"Microbial communities and TIC" project DMP

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).

Plan Details

Plan title	"Microbial communities and TIC" project DMP
Fields of science and technology (from OECD classification)	Computer and information sciences, Biological sciences (Natural sciences)
Language	eng
Creation date	2023-03-24
Last modification date	2023-07-27
Associated documents (publications,reports, patents, experimental plan...), website	• Web Site : https://project.inria.fr/mistic

Project Details

Project title	Microbial communities and TIC
Acronym	MISTIC
Abstract	Digital technologies have tremendous potential for facilitating and accelerating the development and deployment of agroecological innovations: optimized varietal selection taking into account plant-microbiome interactions, biostimulation and management of plant immunity, biocontrol based on the use of bioinputs or management of resident beneficial organisms, plant nutrition. Agro-ecological cropping systems experience a wide range of biotic interactions with complex microbial communities: **beneficial** to the plant, providing important nutritional and biodefense functions, or **detrimental**, notably microbial parasites and pathogens exploiting plant resources. The diversity and dynamics of these interactions depend on ecological conditions, including the phenotypes of the interacting species, on physiology, and on the abiotic environment. Deciphering the links between interspecific diversity, community structure, and biological functions is key to understanding, maintaining, diagnosing, and exploiting the community dynamics underlying the health or illness of a crop, and adapting agroecological systems to environmental stresses. We will design new multi-omic data analysis tools and develop

multi-scale spatio-temporal models of microbial communities in crop plants. This will require significant development of algorithmic, dimension reduction, and machine learning methods for data analysis; significant advances in ODE and PDE based system dynamics, hybrid numerical and discrete modeling of the complex biological systems; and significant advances in **transversal AI and HPC techniques**. Acquisition of novel data to support AI machine learning and numerical modeling is necessary to provide methodological guarantees and validation. We will capitalize on existing culture systems to anchor our work to pertinent challenges in agroecology, while acquiring novel data specific to this PEPR.

We will use **large-scale, genome-resolved metagenomic analysis** to address key questions about the functional properties of plant-microbiome biochemical reaction networks. We will implement community-scale metabolic network analyses to identify key species and metabolic functions that mediate plant-microbiome interactions. Advances in AI will be needed to cope with noisy data as well as to refine knowledge-based reasoning methods for network analysis. We will construct **digital twins of reduced microbial communities** and plant–microbiome systems, through spatio-temporal models fitted to experimental data. We will use culturomics to design and cultivate microbial communities as controlled and repeatable experimental models of natural communities, for each pathobiont or symbiont system of interest. Plant-microbiome feedbacks will be characterized through the evolution of microbial communities and plant response.

The two-pronged **systems biology** strategy of this project addresses methodological challenges up- and downstream of specific applications in biocontrol. An expected outcome of this proposal is development of computer software and mathematical tools for a deeper understanding of the links between microbial community structure and crop response. This entails identifying key drivers of plant microbial communities that impact plant health and crop traits to design and assess ideotypes and species associations of agronomic interest. New modeling tools will foster knowledge of crop-microbiome diversity and interactions, with translational applications to be tested in the framework of the SADEA big challenge on plant biocontrol. The data acquired for this project will constitute a unique reusable resource for French scientists: those developing methods, those developing applications, and those representing agroecological stakeholders.

Funding

- Agence Nationale de la Recherche : ANR-22-PEAE-0011

Start date 2022-11-07

End date 2027-11-06

Partners

- Institut national de recherche en informatique et en automatique https://ror.org/02kvxyf05
- Institut national de recherche pour l'agriculture, l'alimentation et l'environnement https://ror.org/003vg9w96

Research outputs :

1. Experimental data produced by service platforms (Dataset)
2. Databank data extracted from public data banks (Dataset)
3. Partner data shared with the project but not otherwise available in a public data bank (Dataset)
4. Calculated data produced by computer programs from other data (Dataset)
5. Source code for computer software (Software)
6. Computational workflows that automate analyses (Workflow)

Contributors

Name	Affiliation	Roles
Sherman David James - https://orcid.org/0000-0002-2316-1005	Inria - https://ror.org/02kvxyf05	• DMP manager • Personne contact pour les données (Calculated Data, Experimental Data, Databank Data, Partner Data, Software, Workflows) • Project coordinator

"Microbial communities and TIC" project DMP

1. Data description and collection or re-use of existing data

Experimental data produced by service platforms

1a. How will new data be collected or produced and/or how will existing data be re-used?

Experimental data will be produced by service platforms by performing biological analyses on samples collected from full field or from horticultural plant crop cultures. Biological samples may be directly processed or extracted from biobanks. Biological analyses will follow protocols defined by the service platforms. Specific cases are as follows.

Metabolomic data:

Methodologies or software: New metabolomic data will be acquired using state-of-the-art targeted and untargeted metabolomics at the Bordeaux metabolome facility (10.15454/1.5572412770331912E12) following protocols for LCMS-based metabolomics and targeted biochemical phenotyping described previously (Luna et al., 2020, 10.3390/metabo10030096; Dussarrat et al., 2022, 10.1111/nph.18095). MSDIAL will be used for the deconvolution of MS spectra.

Documentation of data provenance: Metadata identifying the samples and generated metabolomics variables will be documented.

DNA sequencing:

Methodologies or software: DNA sequences will be acquired using medium- and high-throughput sequencing technologies providing long and short reads using Illumina NextSeq2000, Illumina MiSeq, and Oxford Nanopore Technologies P2 solo sequencers.

Documentation of data provenance: Depending on the sequencing machine, the NextSeq 1000/2000 Sequencing System and the PromethION Data Acquisition Unit A100 instrument documentation will be used.

Soil samples:

Methodologies: Soil samples will be acquired from crop cultures and processed following Genosol protocols. Additional processes for sample treatment may be developed in MISTIC. DNA extracted from soil samples will be processed as above in ¶ *DNA Sequencing*.

Documentation of data provenance : Metadata identifying the samples will be documented, including date, host plant, and location. Samples may be acquired from existing crop culture experiments performed by partner projects, such as MetaNema. Horticultural samples are collected from three different agricultural greenhouses (Crop Lambesc, Crop Le Thor, and Crop Pernes-les-Fontaines) with five different types of plants (eggplant, eggplant rootstock, salad, cucumber and peppers).

Leaf samples:

Methodologies: Leaf samples will be acquired from crop cultures. Grapevine leaf samples are collected in Aquitaine as described in Paola Fournier's PhD thesis (PPR-CPA VITAE project), and analyzed using culture-dependent methods as described in Aarti Jaswa's PhD thesis (VITAE/BSA) and metabarcoding approaches (VITAE). Strains will be used to build synthetic microbial communities (SynComs) antagonists of downy mildew (VITAE). Time series data of strain abundance within SynComs will be collected using a quantitative microfluidic chip (VITAE). DNA extracted from leaf samples will be processed as above in ¶ DNA Sequencing.

Documentation of data provenance: A brief description of the sampling design and the culture collection can be found here: Fournier et al., 2023, https://ives-openscience.eu/34866/

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

MISTIC has acquired 200 TB of storage space for all project data.

DNA sequences: Sequence data in EMBL format will be generated from FastA and FastQ raw sequences; 30 terabytes. Genomic sequences will be generated from soil samples in French southern soil and rhizosphere microbiomes, from leaf samples for approximately 1000 genomes (bacteria, yeasts and filamentous fungi) isolated from grapevine leaves. DNA sequencers at the PGTB and Gentiane facilities generate fastQ files, which will be staged on the platform NAS in a MISTIC-specific folder for subsequent access by project participants.

Metabolomic data (molecular species concentrations over time): open-source tabular formats, 75 gigabytes.
Raw data are generated in proprietary format by the machines (.raw, .dcl, .pai2, etc.), directly used by MSDIAL or then converted to the open mzML format. After treatments, data are shared in CSV format. Raw data are gigabyte-sized while processed CSV files are megabyte-sized. Data will be generated at the Bordeaux metabolome facility. If needed and relevant, data coming from the Metabolights or Metabolomic Workbench databases could be used (usually stored

as raw or mzML files).

Data tables: Text UTF-8 using nonproprietary tabular format (CSV or R object), or the nonproprietary HDF5 format, 10 gigabytes.

Databank data extracted from public data banks

1a. How will new data be collected or produced and/or how will existing data be re-used?

Databank data extracted from public data banks, such as the EMBL database of DNA sequences, will be collected with their associated metadata, including unique identifiers, provenance, experimental context, associated ontology terms, and associated scientific publications.

Methodologies or software: Web sites will be used to prepare queries that will be submitted to public APIs to extract formatted data.

Documentation of data provenance: Each extracted dataset will be associated with the specific query against an API endpoint that provided the result.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Data volume and formats as described in ¶ *Experimental data* and ¶ *Calculated data*.

Partner data shared with the project but not otherwise available in a public data bank

1a. How will new data be collected or produced and/or how will existing data be re-used?

Partner data will be provided by academic or industrial partners.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Data volume and formats as described in ¶ *Experimental data* and ¶ *Calculated data*.

Calculated data produced by computer programs from other data

1a. How will new data be collected or produced and/or how will existing data be re-used?

Calculated data will be produced by computer programs from other data. Workflows may be used to automate calculations. Specific cases are as follows.

Assembled genomes: Methodologies or software: Assembled genomes will be produced by running de novo assembly and metagnomic binning tools on raw sequencing datasets either produced by the project or publicly available. Tools will be chosen among the state of the art tools (e.g. MetaFlye, MetaBat) and/or developed during the project. The quality of produced assemblies will be evaluated using standard metagenomic evaluation tools, such as metaQuast and CheckM. Developed tools and the full pipeline to obtain assembled genomes will be made available open source on github repositories. Parameters for assembly may be chosen based on input sequence metadata, such as Kingdom,

sample preparation, or data source. Documentation of data provenance: Sequencing data provided by MISTIC will be documented as described in ¶ *Experimental data*. When re-using existing raw sequencing data extracted from databanks, their provenance will be documented by citing the journal articles and giving their accession ids in the databases that they were retrieved from. The RO-Crate for each assembled genome will specify the input data provenance and the workflow, or the specific query against an API endpoint that provided the result. **Genome-scale metabolic networks:** Methodologies or software: Genome-scale metabolic models will be calculated using reusable workflows, such as Metagenopic developed in Pleiade, or by using public web services. Documentation of data provenance: The RO-Crate for each calculated model will specify the input data provenance and the workflow, or the specific query against an API endpoint that provided the result. **Simulation results and predictions:** Methodologies or software: Simulation results and predictions will be computed using reusable workflows. Documentation of data provenance: The RO-Crate for each calculated model will specify the input data provenance and the workflow, or the specific query against an API endpoint that provided the result.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

MISTIC has acquired 200 TB of storage space for all project data. **Assembled genomes**: FastA or FastQ sequences, EMBL format when annotated. Volume estimated at 5 terabyte. **Genome-scale metabolic networks**: SBML format and ad hoc RDF formats. Volume estimated at 10 gigabyte. **Simulation results and predictions**: Text UTF-8 using appropriate surface syntax (CSV, R object), or the nonproprietary HDF5 format. Volume depending on results. **Generated or simulated training data**: Text UTF-8 using nonproprietary tabular format (CSV or R object), or the nonproprietary HDF5 format. Volume depending on results.

Source code for computer software

1a. How will new data be collected or produced and/or how will existing data be re-used?

Software is developed by MISTIC participants of all profiles, including researchers, engineers, and students. Software development entails specification, programming, and testing of computer programs. MISTIC will encourage the adoption of institutional project-specific best practices for software quality assurance. MISTIC will use Inria's GitLab instance https://gitlab.inria.fr to host specifications, source code, and test results.

Software specifications may be informal, or may adhere to different software engineering formalisms. MISTIC will encourage software developers to include specifications in source code repositories; for example, for software projects using behavior-driven development, the feature files will be included in the source code.

Software programming involves writing source code for computer programs. MISTIC will encourage the use of GitLab tools for assisting program writing and modification for evolutive or corrective maintenance, including *issues*, *labels*, *milestones*, and *merge requests*.

Software testing improves the quality of software and confidence in its correctness by comparing the results of program executions to expected results. MISTIC will encourage the adoption of *continuous integration* practices, using GitLab CI in particular.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Software source code may be produced in a variety of programming languages, including the Python general-purpose language and the R language for statistical computing in particular.

Software source code volume is measured in *kloc*, thousand lines of code, and each software system developed for MISTIC is estimated to be between 5-35 kloc.

Software artifacts may be generated from source code by compilation or packaging, in formats specific to their respective programming languages.

Software artifacts and their configuration may also be packaged in *container images* to facilitate their use on computing platforms and in reusable workflows.

Computational workflows that automate analyses

1a. How will new data be collected or produced and/or how will existing data be re-used?

Workflows developed by MISTIC entail the specification of computation steps used to transform input data into results.

Workflow development will adhere to the same principles described in ¶ *Software*. MISTIC will use Inria's GitLab instance https://gitlab.inria.fr to host workflow development and will encourage the use of GitLab tools for assisting program writing and modification for evolutive or corrective maintenance, including issues, labels, milestones, and merge requests.

Workflow testing improves the quality of workflows and confidence in its correctness by comparing the results of program executions to expected results. MISTIC will encourage the adoption of continuous integration practices, using GitLab CI in particular.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

Workflows may be defined using a variety of formalisms, depending on the computing environment in which the workflow will be executed.

MISTIC will promote the use of the Common Workflow Language (CWL) for the specification of reusable and platform-agnostic workflows.

2. Documentation and data quality

Experimental data produced by service platforms

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Metabolomic data:

Metadata provided to help others find the data: All required metadata on the samples and metabolomic variables will be shared (including exact m/z and retention time, and InChiKey if metabolites are annotated/identified)

Documentation needed to enable re-use: Sample metadata, metabolomic variables metadata.

Where metadata information is recorded: Metadata for Sample/Class ID are recorded by the MSDIAL software and are stored in the CSV files shared with the partners during the project and stored in Metabolights database once research papers are published. Other Metadata will be stored in separate CSV files.

Soil samples:

Metadata provided to help others find the data: We will use Genome metadata format.

Documentation needed to enable re-use: Genome metadata format would include localisation and date of sampling, DNA extraction protocol, sequencing technology, data processing tools inventory, and genomic features such as GC% or gene content.

Where metadata information is recorded: Data will be available on Dataverse INRAe (entrepot.recherche.data.gouv.fr)

DNA sequences:

Metadata provided to help others find the data: Sample name, library kit, sequencer, sequencing reagent kit, flow cell type, sample sheet.

Documentation needed to enable re-use: No specific documentation is necessary, the fastQ format is ready to use and universal.

Metadata will be recorded in the header of EMBL sequence file.

In addition to sample provenance metadata, DNA sequences will be accompanied by a MultiQC (https://multiqc.info/) sequencing report and an analysis with Dragen Metagenomics, for Illumina, or WIMP, for Oxford Nanopore.

Leaf samples:

Metadata provided to help others find the data: For each microbial genome assembled during the MISTIC project, we will provide the metadata requested by the NCBI: isolate taxonomy, isolate identifier, host, isolation source, collection date, person in charge of the collection, latitude and longitude.

Documentation needed to enable re-use: Raw sequence data will be available from the NCBI Bioproject associated with each genome. Sequencing and assembly methods will be described in a document deposited in Research Data Gouv.

Metadata will be recorded in the NCBI BioSample associated with each genome.

2b. What data quality control measures will be used?

Experimental data will be subject to data quality measures defined by experimental platforms.

Metabolomic data:

Calibration, repeated measurements, standardized data capture, data entry validation, controlled vocabularies: State of the art calibration of the equipment will be used. Sufficient replicates (at least 3, if possible 5) will be analyzed. When metabolites are annotated/identified, chemical ID such as InChiKey, SMILE, ChEBI identifiers will be used with the ontology linked to those identifiers. RSD for QC samples (repeated measures of pooled samples) will also be used to account for variation coefficient of metabolomics variables.

DNA sequences:

Sequencing runs include an internal control (for example PhiX for Illumina runs) which is used to calibrate and/or validate the sequencing. All runs are validated only if the expected specifications are met (output, quality, …)

DNA sequences will be accompanied by a MultiQC (https://multiqc.info/) sequencing report.

During an early phase of MISTIC, a single sample will be sequenced on both systems, in order to calibrate future experiments in terms of the reading depth required to obtain accurate results.

Soil samples:

Calibration, repeated measurements, standardized data capture, data entry validation, controlled vocabularies:
We will use a standard genomic and metagenomic quality assessment protocol supported by multiple sequencing of same samples.

Leaf samples:

Calibration, repeated measurements, standardized data capture, data entry validation, controlled vocabularies:
Quality of sequence data will be verified by the sequencing facilities (Gentiane, Clermont-Ferrand and PGTB, Bordeaux, France).

Databank data extracted from public data banks

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Metadata and documentation for databank data are defined by the databank.

2b. What data quality control measures will be used?

Databank data will be subject to data quality measures defined by the databank.

Partner data shared with the project but not otherwise available in a public data bank

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

In the specific case of grapevine leaf samples, partner data from the EPIcure information system will be used to choose open field plots, but this metadata may not be shared and cannot be associated with the leaf sampledata the MISTIC will produce.

2b. What data quality control measures will be used?

Partner data will be subject to data quality measures defined by experimental platforms.

Calculated data produced by computer programs from other data

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Metadata provided to help others find the data: Metadata for calculated data will include provenance of input data, parameters, versions of software, and workflows. Documentation needed to enable re-use: Calculated data and accompanying metadata will be formatted using RO-Crates, which will provide a form of documentation, instructions for recomputation, and a concrete example inspiring reuse. Where metadata information is recorded: Metadata are recorded in the RO-Crate. Specific metadata requirements are as follows: **Assembled genomes:** Parameters used for assembly, input sequence provenance and metadata. If raw sequence data are generated by MISTIC, their metadata will include documentation of the biological system, parameters of the experiments, and identifiers in sequence read repositories (ENA, SRA). For data re-used from publicly available databases, their metadata will include references to the corresponding databases and publications.
Genome-scale metabolic networks: Parameters for the construction pipeline, input data provenance and metadata.

2b. What data quality control measures will be used?

Computed data will obey two data quality measures. First, procedures for capturing computation provenance in RO-Crates will guarantee that computed data are produced in a reliable and reproducible fashion. Second, algorithms and software for computing data should be published with validation of their correctness.

Biomedical ontologies are used to provide controlled vocabularies.

Assembled genomes: Calibration, quality measures, controlled vocabularies: The quality of assembled genomes will be assessed by estimating its sequence completeness and level of contamination using standard tools, such as CheckM that relies of universal sets of genes that must be present in unique copy in bacterial genomes. **Genome-scale metabolic networks:** Calibration, quality measures, controlled vocabularies: Workflows for constructing genome-scale metabolic networks include quality assessment steps chosen from the literature. Network labels are chosen from biomedical ontologies.

Source code for computer software

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Metadata about the software development process are provided by the history of GitLab issues, issue comments, merge requests, merge request comments, milestones, git commit logs, and releases.

Software documentation will distinguish between *user documentation*, describing how the software is installed, used, and parameterized; and *developer documentation*, describing the software architecture and how corrective and evolutive maintenance could be performed. Software documentation will be available online and in PDF format. MISTIC will encourage software authors to include tutorials, in preference in the form of executable Jupyter notebooks.

Software that relies on third party open source modules will be associated with a *software bill of materials* (SBOM).

2b. What data quality control measures will be used?

Software quality control will be promoted through review of specifications, use of programming best practices, and testing, as described in ¶1. Behavior-driven development and continuous integration processes will be employed.

Computational workflows that automate analyses

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Workflow data are treated as computer software and development metadata and documentation will follow the guidelines in ¶ Software.

2b. What data quality control measures will be used?

Workflow data are treated as computer software and quality control measures will follow the guidelines in ¶ Software.

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

For day-to-day operations, data and metadata for MISTIC will be stored in a secure data lake maintained by Inria in the Pleiadès Kubernetes cluster located in the Inria centre at the university of Bordeaux.
The cluster uses a triply redundant, resilient, and self-healing Ceph storage infrastructure that does not require back up in the traditional sense.

3b. How will data security and protection of sensitive data be taken care during the research

MISTIC will not acquire or use sensitive data.

---

4. Legal and ethical requirements, code of conduct

Experimental data produced by service platforms

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Intellectual property will be protected according to PEPR and institutional policies.

Intellectual property rights to data produced by MISTIC will be shared by Inria and INRAE as stipulated in the Consortium Agreement. Data that are made public will be associated with a license decided by Inria and INRAE, by default CC-BY.

Specific rights attached to biological materials produced by INRAE will be managed by the INRAE. Biological materials such as soil and leaf samples, DNA preparations, cultures will be conserved in biobanking facilities according to INRAE protocols and standards. Soil samples will be conserved at the Genosol facility.

Know-how developed during the course of this project will be protected pursuant to EU Commission Regulation No 316/2014. In order to create the greatest positive impact on best practices and to promote uptake by stakeholders, we will seek to develop and publish experimental protocols capturing acquired know-how.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.

General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

Databank data extracted from public data banks

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Databank data will use used only as permitted by the databank data provider's license and no intellectual property rights for Inria and INRAE are created by this use.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.

General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

Partner data shared with the project but not otherwise available in a public data bank

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Intellectual property will be protected according to PEPR and institutional policies.
No intellectual property rights to partner data furnished to MISTIC are created for Inria and INRAE by this use.
Intellectual property rights to calculated data or to software developed or validated using partner data are defined in ¶ *Calculated data* and ¶ *Software*.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.
General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

Calculated data produced by computer programs from other data

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Intellectual property will be protected according to PEPR and institutional policies.
Intellectual property rights to calculated data produced by MISTIC will be shared by Inria and INRAE as stipulated in the Consortium Agreement. Data that are made public will be associated with a license decided by Inria and INRAE, by default CC-BY.
Intellectual property rights to calculated data produced by MISTIC **using** partner data obtained under a confidentiality agreement, furnished to MISTIC by an industrial partner, or acquired by MISTIC through industrial collaboration, will be shared by Inria and INRAE as above unless otherwise stipulated by a specific contract with the partner providing the data.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.
General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

Source code for computer software

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data. Software will not be developed that uses personal data.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Intellectual property will be protected according to PEPR and institutional policies.

Software will be co-owned by partner institutions *pro rata* to employee contributions recorded using Inria's BIL database. Releases will be deposited in APP.

Software may re-use third party software available under an open-source license. Auditing of open sources licenses will be performed using FOSSA.

Know-how embodied in software developed during the course of this project will be protected pursuant to EU Commission Regulation No 316/2014. In order to create the greatest positive impact on best practices and to promote uptake by stakeholders, we will seek to develop and publish experimental protocols using this software and demonstrating the acquired know-how.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.

General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

Computational workflows that automate analyses

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

MISTIC will not acquire or process personal data. Workflows will not be developed that use personal data

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Intellectual property will be protected according to PEPR and institutional policies.

Workflows will be co-owned by partner institutions *pro rata* to employee contributions.

Know-how embodied in workflows developed during the course of this project will be protected pursuant to EU Commission Regulation No 316/2014. In order to create the greatest positive impact on best practices and to promote uptake by stakeholders, we will seek to develop and publish experimental protocols using these workflows and demonstrating the acquired know-how.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

No specific ethical issues are identified for MISTIC.

General ethical issues and codes of conduct as established by Inria and INRAE will be followed.

5. Data sharing and long-term preservation

Experimental data produced by service platforms

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Experimental data produced by MISTIC will be made available according to Open Data principles except where forbidden or embargoed for PPST reasons. Experimental data provided as a MISTIC deliverable will be recorded in a trustworthy data repository, by default https://entrepot.recherche.data.gouv.fr/dataverse/inrae.

Experimental data associated with a scientific publication will be made available, as described above, at the **latest** at the date of publication of the article. Unless an earlier date is chosen by the MISTIC data steward, experimental data will be under embargo until the publication of the princeps article.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Experimental data associated with a scientific publication or provided as a MISTIC deliverable, are chosen implicitly for preservation and will be preserved in a trustworthy data repository, as described above.

Other data may be chosen explicitly for preservation with the accord of the holders of intellectual property rights.

5c. What methods or software tools are needed to access and use data?

Controlled access to secure data lake will be provided to partners through a data portal providing Ceph object store access using an S3-compatible REST API. There are many clients for using this interface, both from the command line or through we portals.

Data for individual studies may be stored and accessed within the Pleiadès cluster using shared disk volumes, which may be mounted into batch computations or interactive Jupyter notebooks deployed in the cluster.

Software for using specific formats, such as HDF5 or sequence data, will be chosen *ad hoc*. The choice of software tools and their versions will be documented in reusable workflows defined in standard formats, such as CWL.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Experimental data preserved in a trustworthy data repository will have a DOI associated with them.

Databank data extracted from public data banks

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Databank data is already shared and does not require any management by MISTIC.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Databank data will not be preserved long-term, as it can be retrieved as necessary.

If a databank becomes unavailable permanently or through the end of MISTIC, the data will be reclassified as *partner data* and managed as such.

5c. What methods or software tools are needed to access and use data?

Controlled access to a secure data lake will be provided to partners through a data portal providing Ceph object store access using an S3-compatible REST API. There are many clients for using this interface, both from the command line or through we portals.

Data for individual studies may be stored and accessed within the Pleiadès cluster using shared disk volumes, which may be mounted into batch computations or interactive Jupyter notebooks deployed in the cluster.

Software for using specific formats, such as HDF5 or sequence data, will be chosen *ad hoc*.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Databank data will be identified by the unique and persistent identifier used by the databank, including the namespace of the databank.

Partner data shared with the project but not otherwise available in a public data bank

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Partner data shared with MISTIC will be made available by MISTIC to all its participants, but not made available outside of the project. Partners who wish to make their data more broadly available may do so themselves using a trustworthy data repository.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Partner data that cannot be retrieved at will, will be preserved by MISTIC only as long as the project is active and that preservation is permitted by legal agreement with the partner.

5c. What methods or software tools are needed to access and use data?

Controlled access to a secure data lake will be provided to partners through a data portal providing Ceph object store access using an S3-compatible REST API. There are many clients for using this interface, both from the command line or through we portals.

Data for individual studies may be stored and accessed within the Pleiadès cluster using shared disk volumes, which may be mounted into batch computations or interactive Jupyter notebooks deployed in the cluster.

Software for using specific formats, such as HDF5 or sequence data, will be chosen *ad hoc*.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Partner data that already has a unique and persistent identifier provided by the partner with the data set, will use that identifier. We will encourage partners to obtain such identifiers if they do not already exist.

Partner data that does not already have an identifier will be assigned a name-based version 5 UUID obtained by hashing the partner namespace designator and a dataset identifier. This identifier will be unique and persistent within MISTIC.

Calculated data produced by computer programs from other data

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Calculated data published in support of a method or as a MISTIC deliverable, will be formatted using RO-Crates and deposited in a trustworthy data repository, by default https://entrepot.recherche.data.gouv.fr/dataverse/inria.

Calculated data associated with a scientific publication will be made available, as described above, at the **latest** at the date of publication of the article. Unless an earlier date is chosen by the MISTIC data steward, calculated data will be under embargo until the publication of the princeps article.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Calculated data associated with a scientific publication or provided as a MISTIC deliverable, are chosen implicitly for preservation and will be preserved in a trustworthy data repository, as described above.

5c. What methods or software tools are needed to access and use data?

Controlled access to a secure data lake will be provided to partners through a data portal providing Ceph object store access using an S3-compatible REST API. There are many clients for using this interface, both from the command line or through we portals.

Data for individual studies may be stored and accessed within the Pleiadès cluster using shared disk volumes, which may be mounted into batch computations or interactive Jupyter notebooks deployed in the cluster.

Software for using specific formats, such as HDF5 or sequence data, will be chosen *ad hoc*.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Calculated data preserved in a trustworthy data repository will have a DOI associated with them.

Source code for computer software

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Computer software will be shared within the project during development using https://gitlab.inria.fr, with the

expectation that the repository itself will be made public, at the latest at the time of publication of any scientific publication or RO-Crate what uses the software.

Computer software provided as a MISTIC deliverable will be published in the trustworthy repository https://hal.inria.fr and thus in Software Heritage.

Computer software associated with a scientific publication, for which the source code repository is not made available as described above, will be made available in HAL at the latest at the date of publication of the princeps article.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Computer software associated with a scientific publication or provided as a MISTIC deliverable, are chosen implicitly for preservation and will be preserved in a trustworthy data repository, as described above.

5c. What methods or software tools are needed to access and use data?

Computer software and compiled packages or artifacts will be accessible through the Inria GitLab instance. Authorized people will authenticate to the web interface https://gitlab.inria.fr using their individual credentials, which will allow them to discover and retrieve software data. Authorized computational processes including workflows will be provided with revokable tokens, which will allow them to retrieve software data.

Computer software that has been chosen for inclusion in a public repository, such as PyPI, CRAN, or container image repositories, can be retrieved using publicly available repository-specific tools.

Computer software that has been chosen to be preserved in a trustworthy data repository, can be retrieved from the repository.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Computer software preserved in the trustworthy data repository Software Heritage will have a SWHID associated with them. Computer software deposited in the APP will have a unique identifier associated with them.

During development, software will use the Git distributed version control system and each version of the software source code will be identified by a *commit hash* that will be unique and persistent within the project.

Versions of computer software provided as MISTIC deliverables will be identified by a *semantic version*.

Computational workflows that automate analyses

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Workflow data are treated as computer software and will be shared following the guidelines in ¶ *Software*.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Workflow data are treated as computer software and will be preserved following the guidelines in ¶ *Software*.

5c. What methods or software tools are needed to access and use data?

Workflow data are treated as computer software and will be used following the guidelines in ¶ *Software*.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Workflow data are treated as computer software and will be given unique and persistent identifiers following the guidelines in ¶ *Software*.

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

The project coordinator, David Sherman, is ultimately responsible for data management. He will be assisted by an engineer from the Inria DSI co-funded by MISTIC.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Adherence to FAIR principles and commitment to open science is global and transversal in MISTIC: part of all development efforts will include time dedicated to these commitments.

MISTIC will fund a computer engineer working at 20% over the five years of the project, specifically for managing computing resources, managing data, and for assisting scientists in meeting MISTIC commitments to open science