



CABI POLICIES & PROCEDURES

CABI Policy for Data Management and Governance

Implementation of Responsible Data Practice and FAIR Principles

June 2026

KNOWLEDGE FOR LIFE



A. Introduction

With the proliferation of AI and big data over recent years, good organizational data management is becoming increasingly important. It is imperative that CABI as an organization is responsive to emerging understanding of how data should be effectively managed. Therefore, we are developing a CABI-wide implementation plan to better organize our data pipelines. In doing so we create more opportunities for us and our member countries to benefit from responsible and FAIR (Findable, Accessible, Interoperable and Reusable) data management, using data assets more effectively to build sustainable new services, applications and decision support tools. FAIR data equally supports CABI and partner aspirations to both open and to commercialize assets. A key impetus for developing a new Data Implementation Framework and this accompanying Policy, has been CABI's interest in the development and application of AI models, both in-house, and with external partners. For this, better data management that yields "AI-ready" datasets, is a prerequisite.

CABI, since its inception in 1910, has undertaken vast amounts of primary research, and consequently, today, is a data-rich organization. The new framework that is being formulated, specifies the economic, commercial, organizational, and environmental benefits for CABI to improve its data ownership and management. CABI envisions a future where we are leaders in the production and utilization of research data that is FAIR. This implementation of FAIR data principles within CABI will exhibit its aspiration to share knowledge as an organization, aligned to the goals laid out in its medium-term and science strategies.

While the CABI Data Implementation Framework has enunciated the details of CABI's data vision and mission, and key considerations around why better data management needs to be implemented, the CABI Policy for Data Management and Governance (hereinafter referred to as "Policy") aims to give effect to some of those ideas. For this, the Policy stipulates key principles for CABI's internal data management and complements these with an implementation plan. Both these documents draw upon findings from the internal Scientific Review and seek to align with CABI's Strategy.

It is in this context that this Policy aims to establish some baselines in terms of good data management practices within CABI at an organizational scale.

B. Scope

This Policy, along with CABI's Data Implementation Framework, aims to streamline the collection, reuse, storage and retention, and safety and processing of primary and novel data gathered during projects. This policy shall apply to CABI staff, and any third-party subcontractors who are responsible for collecting or managing:

- a. Science and research data; and
- b. Digital and data assets produced under CABI projects.

This Policy supersedes the previous Knowledge Management policy for CABI.

C. Guiding principles for CABI's internal data management and governance

- a. **Alignment with FAIR data principles** – FAIR data principles aim to make data assets findable, accessible, interoperable, and reusable. The previous and now retired CABI Knowledge Management policy focused on “Open Data” access, a paradigm that has undergone considerable shift in favour of FAIR data principles as a more viable and actionable framework for data management. CABI has been a proponent of FAIR data with external stakeholders, given their general recognition as a “gold standard” for good data management and practice(s). The overall objective of FAIR principles is to improve availability of robust data (and metadata) in a scalable and sustainable manner. Much data can benefit from being Open Data, while all data can benefit from being FAIR Data. Given CABI's data-rich status, it is vital to ensure data generated in different teams and practice areas at CABI is not existing in silos but is increasingly FAIR. Tools such as a Data Management and Access Plan (DMAP) can aid in delivering this alignment.
- b. **Restatement of commitment to Open Access Publication** – CABI as stated in its Science Strategy (2026-2028) is committed to open access publishing for all publications arising from projects, where CABI staff are the lead or corresponding author. Article processing charges (APCs) will be costed into projects where permitted by the funder. Where costs are a consideration self-archiving of pre-prints and/or author-accepted manuscripts is expected. Publication in journals who do not accept self-archiving should be avoided. CABI will also continue to encourage the use of open access for other co-authored publications. CABI will continue to review its older scientific publications (legacy material) and identify important and relevant material to digitize to become open access.
- c. **Ensuring responsible data practices** – In addition to FAIR principles to improve access and sustainability, it is imperative for CABI to also ensure that any personal or sensitive data it collects, collates, processes, or shares, is done so in line with responsible data practices and the data protection regulations of the countries in which the data is collected and processed. The notion of privacy-by-design is useful in safeguarding individual privacy and further establishing organizational and technical controls to prevent data breaches and/or unauthorized processing of personal and sensitive data. Data retention periods should align with funder requirements, legal obligations and CABI retention policies.
- d. **Controlling risk** – governance controls put in place shall be proportionate to the sensitivity, value and risk associated with the type of data being handled.
- e. **Interface with other intra-CABI policies** – CABI is also developing (or has existing) policies that will interact with this Policy and the Data Implementation framework. This Policy must be implemented alongside existing and proposed policies, strategy and implementation plans (such as the Science Strategy, the AI Policy).
- f. **Ensuring AI-readiness for CABI data** – Given a high priority for CABI is to bolster the development and use of responsible, safe, and ethical AI systems to enhance digital agriculture, it is imperative that CABI's large data and content resources be made “AI-ready”. This will include developing the necessary technical infrastructure for collating, storing and sharing this content; creating internal capacity at CABI to ensure data access in alignment with the FAIR principles (also discussed in principle a. above), responsible practices to ensure appropriate transparency, traceability and human oversight of datasets; and setting out clear processes and protocols to enable the transformation of CABI data and content into becoming AI-usable.
- g. **Internal capacity building** – To ensure better data management and practices across CABI, internal capacity building is crucial. This should target training and sensitization, and in parallel, creating roles for data stewards, data officers, project managers and other similar roles, and where possible using project funding to do so. Actionable plans, checklists and guides will be used to drive tangible change.

- h. **Periodic review of data policy and implementation framework** – This Policy and the accompanying Data Implementation Framework need to be reviewed annually by the Data Taskforce, in consultation with the EMT, regional directors and other relevant stakeholders, to ensure they present a dynamic and living data management and governance framework.

D. Data Policy

- a. Making CABI data FAIR – Data assets covered by the Data Implementation Framework and this Policy, shall be managed in accordance with the FAIR data principles. A pre-requirement for 'data-rich' projects is to formulate a Data Management and Access Plan (DMAP) and periodically review and evolve it through the project lifecycle, to attain FAIR data assets. Data-rich projects are defined under section F.
- i. **Findability**
- The CABI data repository or suitable alternative repository should be utilized by project managers (and relevant data stewards within a project team) to collate, organize, annotate, and inventory datasets, within data-rich projects. Datasets should be well-described with metadata to improve retrievability.
 - No project data (defined here as data generated by the project) shall be permanently stored outside an appropriate repository (on external hard drives, SharePoint, etc), and all significant project data assets shall, as a minimum, have a metadata record in the CABI data repository.
- ii. **Accessibility**
- Improve accessibility of project data by compiling and storing it in a central space, which this Policy indicates should be the CABI data repository or suitable alternative with a metadata record on the CABI data repository. Metadata records should be easily accessible even when the data may be less so (because of privacy, commercial concerns, etc.).
 - For Bioinformatics, a separate storage system should be determined by the Data Taskforce, in consultation with EMT.
- iii. **Interoperability**
- Relevant data assets (e.g. datasets) and metadata shall be interoperable and fit for reuse by using open and widely used file formats, rich metadata, and controlled vocabularies.
 - DMAPs (Data Management and Access Plans) shall be developed at project formulation in such a way as to ensure that data assets conform to accepted interoperability standards (e.g. specified use of ontologies, controlled vocabularies).
 - CABI will maintain a list of core sets of standards, ontologies and vocabularies to be used in CABI projects.
- iv. **Reusability**
- DMAPs must set clear plans of action and budgetary requirements to ensure long-term reusability and sustainability of data assets created in a project.
 - Data stewards (or project managers, if a project does not have a dedicated data steward role defined) shall be responsible for formulating and implementing reusability strategies.
 - Data assets should be labelled with metadata describing clear and accessible licences and provenance.
- b. Responsible and ethical data practices: CABI's data management and governance must ensure that data-rich projects establish clear steps to process and store data in a responsible and ethical manner. As with FAIR related milestones, the DMAP shall contain milestones and measures that will be implemented to ensure data safety, privacy-by-design, and ethical data practices.
- i. **Data Safety**
- Data must be collated, compiled and stored in a manner that is secure.
 - Technical infrastructure (such as the CABI data repository) must deploy well-tested and robust data protection and security measures.
 - Data breaches must be identified and tackled promptly by the IT team, in consultation with the Data Taskforce and where relevant the CABI Ethics Review Board.
 - Personal and sensitive data will be stored and processed only in strict compliance with applicable local laws and regulations (such as data protection, cyber security, or AI safety laws), as well as internal CABI policies around data transfer and retention.

ii. **Privacy-by-design**

- Data-rich projects shall focus on developing and managing data assets in a manner that integrates privacy-by-design. This is primarily for the collection, storage, and processing of personal and sensitive personal information.
- Privacy-by-design will require appropriate consent requirements, preservation of autonomy of data subjects (including without limitation their right to withdraw consent during research), and compliance with applicable data protection and privacy laws and regulations, such as the UK GDPR, the laws of countries where we operate, including those of member countries, as well as with internal CABI policies around data protection.
- Data stewards of data-rich projects (or in their absence, project managers) must submit an annual report to the Data Taskforce for each year of the project cycle, specifying what personal or sensitive personal data was collected, stored, processed in that period, and the measures that were implemented to comply with this Policy.

iii. **Ethical data practices**

- Data-rich projects should collect personal and sensitive data only to meet research objectives after prior informed consent to participate is received. Data retention and access policy should be explicit and limited. Pseudonymization is prescribed.
- Data collectors and users must sign non-disclosure and data protection agreements which should continue to survive even beyond their employment or tenure within such project(s).

c. Making CABI's data AI-usable and AI-ready:

- The policy seeks to create an enabling environment where data is of sufficient quality for onward use with AI applications. AI requires well curated interoperable data that carries a clear licence that permits use in AI contexts.

E. **Roles and responsibilities**

- a. **CABI Data Taskforce** – The CABI Data Taskforce shall continue to oversee the implementation of this Policy, as well as undertake periodic review of the same, ensuring there is at least one such review annually. The Data Taskforce is also responsible for putting in place protocols, processes and any additional frameworks to ensure compliance and specifically:
- i. That CABI's data is aligned with FAIR principles;
 - ii. That CABI maintains lists of agreed standards, repositories, metadata models and access rules
 - iii. That this Policy interacts harmoniously with other relevant CABI policies, and;
 - iv. Working with IT and Publishing, Data Stewards and other relevant CABI departments, to improve and upgrade CABI's data and content, to make relevant data AI-ready and AI-usable within a specific time frame, as determined by it.

The Data Taskforce will work in tandem with the EMT which will be responsible for the overall strategic planning and direction of CABI's data vision and aspirations. Proposed exceptions to the policy should be put to the Data Taskforce for review.

- b. **People Team** – The People Team will be responsible for ensuring that standard induction materials are prepared, in consultation with the Data Taskforce, to create a baseline of common basic understanding and knowledge around FAIR and responsible data practices, especially for non-technical staff members. This capacity building must in parallel be supported by appointing data specialists, as determined by the EMT, in consultation with the Data Taskforce. More specialist training will be led by specific teams, for example those working on digital sequence information will need to know their data responsibilities under the Nagoya Protocol.

i. **Recruiting**

1. Data stewards shall be appointed where possible in a permanent capacity to ensure proper management of CABI's data assets.
2. Project-specific data specialists can be recruited for fixed terms. These resources shall be identified and budgeted within the PFA and the DMAP for data-rich projects.
3. Ensuring proper provisions for data governance and capacity building are incorporated into CABI flagship programmes (PlantwisePlus) and large data-rich projects (Global Burden of Crop Loss, etc.)

ii. Baseline training and understanding

1. CABI staff shall be required to undergo periodic training and workshop sessions focused on relevant data issues. The frequency and content of such training shall be determined by the Data Taskforce in collaboration with the People Team.
2. CABI staff shall be incentivized and encouraged to undertake workshops, online courses, seminars and other modes of virtual or in-person training around relevant data issues. CABI EMT shall endeavour to allocate adequate funds to allow for at least one such training per year for every science, project and technical assistance-focused staff member.

F. Key Definitions

- a. **Artificial Intelligence** – (AI) means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.
- b. **AI Ready** – data prepared to be used to train or run an AI model for a specific purpose. Data which is appropriate to a use case, including patterns, errors, outliers, and unexpected occurrences. Data which is AI ready should carry an appropriate licence such as one based on the Model Content Licence for AI (CABI. Model content license agreement for AI – special terms and general terms. Gates Open Res 2026, 10:8 (<https://doi.org/10.21955/gatesopenres.1117269.1>). The Framework for AI Ready Data developed by the ODI and used by the UK Government usefully describes what it means to make data AI ready in practice (see https://theodi.cdn.ngo/media/documents/A_framework_for_AI-ready_data.pdf).
- c. **Data assets** – including but not limited to peer-reviewed journal articles; non-reviewed articles, reports, briefs, extension and training and other materials; books and book chapters; data, datasets and databases, digital sequence information; data collection and analysis tools (including models and survey tools); video, audio and images; computer software (including models, code, APIs); web tools (e.g. data portals, dashboards), and the metadata used to describe other data assets.
- d. **Data-rich project** – Those that generate new data, datasets, or research outputs (e.g., survey results, genetic data, sensor data, or experimental data) that could be valuable to future research, development, AI tools, or policymaking. Or projects that manage significant volumes of secondary data for analysis. Or projects with anticipated collaborations that require data sharing, where data standards or protocols must be used. Or where there are privacy, confidentiality, or Intellectual property concerns that need to be addressed for sensitive data. Data-rich projects typically signal the need for a Data Management and Access Plan (DMAP).
- e. **Data Management and Access Plan** (DMAP) – A detailed outline of the operational activities required to treat data as a valuable asset throughout its entire lifecycle, including publishing and sharing data using processes that promote openness and accessibility.
- f. **FAIR Data Principles** – FAIR is an acronym referring to making data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. FAIR principles seek to improve overall data access and long-term sustainability of data assets. See <https://fairprocessframework.org/> for more information.
- g. **Generative AI** (Gen AI) – A subfield of AI that can create new data, such as text, images or code. GenAI models are trained on large amounts of existing data and can be used for a variety of purposes, such as generating creative text formats, translating languages or creating realistic images.
- h. **Personal data** – any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. See <https://ico.org.uk/> for more information.