2) a position paper on the precision agriculture in smallholders' farming, jointly developed with the Data-Driven Agriculture CoP, and 3) GARDIAN's geospatial data cataloging and visualization to be tightly integrated with the backend system.

- **Livestock Data for Decisions CoP**, led by the University of Edinburgh, will begin to develop livestock specific data ontologies in collaboration with the other CoPs, and will work with the Platform to make more open livestock data discoverable to CGIAR researchers and the wider community.

- **Ontologies CoP** will continue to develop resources for the Big Data and other platforms. The following products from this CoP are expected in 2018: 1) A map of available expertise in the CoP and identification of additional expertise needed to better harness Big Data capabilities in agriculture, 2) Matrix of existing standard vocabularies /ontologies with guidelines/tutorials, mappings solutions between ontologies & current gaps for agriculture knowledge domains, 3) Definition of the key features of an ontology hub as a service to ontology-driven annotation tools and query systems, 4) Expert advice on demand to other CoPs and gap-filling where possible on other capacity building/training needs, and 5) Compilation of the Data-driven Agronomy CoP's perspectives on the Agronomy Ontology and the field book prototype. Finally, this CoP is preparing the 2018 iteration of the PhenoHarmonIS workshop to discuss harmonization of agronomic and plant phenotypic data across CGIAR and the wider community of researchers.

- **Socio-Economic Data CoP** aims to bring together CGIAR centers and external partners to tackle major issues related to socio-economic (survey) data and moving towards making the data FAIR. By doing so the SED CoP can enhance CGIAR impact and the use of CGIAR socio-economic data for partners in development. The key priorities for the SED CoP in 2018 are therefore the workgroups: 1) 100 questions, to identify key indicators already incorporated in many if not most surveys, which can be standardized at some level to make interoperability possible, 2) SociO, to create a socio-economic ontology of controlled vocabularies, classifications and concordances that will allow standardization of key indicators, and 3) Ontology-independent structural metadata schema. For truly interoperable and reusable data, the structural metadata should be machine-readable. Because socio-economic data are highly varied and contain all types of data, and often do not abide by fixed ontologies, any metadata schema must be highly flexible, extensible and ontology-independent.

2.3.3 New Cross-Platform Interactions:

The Gender Platform

In 2018 the Platform will pursue more active collaboration with the Gender Platform, including Gender review on the expanding Big Data Platform program as well as specific collaborations in research, communications, and innovation processes. In addition, in 2018 datasets made open through Module 1 activities will be gender-disaggregated wherever applicable; this will be achieved in close cooperation with the CGIAR Community of Practice on Gender. Data standards developed in Module 1 will also include appropriate metadata for representing the gender dimension.

The Genebank and Excellence in Breeding (EiB) Platforms

The Big Data Platform will assist the Genebank and EiB platforms with the development of data sources and intelligence around the agronomic and socioeconomic impact of breeding programs, and collaborate on designing a pilot for tracking Digital Object Identifiers from Genebanks, through breeding programs to their impact in food systems. In addition, the Big Data Platform will continue to develop synergistic investments in 2018 that include:

- Funding proposals to pilot improved traceability of seeds and to capture data on their on-farm performance and socioeconomic impact.
- Software development for ontology-based electronic field books for agronomic data collection that link to ontology-based breeding tools and mobile device apps already successfully tested or implemented by these tools.
- Ensuring discoverability of Genebank and EiB outputs by the Big Data Platform's GARDIAN.

2.3.4 Expected Efforts on Country Coordination

The Big Data in Agriculture Platform is rooted in global action, supporting all the CRPs in their impact pathways, and working across all Centers. Hence, the Big Data Platform itself is not making any particular geographies a priority, nor strongly engaging in country coordination efforts. However, should there be demand from CRPs for place-based work in priority countries, the Platform will ensure collaboration and integration with ongoing integration efforts.

2.4 Monitoring, Evaluation, and Learning

In 2018 Big Data will work with MARLO, CCAFS' interoperability standards-compatible planning and reporting platform adopted by a majority of the CRPs, to test the discoverability via the Platform's harvester of non-sensitive indicator reporting. This information will be available for mapping and visualization through the Platform's mapping tool to allow easy visualization of CGIAR's impact at scale. The Big Data Platform will also consult with CGIAR's Monitoring, Evaluation, and Learning Community of Practice to develop an impact assessment framework that demonstrates the value of the Platform's activities and allows for agile development and course correction.

3. Platform Management

3.1 Management of Risks to Your Platform

The Platform will keep an important focus on Module 1 activities in 2018. As a result, there are potential risks with making data open and accessible. For example, there could be a risk of failure of CGIAR centers to deliver on CGIAR Open Access Open Data Policy resulting in erosion of reputation

of the Platform to deliver. Centers will continue to receive funds in 2018 to move towards implementation of the center OA/OD strategy, and distribution of those funds are subject to having an implementation plan in place. Regular meetings of the Data Management Task Force will be held. Higher level management will be engaged to ensure the correct incentives are in place in centers.

Similarly, there is a risk of lack of high-level governance buy-in across the CGIAR to serious management of data and center-level investment in the correct incentives and structures to build data-related capacity; a key enabler of OA/OD compliance. Every 3 months, an email will be sent to DGs highlighting best practice across the centers to ensure this gets some high-level management bandwidth. Every year the SMB will also be updated on progress.

3.2 Platform Management and Governance

No changes in Management or Governance processes and bodies are expected from those defined in the approved Big Data Platform proposal. The Steering Committee will formalize the International Advisory Board in the first quarter of 2018.

TABLES

Table A: Planned Milestones

Module	2022 Platform outcomes (from proposal)	Milestone*	Budget by outcome		Assessment of risk to achievement** (L/M/H)	Means of verification
			W1/2	W3/ bilateral		
1. Organize	1.1. A demand- driven data analytics environment is available.	1.1.1. Prototype data analytics environment developed and tested. 1.1.2. Seamless integration with key data visualization and analytical tools enabled. 1.1.3. Case studies using the analytical environment developed and presented.	800,000		Risk: Technical difficulties to securely store/handle sensitive data; cultural difficulty in (re)training scientists to use new data analytics environment while dealing with privacy/ ethics effectively. Risk level: Low	Prototype backend tools and services leveraged for analytics and visualization. Sensitive data securely stored and accessed. At least 2 case studies developed using the data analytics environment.

| | 1.2. CGIAR resources are discovered and reused. | 1.2.1. GARDIAN (Global Agricultural Research Data Innovation and Acceleration Network) launched with tools for data cataloging and visualization and services to accommodate data privacy, ethics, and licensing issues.

1.2.2. Fifty datasets in institutional repository quality-checked, and both data and metadata annotated with ontology and/or AGROVOC/GACS terms.

1.2.3. Sustainable model for data annotation/ curation tested.

1.2.4. Data mining and machine learning methodologies tested for improving data quality or searchability. | 2,361,671 | | Risk 1: Technical and workflow issues related to the who and how of data annotation and curation.

Risk 1 level: Moderate

Risk 2: Privacy and ethics issues relating to farmer information in datasets.

Risk 2 level: Moderate

Risk 3: Immaturity –despite the promise—of data mining, machine learning protocols.

Risk 3 level: High | Curated CGIAR datasets (at least 500) are discoverable and reusable by humans and machines, including key legacy and high-value datasets. |

	1.3. Standards and semantics are utilized to enable FAIR (Findable, Accessible, Interoperable and Reusable) agricultural data.	1.3.1. CG Core Metadata Schema v.2.0 finalized and implemented, and/or mapped across Center publications and data repositories. 1.3.2. Agronomy Ontology completed and prototype field book tested for data collection based on semantic standards. 1.3.3. Ontology (Crop and Agronomy Ontology at minimum) and/or AGROVOC/GACS terms adopted as metadata descriptors by Center repositories/ platforms. 1.3.4. Draft for key classes and sub-classes for socioeconomic ontology developed. 1.3.5. Links to agrisemantics efforts external to CGIAR maintained or enhanced.	1,931,671	Risk: CGIAR centers do not prioritize FAIR data or the implementation of standards in data management and sharing. Risk level: Moderate	Metadata, data and publications that score high (over 4/5) on the DANS FAIR scale (in GARDIAN), particularly for interoperability and reusability. A mature agronomy ontology and field book that produces semantically enabled, standards-compliant metadata and data. Collaborations created/ maintained to build on, leverage funds, and avoid duplication in agrisemantics.

	1.4. Enhance capacity, catalyze culture change to further CGIAR OA/OD compliance and public goods mandate.	1.4.1. Materials and webinars developed and/or shared with appropriate CGIAR communities on licensing resources, data privacy and ethics, and data management and standards for maximizing FAIRness (Findability, Accessibility, Interoperability and Reuse) of CGIAR resources. 1.4.2. At least 2 data sprints or 1 relevant training and 1 data sprint held for researchers to increase the number of well-annotated datasets in Center data repositories. 1.4.3. At least two workshops/ trainings for data/ information/ ontology managers and researchers held on ways to render datasets FAIR.	649,170	Risk: CGIAR centers do not apply learnings from capacity building efforts, but continue along a "business as usual" trajectory. Risk level: Moderate	Guides on dealing with data to achieve Big Data capabilities, including: licensing, privacy/ ethics, and best management. OA/OD Support Pack Version 2.0, including: best practices, training and reference docs, standards.
2. Convene	2.1. CGIAR is more broadly engaged in big data community.	2.1.1. Communities of Practice around geospatial data, socioeconomic data, ontologies, data-driven agronomy, livestock data, and crop modelling establish CoP networks across CGIAR and produce outputs addressing key constraints in data and analytics.	600,000	Risk: Failure of leadership to build a functioning team across the Centers. Risk level: Low.	Products from Communities of Practice (shareable, harmonized data across disciplines, position papers, small pilots).

		2.1.2. Hold high-level Convention on Big Data in Agriculture, with wide participation of CGIAR and non-CGIAR actors, establishment of collaborative agreements.				Convention, convention reports and communications materials.
						At least two new high priority data products.
	2.2. CGIAR increases its capacity to work on priority topics more quickly, more effectively and at greater scale.	2.2.1. Identify and produce high priority, high impact new data products and develop methodological plan to produce them with initial implementation.	400,000			
	2.3. CGIAR develops as a learning organization.	2.3.1. Map out CGIAR needs for common big data related computing and storage infrastructure.	1,000,000		Risk: Diverse requirements and capacities across the organization	Survey of CGIAR big data storage, data products or services, computation, and capacity building. At least two high-value shared services investments that

		2.3.2 Establish shared services for CGIAR by negotiating with external data utility partners.			make it difficult to prioritize storage and computation investments.	expand our capacity to work with large datasets.
		2.3.3 Develop capacity building activities linked to Centers' needs.		Risk level: Moderate	At least two workshops (online and in person) for to build CGIAR big data capacity.
		2.3.4 Build the capacity of CGIAR to meet the data needs of the agriculture development sector.			"State of the Network"report of CGIAR digital transformation and key recommendations.

| 3. Inspire | 3.1 CGIAR shows how data-driven approaches yield results in poverty reduction, enhanced nutrition or environmental benefits. | 3.1.1. New Pilot Inspire projects around Big Data related innovations.

3.1.2. Scale-up one successful pilot Inspire project (from winners 2017) around Big Data related innovations.

3.1.3. Synthesis of Inspire project successes and failures, policy documents, best-practice guidance. | 800,000 | | Risk: Targeting of innovations does not adequately address bottleneck issues for the sector.

Risk level: Moderate. | Completed selection process and awards for 5 new Inspire projects.

Completed evaluation of impact and selection process for one scale-up grant.

Synthesis report examining the Inspire Challenge in light of digital innovation strategy. |

* Milestones include outputs, output use and outcomes along the impact pathways as appropriate to the scale and maturity of the work. In this table A, please focus as much as possible on significant milestones towards outcomes which can be justified the completion at reporting.

**Please list the major risks focusing more on technical or geographic considerations that may hinder the expected delivery of results by the Platform.

Table B: Cross-cutting Aspect of Expected Outputs

Cross-cutting	Number (%) scored 2 (Principal)	Number (%) scored 1 (significant)	Number (%) scored 0	Total overall number of outputs
Gender	5%	5%	90%	17
Youth			100%	
CapDev	23%		77%	

Table D: Platform Planned Budget

	Planned budget 2018			Comments on major changes
	W1/2	W3/bilateral	Total	
Module 1	5,742,513		5,742,513	W1/2 figure includes carry-over from 2017
Module 2	2,000,020		2,000,020	W1/2 figure includes carry-over from 2017
Module 3	1,282,866		1,282,866	W1/2 figure includes carry-over from 2017
Strategic Competitive Research grant	0			
Platform Management & Support Cost	269,325		269,325	We aim to stay within budget on Management and Support costs, but do note significant workload and demand for the Secretariat.
Platform Total	**9,294,724**		**9,294,724**	The identified Funding Sources are primarily W1+2. The Platform is still a relatively a new initiative and a large portion of the funding is contributing to organizational strengthening, we expect some bilaterals to be aligned in 2018 and for this to grow in 2019 and beyond.

Table F: List of New Key External Partnerships

Module	Stage of research*	Name of partner	Partner type*	Main area of partnership*
1,2,3		IBM Research	Private Sector	Collaborative research into digital interventions in food systems.
1,2,3		Digital Impact Alliance	Non profit	Liaise with the mobile sector on behalf of the agriculture development sector, source mobile network data and collaborate on innovative data analyses
2		Digital Globe	Private sector	Provide access to the global archive of high resolution satellite imagery and cloud computing tools to accelerate geospatial analysis
1,2		Amazon Web Services	Private sector	Cloud storage and computing, collaboration on open geospatial research.
1		Minnesota Super Computing Institute	Public sector	Collaborative research on data ontologies, secure data sharing, and data integration for food systems.

* See instructions in the common results indicators manual (available early 2018).

Table G: New Internal (CGIAR) Collaborations between the Platform and Programs and among Platforms

Name of CRP or Platform	Brief description of collaboration (give and take between Platform and CRPs) and value added*	Relevant Module
Genebanks and Excellence in Breeding Platforms Gender Platform	Collaboration on tracing Digital Object Identifiers assigned to genetic materials to enable deeper understanding of impact. Ontology include genebank morpho-taxonomic descriptors and germplasm management relevant concepts. Co-design of an innovation process or strategic research to use data in novel ways to reveal gender in food systems.	1 3

*E.g. scientific or efficiency benefits

Table H: Planned Monitoring, Evaluation, and Learning Exercises

Planned studies/learning exercises in 2018	Comments
Awardees of Inspire Challenge grants will be reporting against project specific impact metrics, which will inform and help refine Platform learnings about digital innovation strategies for agriculture development. The Platform data harvester will be connected to the MARLO results based management / M&E platform to be able to monitor progress towards data sharing and enable new analyses of potential impact.	

Table I: Main Risks to Performance and Planned Mitigation

	Key area of risks	Mitigation strategy
STRATEGIC RISKS: risks voluntarily accepted to achieve the BDP strategy	Strategy doesn't address the salient research issues relevant to the mission.	Strategy reflections once per year at the time of annual reporting and generation of POWB.
	Failure of CGIAR centers to deliver on CGIAR Open Access Open Data Policy resulting in erosion of reputation of the Platform to deliver.	Each center will receive funds to move towards implementation of the center OA/OD strategy, and distribution of those funds are subject to having an implementation plan in place. Regular meetings of the Data Management Task Force will be held. Higher level management will be engaged to ensure the correct incentives are in place in centers.
	Lack of high-level governance buy-in across the CGIAR to serious management of data and center-level investment in the correct incentives and structures to build data-related capacity.	A key enabler of OA/OD compliance. Every 3 months, an email will be sent to DGs highlighting best practice across the centers to ensure this gets some high level management bandwidth. Every year the SMB will also be updated on progress.
	Ethical failures leading to erosion of reputation, or at worst, legal objection due to improper treatment of privacy issues in datasets made available.	Best practice guidelines developed and shared with all centers on adequate treatment of privacy in research data; Facilitation of discussion around ethical issues related to data; Active engagement and 25% contracting of time of SMO Legal advisor expert on ethical considerations in data. Thought leadership on the topic established.
	Mismatch of skills with business needs (including on outcome orientation).	Building skilled teams at the Secretariat, and generating a dynamic and highly skilled network of skilled staff across the centers, permitting flexible tapping into relevant expertise when needed. Strong partnership models.
	Significant over or under expenditures.	Quarterly reviews of expenditure and communication with Center focal points to enable timely management of expenditures. Linking expenditure planning to performance reviews.

	Centers not aligning bilateral projects to the Platform.	Put in place incentives to get Centers to allocate resources to BDP and increase fund-raising efforts. Controls need to be put in place.
	Loss of funds due to research partners' failure to account (technically or financially) for the resources allocated to them by Center.	Annual reports checked each year and feedback given.
EXTERNAL RISKS: risks arising from events outside the Platform that are beyond BDP's control and/or influence.	Big Data goes out of fashion, or becomes a meaningless buzzword with negative reputation.	Constant monitoring of the external environment; Maintenance of flexible framing and direction of the Platform to adjust to external environment.
	Major privacy issues of farmer-based data create controversial environment for working on Big Data related efforts.	Constant monitoring of external environment and proactive communications and maximum transparency to protect reputation of Platform; Put in place controls internally to ensure privacy issues respected. Engagement of subject matter experts and intellectual property lawyers to develop guidelines for responsible data management and use.
	Funding stability year-to-year negatively impacting Platform operations and activities.	Concerted efforts with donors and potential investors to: 1. Protect W1 investment by ensuring SMO, SMB and W1 donors see value of the Platform to delivering on the CGIAR SRF and contributing to innovation within the CGIAR; 2. Generate some W2 investment to diversify funding sources, providing donors with a window to support specific components of the Platform; 3. Leverage the Platform partners to bring value addition to the platform through in kind support and provision of data and services at discounted rates; 4. Support centers in the raising of bilateral funds to align to the Platform to diversify risks.

INTERNAL RISKS: risks arising within the Platform from staff's unauthorized/undesirable behavior or actions breaking down routine operational processes that do not generate strategic benefits.	Staff injury or death due to work in difficult locations e.g. - Kidnappings - Staff caught up in political or social unrest / crime waves - Staff exposed to dangerous conditions and diseases while traveling - Terrorism	Work with partner security systems including the UN and national government systems. Staff relocation when necessary. Not organizing meetings in countries with political instability.
	Stress related loss of productivity in core Platform team due to excess work demands.	Platform Management Team reiterates need for achieving good work-life balance; Contracting of internal and external expertise as needs arise to balance workload; Increase the base staffing level of the "Secretariat" to manage high workload.
	Failure of leadership to build a functioning team across the Centers.	Maintain good communications with centers for team building.

*Please refer to CGIAR Risk Management Guidelines.

Table J: Platform Specific Indicators or Targets

Module	2022 Platform outcomes (from proposal)	Immediate Outcomes (= Milestones)	Targeted products for 2018
1. Organize	1.1. A demand-driven analytics environment is available.	1.1.1. Prototype data analytics environment developed and tested. 1.1.2. Seamless integration with key data visualization and analytical tools enabled. 1.1.3. Case studies using the data analytics environment developed and presented.	Prototype data platform with at least two backend tools and services integrated to streamline data analysis and visualization. Sensitive data securely stored and accessed. At least two case studies developed using the analytics environment and presented at the Convention.
	1.2. CGIAR resources are discoverable and reused.	1.2.1. GARDIAN (Global Agricultural Research Data Innovation Acceleration Network) launched with tools for data cataloging and visualization and services to accommodate data privacy, ethics, and licensing issues.	Curated CGIAR datasets (at least 500) are discoverable and reusable by humans and machines, including key legacy and high-value datasets.

		1.2.2. Fifty datasets in institutional repository quality-checked, and both data and metadata annotated with ontology and/or AGROVOC/GACS terms. 1.2.3. Sustainable model for data annotation/ curation tested. 1.2.4. Data mining and machine learning methodologies tested for improving data quality or searchability.	
	1.3. Standards and semantics are utilized to enable FAIR (Findable, Accessible, Interoperable and Reusable) agricultural data.	1.3.1. CG Core Metadata Schema v.2.0 finalized and implemented, and/or mapped across Center publications and data repositories. 1.3.2. Agronomy Ontology completed and prototype field book tested for data collection based on semantic standards. 1.3.3. Ontology (Crop and Agronomy Ontology at minimum) and/or AGROVOC/GACS terms adopted as metadata descriptors by Center repositories/ platforms.	Metadata, data and publications that score high (over 4/5) on the DANS FAIR scale (in GARDIAN), particularly for interoperability and reusability. A mature agronomy ontology and field book that produces semantically enabled, standards-compliant metadata and data. Collaborations created/ maintained to build on, leverage funds, and avoid duplication in agrisemantics.

		1.3.4. Draft for key classes and sub-classes for socioeconomic ontology developed. 1.3.5. Links to agrisemantics efforts external to CGIAR maintained or enhanced.	
	1.4. Enhance capacity, catalyze culture change to further CGIAR OA/OD compliance and public goods mandate.	1.4.1. Materials and webinars developed and/or shared with appropriate CGIAR communities on licensing resources, data privacy and ethics, and data management and standards for maximizing FAIRness (Findability, Accessibility, Interoperability and Reuse) of CGIAR resources. 1.4.2. At least 2 data sprints or 1 relevant training and 1 data sprint held for researchers to increase the number of well-annotated datasets in Center data repositories. 1.4.3. At least two workshops/trainings for data/ information/ ontology managers and researchers held on ways to render datasets FAIR.	Guides on dealing with data to achieve Big Data capabilities, including: licensing, privacy/ ethics, and best management. OA/OD Support Pack Version 2.0, including: best practices, training and reference docs, standards.

2. Convene	2.1. CGIAR is more broadly engaged in big data community.	2.1.1. Communities of Practice around topics of geospatial data, socioeconomic data, ontologies, data-driven agronomy, livestock data for development and crop modelling produce outputs addressing key constraints of the sector and establish CoP networks.	Products from Communities of Practice (e.g., shareable, harmonized data across disciplines, position papers, small pilots).
		2.1.2. Hold high-level Convention on Big Data in Agriculture, with wide participation of CGIAR and non-CGIAR actors, establishment of collaborative agreements.	Convention, convention reports and communications materials.
	2.2. CGIAR increases its capacity to work on priority topics more quickly, more effectively and at greater scale.	2.2.1. Identify high priority, high impact new data products and develop methodological plan to produce them,with initial implementation.	At least two new high priority data products.
	2.3. CGIAR develops as a learning organization.	2.3.1. Map out CGIAR needs for common big data related computing and storage infrastructure.	Survey of CGIAR big data storage, data products or services, computation, and capacity building.
		2.3.2 Establish shared services for CGIAR by negotiating with external data utility partners.	

		2.3.3 Develop capacity building activities linked to Centers' needs.	At least two high-value shared services investments that expand our capacity to work with large datasets.
		2.3.4 Build the capacity of CGIAR to meet the data needs of the agriculture development sector.	At least two workshops (online and in person) for to build CGIAR big data capacity. "State of the Network"report of CGIAR digital transformation and key recommendations.
3. Inspire	3.1 CGIAR shows how data-driven approaches yield results in poverty reduction, enhanced nutrition or environmental benefits.	3.1.1. New Pilot Inspire projects around Big Data related innovations.	Completed selection process and awards for 5 new Inspire projects.
		3.1.2. Scale-up one successful pilot Inspire project (from winners 2017) around Big Data related innovations.	Completed evaluation of impact and selection process for one scale-up grant.
		3.1.3. Synthesis of Inspire project successes and failures, policy documents, best-practice guidance.	Synthesis report examining the Inspire Challenge in light of digital innovation strategy.