Data manangement plan - IBISBA 1.0 (Version 3)

Michael O'Donohue

IBISBA 1.0

Industrial Biotechnology Innovation and Synthetic Biology Accelerator

---

Deliverable D8.3
IBISBA 1.0 Data Management Plan

Planned delivery date (as in DoA): M6

Actual submission date: 01/06/2018, month M7

Revision date: 01/12/2020 (Version 2.0), 31/05/2022 (Version 3.0)

Workpackage: WP8

Workpackage leader: INRAE

Deliverable leader: INRAE

Version: 3.0 (final version)

Dissemination Level	
PU Public	
CI Classified, as referred to Commission Decision 2001/844/EC	
CO Confidential, only for members of the consortium (including the Commission Services)	**X**

Table of contents

Executive summary

This deliverable addresses the request by the European Commission to establish a data management plan. It is intended that this plan will evolve over the project duration. This present document constitutes the 2nd version of the data management plan, which provides overall guidelines and a basis for further evolution before the end of the project. In particular, it covers the question of data identification, data description, standards and metadata, data sharing and data archiving. In the IBISBA 1.0 project, a complete workpackage is dedicated to the development of e-tools for data and knowledge management and the project is operating a data management hub in cooperation with the Fairdom initiative, a leading European association that promotes the principles of data FAIRization (i.e. making data Findable Accessible Interoperable Re-usable).

1   Introduction

1.1.   *Open Research Data Pilot*

In the framework of Horizon 2020 the European Commission is running a flexible pilot called the Open Research Data Pilot (ORD pilot). The ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects and considers the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.

As a Horizon 2020 project, IBISBA 1.0 aims to be fully compliant with the aims of the ORD. For this reason, the project's beneficiaries have defined an evolutive Data Management Plan that is described herein.

1.2   *IBISBA a distributed European research infrastructure*

IBISBA is an Engineering Biology Research infrastructure that benefits from ESFRI recognition. Presently, IBISBA is classed on the ESFRI roadmap (since 2018) as an infrastructure project.

Currently IBISBA is in the project phase and is intended to be in full operation before 2028. To support the growth and implementation of IBISBA, funding is being received from IBISBA stakeholders and in the framework of two Horizon 2020 projects, IBISBA 1.0 and PREP-IBISBA, which carry the grant agreements N°s 730976 and 871118 respectively.

IBISBA supports the growth of Industrial Biotechnology, combining biotechnological processes with chemical processes, translating early scientific results into prototypes, innovation and, ultimately, industrially workable solutions. IBISBA is unique in that it brings together, within a coordinated network infrastructure, services that cover the different steps in R&D project pipelines, from bioprocess conceptualization and design all the way through to pilot phase testing.

Within the scope of IBISBA, its primary stakeholders are building several shared functions that aim to materialise IBISBA services:

● A one-stop shop – IBISBA offers users a single access to best-in-class infrastructure across Europe, through which finding a comprehensive research infrastructure service offer no longer requires endless, multi-contact searches. IBISBA can work with users to identify the right combination of services to suit their R&D needs.

● Harmonized practices and delivery of end-to-end bioprocess development – IBISBA founders are working hard to harmonize practices within the different infrastructures that contribute services. The aim is to deliver seamless project R&D pipelines that are accelerated by greater synergy and common practices, including shared standardized procedures.

● Europe-wide access to infrastructure – IBISBA aims to provide access to researchers across the private and public sectors in all member states and target third countries, thus reducing regional and national inequalities and helping innovators to get access when they most need it.

● Digital tools to support R&D projects – IBISBA provides users with advanced capabilities to collect and manage data and project-related knowledge assets in a secure digital environment, while ensuring compliance with FAIR criteria.

1.3   How is this data management plan related to IBISBA's overall data management strategy?

As a distributed European research infrastructure, IBISBA requires an overarching data management plan (DMP) that will be 'greater than the sum of its parts'. However, although an overarching DMP doesn't currently exist, preliminary guidelines are laid out in the first versions of the DMPs of the IBISBA 1.0 and PREP-IBISBA projects. This current version of the DMP for IBISBA 1.0 is thus a revised version of the deliverable 8.3 and a part of the overarching DMP that will be prepared and made public in 2022.

1.4   The specific purpose of this DMP

The IBISBA 1.0 project is structured into several workpackages with activities that address (i) networking (ii) communication (iii) research (iv) transnational access and (v) management. Networking and communication activities will collect knowledge and generate community standards, including those that can be used to organize and store data, and principles to guide IPR handling. The research and transnational access activities will produce experimental data and tools to structure, manage and archive data, while the management activity will provide the framework for the implementation of good practices, additionally producing project-related documents that constitute sources of data and knowledge.

Importantly, in the framework of IBISBA 1.0, a research and development action (WP7) is developing digital tools for IBISBA. Regarding IBISBA's ambitions to build in-house data management capacity, the project consortium includes scientists who are members of the FAIRDOM initiative (fair-dom.org). FAIRDOM is a non-profit association that grew out of a joint action of ERA-Net project ERASysAPP and the European Research Infrastructure ISBE, its aim being to establish a data and model management service facility for Systems Biology. Presently, FAIRDOM has achieved this aim, and now provides the tools necessary to make systems biology-related data, operating procedures and models, Findable,

Accessible, Interoperable and Reusable (FAIR). In IBISBA 1.0, the tools offered by FAIRDOM, particularly the SEEK platform, have been used to build a project-specific data management hub, designated IBISBAHub and hosted at the address https://hub.ibisba.eu/. This hub is an instance of FAIRDOMHub tailored for workflows. Accordingly, most elements of the data management plan are based on the capabilities offered by IBISBAHub, which like FAIRDOMHub, is fully in conformity with European standards and best practices. As repertoire for IBISBA 1.0 project assets and as an instance of FAIRDOMHub, IBISBAHub is linked to a variety of external resources, including OpenAire.

2 Glossary

Hereafter is a list of terms and a proposal for their definition. Some terms (in bold characters) are taken from the experimental IBISBA termbase available here:

Archiving - is the long-term preservation of data. Achieving this implies that the data is accompanied by metadata, formatted using a stable (standardized) format and held on a storage medium that will be readable in 20-30 years or more.

Asset – Any (tangible or intangible) output of the action such as data, knowledge or information — whatever its form or nature, whether it can be protected or not — that is generated in the action, as well as any rights attached to it, including intellectual property rights.

Asset embargo – this is a fixed time period throughout which data remains concealed and thus only available to its owner or the project consortium. During this period Data is not FAIR.

Data – according to the Merriam-Webster dictionary this word refers to "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation". Moreover, the Merriam-Webster dictionary goes on to say that data is "information in digital form that can be transmitted or processed".

Data accessibility – *related to Accessibility in the IBISBA termbase -* **A resource is Accessible if it meets the following criteria. A1. (Meta)data are retrievable by their identifier using a standardised communications protocol. A1.1 The protocol is open, free, and universally implementable. A1.2 The protocol allows for an authentication and authorisation procedure, where necessary. A2. Metadata are accessible, even when the data are no longer available.**

Data findability – *related to Findable in the IBISBA termbase -* **A resource is Findable if it meets the following criteria. F1. (Meta)data are assigned a globally unique and persistent identifier. F2. Data are described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the data they describe. F4. (Meta)data are registered or indexed in a searchable resource.**

Data interoperability – the property implies that datasets can be transferred into different systems and processes and still be usable and comprehensible.

Data reusability – this data property implies that existing data can be reused by someone to further his/her research, even when the scientific question being asked is different from the original for which the data was generated.

Digital – digit refers to counting and thus information that is represented by a series of numerical characters from 0 to 9.

FAIR – an acronym used to describe the principles of Data findability, accessibility, interoperability and reusability, as defined by the FAIR Guiding Principles for scientific data management and stewardship (https://www.nature.com/articles/sdata201618)

Governance – the processes and constituted bodies that guide and arbitrate within a multi-actor system.

IBISBAHub - is a database for FAIR sharing of heterogeneous scientific research datasets, models, simulations, processes, and other research outcomes along with information about the projects, people, and organisations that contributed to their creation.

Knowledge – in the classical sense of the word, knowledge is information that is supported by observations and/or measures and is thus held to be true.

Knowledge asset – a synonym of Asset (see above).

Metadata – this term literally means data about data. For example, metadata can include who produced the data, the time and the date when the data was produced, the nature of the sample that is described by the data, the conditions of the experiment and so forth. Standard metadata sets exist for many data domains.

Open archives - archives can contain unpublished data and other assets produced during a project. Open archive platforms provide the means to disseminate publications and associated data.

SOP – meaning Standard Operating Procedure, in the context of IBIBSA 1.0 this term refers to a normalized protocol document that establishes the different steps required to execute an operation (e.g. an in silico or wet bench experiment). SOPs favour experimental reproducibility.

Storage - data storage is the act of writing the data onto a support medium and the creation of backups that ensure that the data is not lost at a later date.

3   Details of the IBISBA 1.0 Data management plan

3.1   Data and project asset governance

The ambitions of IBISBA 1.0 regarding data management in favour of open access, data and asset use and reuse are fully aligned with those of the European Commission. To achieve these ambitions, the IBISBA 1.0 project will implement a series of measures and tools that will require the creation of a data and asset governance.

3.1.1   The IBISBAhub

As mentioned in the glossary, IBISBAHub is IBISBA's central platform for recording, sharing, and accessing assets and workflows related to the IBISBA 1.0 project and its members. IBISBAHub supports the easy deposition of a variety of assets including, SOPs, models and computational workflows, and

the registration of external workflows. This tool is FAIR compliant and thus adopts community standards, such as ISA (Investigation/Study/Assay) metadata standards and Common Workflow Language, so that assets on IBISBAHub can automatically be augmented with relevant metadata.

Registering as a user on the IBISBAHub

Use of the IBISBAHub is on a registered user basis only. Registration requires the user to supply minimal personal data (Family name, given name, and email address). Compliant with GDPR, access to personal data registry is restricted to designated IBISBAHub administrators and users. Upon request, details about personal information held on IBISBAHub will be supplied within one month. Refusal to provide information will occur only in the case of unjustified circumstances (e.g. an elevated number of requests from a single user is received).

Personal data is required to complete registered assets and identify the person responsible for the registration of data/assets. Personal data will only be used if users supply it. It is assumed that users supplying personal data to the IBISBAHub have implicitly given permission for its use. Therefore, explicit consent will not be sought after by the IBISBAHub administrators. Personal data will only be used to describe/complete assets that are generated within the framework of IBISBA. No personal data will be transferred to third parties, unless explicit written permission is obtained from the data owner.

Phased implementation of the IBISBAHub

In the first phase of the project, the IBISBAHub was run in trial mode, meaning that it is mainly available to project beneficiaries. The IBISBAHub is structured in a way that provides users with the means to choose the confidentiality level of the assets that are registered. Each asset can be designated inaccessible, viewable only, downloadable, editable or administrable.

In the second phase of the project, access to the IBISBAHub is being opened to a wider user group, including TNA users.

In a third phase IBISBAHub will become fully open to all users, but will still require user registration.

IBISBAHub governance

In terms of governance, the use and content of the IBISBAHub is under the responsibility of the IBISBA 1.0 General Assembly. However, from an operational standpoint, the General Assembly will delegate governance to the Executive Committee (the ExCom)[1], which is composed of all IBISBA 1.0 workpackage leaders. The ExCom oversees operation of the IBISBAHub and makes recommendations regarding FAIR and GDPR compliancy. The chairperson of the ExCom (the IBISBA 1.0 coordinator) and the IBISBA Coordination Team (a joint management group that oversees both IBISBA 1.0 and PREP IBISBA) takes responsibility for final decisions and their implementation. Nevertheless, project beneficiaries and other depositors reserve proprietary rights with respect to their knowledge assets and thus have the right to protect them from public disclosure as long as decisions are compliant with the IBISBA 1.0 consortium agreement and the embargo policy.

---

[1] Section 6 of the IBISBA 1.0 consortium agreement

Copyright of assets

Regarding copyright, by default assets deposited to the IBISBAHub are attributed a Creative Commons 4.0 licence by the depositor. However, depositors have the option to change the copyright conditions and choose from a range of options including no licence and open public domain. The IBISBAHub administrators and governance assume that depositors have the proprietary rights to the assets and take no responsibility for misuse of assets.

Asset access and embargo policy

Compliant with Horizon 2020 rules, any assets registered on the IBISBAHub will be made FAIR and subjected to open access policy. In minimal terms this means that all assets will be associated with machine-readable metadata that meet international standards and, once publicly available, are associated with a persistent identifier, such as a DOI. In the framework of IBISBA and the use of the IBISBAHub, publicly accessible assets can be attributed an IBISBA specific DOI. The DOI is composed of the prefix 10.34701 and the suffix /ibisba. This functionality is embedded in the IBISBAHub, the DOI minting service being furnished through a contractual relationship with INIST(France), the official French agent of DataCite.

The aim of IBISBA 1.0 is to make all metadata (and preferably the assets that are linked to them) publicly available. However, to protect intellectual property rights and allow exploitation of the assets by their owner specific conditions are proposed:

IBISBAHub depositors have the option to choose the confidentiality option offered by the IBISBAHub. However, the confidentiality option is valid for a maximum period of 48 months, starting from the date of registration on IBISBAHub. Two exceptions to this rule will be any data collected in the framework of projects that have not received public funding and for which at least full costs have been paid. The other exception will be data collected in the framework of projects involving SMEs, as these legal entities benefit from an open access opt out. For these exceptions, data and assets specific to the project can remain undisclosed for an unlimited period of time

3.2 Data production and formats

A wide variety of data will be generated by different project beneficiaries performing various experiments using a range of methods within the IBISBA 1.0 project. These data will be related to different scientific fields including computational biology, chemical engineering, (wet) molecular and systems biology and bioprocess development, including microbial fermentations and downstream processing.

Regarding file formats, a complete list will be established during the project. However, as a guideline, the following file formats will be preferred:

Data typology	Format	Compatibility
Bioparts	Fasta	Text
	SBOL	Text/RDF

	GENBANK	Text
	GBOL	Text/RDF
Biological models	SBML	XML
	PDB	Text
Experimental data	mzML	XML
	Excel	XML
Optimization and design	GAMs	Binary
Flowsheeting and design	Aspen Hysys	
	PSE gproms	
	Intelligen SuperPro	
Chemical data	Mol/MDL	Text
Statistics and regression	ANOVA	
	ALAMO	Text
Documents	PDF	PDF
Workflows	CWL	JSON

3.3 Dataset referencing and naming

General principles

In the IBISBA 1.0 project, a so-called IBISBAHub has been set up. This hub is built on the SEEK platform, a web-based resource for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes, and is an instance of the FAIRDOMHub. As such, the IBISBAHub is designed to catalogue various project data and form associations between them, along with information about the people and organizations involved. The architecture of the platform is designed so that it catalogues all research assets (data, models, SOPs, samples, organisms, publications etc), and the metadata associated with them using a common, related view of the assets organised using the ISA (Investigation, Study, Assay/Analysis) recommendation of Nature and other journals. The platform is flexible regarding the location of content: uploaded to the Hub databases for small datasets and held externally in third party databases where it can be referenced. Thus the IBISBAHub catalogue can span many different datastores including: e-infrastructures installed locally; e-infrastructures provided nationally; and public archives such as the Core Data Resources and Recommended Deposition Databases of ELIXIR[2].

---

[2] European Research Infrastructure for Life Science Data http:/www.elixir-europe.org

Datasets registered on the project hub (hub.ibisba.eu) will be named according to best practices described by Murry et al PLoS Biol 15(6): e2001414. doi:10.1371/journal.pbio.2001414.

Specifically, the format will be URI:object class ID:local ID with the URI name being hub.ibisba.eu. The object classes (e.g. People, data file, etc) will be standardized and made available to partners. The local ID will be a simple numerical code. References to public datasets entries and concepts (Genbank, PDBe etc) will use the identifiers of those datasets, using identifers.org for identifier resolution.

Published data assets can be attributed to an IBISBA specific DOI. The DOI is composed of the prefix 10.34701 and the suffix /ibisba. This functionality is embedded in the IBISBAHub.

3.4 Description of datasets and assets

Data and assets generated within the framework of the IBISBA 1.0 project will be described according to Investigations, Studies and Assays (ISA) recommendations, developed by the ISA-TAB community (http://isacommons.org/) and adopted by FAIRDOM. ISA provides a framework to link the different parts (protocols, data, models) of a project together. A basic experimental description template (.xls format) is available at https://fairdomhub.org/data_files/927?version=2 and new templates will be designed to meet the specific requirements of the different IBISBA 1.0 project activities. These will be made available to the wider scientific community.

3.5 Standards and metadata

Metadata in the IBISBA 1.0 project will follow standards recommended by the SEEK platform, which is a part of the FAIRDOM initiative https://fairdomhub.org/). Regarding SEEK minimal recommendations for metadata, these are:

● Uploader Name (of the person who uploads data or assets). This is the data/asset owner, or the recognized representative of the organization(s) that owns the data/assets. In either case, it is assumed that the named person is invested with rights to upload data. By extension, it is also assumed that the named person has permission from the data/asset owners to set sharing and dissemination. If the uploader is not the author and/or owner, authors and owners can be separately credited. For authors and owners see below 'Names of other people involved'.

● IBISBA ID – (Of the above person). This identifies each person within the IBISBAhub environment. Where users are academics, ORCID will be used as a universal person identifier. If not, the person should simply register to IBISBAhub and enter their details.

● Project – The name of the project is IBISBA 1.0. Typically, assets will only belong to IBISBA 1.0. To reference funding, documents must mention European Commission Horizon 2020 grant agreement number 730976.

● Upload Date (yymmdd). This is the date of uploading, rather than the date of creation. In principle, when assets are uploaded to IBISBAhub, they are automatically tracked. Therefore, new versions of an asset will be identified and timestamped.

● Creation Date (yymmdd). This is the date when the asset was created by its owner(s).

● Title – this a unique IBISBA name for each asset file. Project uploads bearing names that are already used will be automatically rejected by the system. Assets uploaded via a JERM (Just Enough Results Model) harvester without a title will be treated, attributing a unique name.

● Version Number – this is a means to track file iterations. It is proposed to keep this simple, notifying iterations by an n+1 increment, the initial version being "0".

● Names of other people involved – This metadata field provides a means to credit co-workers and recognize all those involved.

Additional metadata will be:

● Dataset type. This is the dataset descriptor (see table in section 5.4) and could be for example "Video", "Slidedeck", "Metabolomics" or "Strain construction"

● Format. This is simply the description of the data file format (*c.f.* section 3.2)

FAIRDOM is participating in the European Open Science Cloud EOSCPilot project and ELIXIR to be compliant to developments in metadata for datasets and repositories including EDMI (https://eoscpilot.eu/edmi-metadata-guidelines) and Bioschemas.org. These are in turn compliant to W3C standards DCAT.

3.6 Data sharing

For data sharing, data generated within the scope of the IBISBA 1.0 project will be registered by project beneficiaries at the IBISBAHub. Access to the IBISBAHub will be available after user registration. In the early stage of the project it is proposed that only Project beneficiaries can register to the IBISBAHub and that three levels of access are defined

1. Data and assets are disclosed to everyone (including public)
2. Data and assets are disclosed to the project consortium.
3. Data and assets are disclosed to workpackage leaders

For restricted access data (levels 2 and 3), an embargo period will be applied up to 48 months after the official end date of the IBISBA 1.0 project or up to the point at which the data is published, whichever is the shortest period. The embargo period can be shortened upon request by the asset owner.

Whenever possible, upon data registration the data itself should be either to the IBISBAHub or to another public (specialized or generalist) repository.

3.7 Archiving and preservation (including storage and backup)

Data assets generated during the IBISBA 1.0 project must be compliant with the following rules:
- a. Data is owned by the organisation(s) creating it (IBISBA 1.0 Consortium Agreement).
- b. The owner (data manager) of the data is responsible to ensure that data is
    - i. registered on the IBISBHub
    - ii. stored in an appropriate location
    - iii. ready for long-term preservation and/or publishing.

Regarding long-term preservation and/or publishing, various recommendations are made to project beneficiaries. These are briefly outlined below:

Publications.

Compliant with H2020 rules, project beneficiaries will be required to adopt best-practice regarding research reports. Specifically for publishing, journals offering an open access (Gold) option should be used. Alternatively, project beneficiaries should use journals that allow self-archiving (green). In this case project beneficiaries are recommended to use a stable archive such as bioRxiv.org. This is a free online archive and distribution service for scientific reports. It is operated by Cold Spring Harbor Laboratory, a not-for-profit research and educational institution. No cost is associated with the use of this archive. All publications will be uploaded to openAIRE, the open access repository for EU projects.

The SEEK platform that implements IBISBAHub has rich capabilities for publication including minting DOIs for snapshots and packaging data to be deposited in long term storage publishing platforms such as Zenodo (part of openAIRE). The FAIRDOM team are collaborators with the openAIRE team.

Data archives and data papers

Regarding complete datasets associated with published work, project beneficiaries will be encouraged to publish these with the publication as supplementary data or as independent data papers, using one of the many platforms that are designed to do this and/or using an open access journal that accepts data papers. Recommended platforms that will be integrated with IBISBAHub are OpenAire and Mendeley data (https://data.mendeley.com/), meaning that these platforms will automatically harvest compatible datasets from IBISBAhub. Other open source web applications for sharing, preserving, citing, exploring, and analyzing research such as Dataverse (https://dataverse.org/) can also be used.

Public data deposition archives will be used for relevant data-types. The EU Research Infrastructure ELIXIR maintains a list of Core Data Resources and Recommended Deposition Databases https://www.elixir-europe.org/platforms/data. Thanks to the SEEK platform underpinning IBISHAHub, data deposited in ELIXIR public archives can be organised and catalogued alongside other project datasets, models and SOPs.

Workflows will be deposited in myExperiment.org. Parts will be deposited in JBEI-ICE.

Official document archiving

Regarding project reports (deliverables and milestones): compliant with European Commission rules, these will be systematically deposited on the H2020 project platform. They will also be conserved by the project management team who will ensure that each document is attributed a DOI and registered on the IBISBAHub. Moreover, the management team will explore ways to archive these key documents at the end of the project using for example the HAL INRAE archive (https://hal.inrae.fr/). Any confidential deliverables will be archived with limited access until the end of the embargo period (48 months after the initial submission to the European Commission) or until the data therein is published, whichever is the shortest.

Other project data and documents

The archiving of other assets pertaining to the project (e.g. raw data, project notes etc.) will be the responsibility of each project beneficiary and subject to H2020 rules.