European Network of Fourier-Transform Ion-Cyclotron-Resonance Mass Spectrometry Centers

Grant Agreement n° 731077

Deliverable D3.6
Data management plan – Iteration #3

Start date of the project: 1st January 2018
Duration: 54 months
Project Coordinator: Christian ROLANDO – CNRS-
Contact: christian.rolando@univ-lille1.fr

“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731077”

Document Classification

Title	Data Management Plan – iteration #3
Authors	P11 CASC4DE – Camille Beluffi Marin
Work package	WP3 – Open Data and e-Infrastructure
Dissemination	PU: Public
Nature	ORDP
Doc ID Code	20210203_EU_FT-ICR_MS_D03.06
Keywords	Data, management, infrastructure

Document History

Name	Date	Comment
P11 CASC4DE – Camille Beluffi Marin	2021-02-03	Third version

Document Validation

Project Coordinator	Date	E-mail
P1 CNRS – Christian Rolando	2021-02-03	christian.rolando@univ-lille1.fr

Neutral Reviewer	Date	E-mail
P1 CNRS – Christian Rolando	2021-02-03	christian.rolando@univ-lille1.fr

The author of this report is solely responsible for its content, it does not represent the opinion of the European Commission and the Commission is not responsible for any use that might be made of the information it contains.

Document Abstract

The Data Management Plan (*DMP*) is a text document that presents how, by whom, how long or where the data produced during the EU FT-ICR MS project will be managed. The *DMP* is provided to all partners of the project and describes the concrete steps that will be followed to handle all the data produced during the project.

For the European *FT-ICR MS* project, the DMP is based on official documents like the Grant Agreement of the project to respect some specific and mandatory rules. It was also completed with information about the solutions that will be built to support the infrastructure.
This *DMP* version is planned to change along with the project to reach its final version at the end of the project. It will be completed one more time (if necessary) it in order to cover for the modifications that could be needed to face possible issues.

Table of Contents

EU_FT-ICR_MS Horizon 2020 Project FAIR DMP - V 0.3

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : Laura Duciel, Delsuc Marc-André, Julia Asencio Hernández, Camille Beluffi

Affiliation du créateur principal : Université de Strasbourg

Modèle du PGD : Horizon 2020 FAIR DMP (anglais)

Dernière modification du PGD : 03/02/2021

Numéro de subvention : H2020-EU.1.4.1.2. - Integrating and opening existing national and regional research infrastructures of European interest

Résumé du projet :

EU_FT-ICR_MS proposal aims to establish a European network of FT-ICR (Fourier Transform Ion Cyclotron Resonance) mass spectrometry (MS) centers in association with a manufacturer and a SME software company. Mass spectrometry (MS) has become the most ubiquitous analytical techniques in use today, providing more information on the composition and the structure of a substance from a smaller amount of sample than any other techniques. For more information on the project: https://cordis.europa.eu/project/rcn/212587_en.html

Chercheur Principal : CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS

Identifiant ORCID : RUE MICHEL ANGE 3 75794 PARIS France

Contact pour les Données : CASC4DE Pôle API - 300 Boulevard Sébastien Brant 67400 ILLKIRCH - FRANCE

EU_FT-ICR_MS Horizon 2020 Project FAIR DMP - V 0.3

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

EU_FT-ICR_MS Data Management Plan

This text is an initial draft of the Data Management Plan of the EU_FT-ICR_MS project
Version 0.1 Date 24 May 2018

Points still to be detailed / things to be done
- finalize 2.4 point about open-data policy and licences
- discuss the DMP in the project team
- finalize text

- Data generation and collection is required in this project in the scope of **keeping a record** of the measurements and experiments performed within the EU_FT-ICR_MS project.
- The project aims at creating an infrastructure which enables the european scientific community to discover, use, and work on FT-ICR MS technology; the DMP allows this community to store, access and analyse data, with the purpose of **improving data quality, interoperability, user open access, reproducibility, data mining,** etc...This implies the collection and conservation of all the produced data. *(See details of project in the GA p.135/308 (PartB - 1: Excellence))*
- This Data Management Plan intends to describe a set of rules allowing the research data to be **findable, accessible, interoperable and reusable** (FAIR), and is developed under the H2020 F.A.I.R. Open Data Guidelines.
- Stored information consist in **raw data, result data, metadata** and **programs**.
    - **Raw data** are the data produced by the measurements using the various spectrometers and physical instruments featured in this project, without any further processing. These are usually binary files as well as spectrometer specific output files, and are kept such as.
    - **Result data** constitute the final results obtained from the processing and analysis of raw and metadata. They may consist in binary data, text files, pictures, or any format that would best describe the final result.
    - **Metadata** are descriptive data associated to raw and result data, adding extra information about the data generation conditions (acquisition and processing). It can contain information about the sample(s), the operator(s), the experiment type, the instrumental set-up, the processing software, as well as any other important information to interpret the content of the raw and results data.
    - **Programs** are used to store, recover, analyse and display the various data. As they are used to access and manage the information, they are integrated

in the DMP.

- Some **data** might be re-used from existing repositories.
- Data originate **from physical measurements, sample description, and may be generated through forms or automatically.**
- Expected size of each data entry is between 1MB to 10GB. Each entry is composed of several sub-entries.
- These data are be useful to:
    - **users of the infrastructure** in order to store, recover, process, analyse the results of measurements.
    - **users external to the infrastructure** to compare their data with the one acquired in this project.
    - **scientists in the Mass Spectrometry** domain to deepen the global knowledge in this scientific domain.
    - from a global point of view to **scientists in biology, chemists or physics,** to gather information on the various systems studied by the infrastructure scientists.
- This plan is meant for the initial infrastructure project and will regularly be updated to follow the advancement.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

- Findability is insured by the use of **keywords** describing the instruments, experiment, sample and data type. The keywords, as well as all other metadata, should be **organised** in a rational manner and fully **searchable**.
- Metadata provision **follows metadata standards** for most of the data, each research domain having its own well documented and described standard. The available platform for data repository and access includes an open Document Type Definition (DTD) describing the structure and the possible content of the metadata. The main structure of the document is described in order to be used as a template; however it is possible to add fields in the files. A method is set up to collect metadata from automatic devices files - like the **Bruker *.d file**.
- Persistent open data are provided with a **DOI** identification. These data include raw and result data, with all the associated metada.
- The file names should contain the date (ex. format **YYYYMMDD**), the **place** where the experiment took place, the **experiment type**, the **sample id**, and the **handler id**. It thus follows the naming convention: YYYYMMDD_PLACE_TYPE_SAMPLEID. This kind of convention allows to get more information about the experiment, sample or handler from databases containing the related details.
  The **keywords list** should be meticulously generated as it is not modifiable and

should contain all the words that could be used by scientists to describe their data. It has to contain the most complete list of vocabulary used in the domains included in the infrastructure.
The approach towards searchable keywords is implemented using a tool such as **iRods** or equivalent with an automatic search **through metadata**.

- A complete **versioning** system is implemented based on different approaches according to the data type.

1. For evolutive data like programs, or specific metadata (samples, operators, procedures, etc), a follow up of their updates is kept and each evolution step recorded, using a **versioning** system like **git** or **hg**. It gives to each program iteration **a version and a revision number** . All modifications are trackable, with information about when, why and by who these modifications have been made.

2. Regarding the **binary data** status:
   1. **raw data** are unique as they come from one experiment, and remain untouched. No versioning system is used.
   2. **result data** are obtained through the processing of raw data. The obtained result is definitive. There is no multiple version of a result: if a new processing is performed on the same raw data, the new result is stored as an independent result, next to the previous one. Therefore, the only requirement for result data is the tamper-proof one. To achieve such thing, the process is compliant with the ***European GMP Annex 11*** **regulation** - equivalent to the US 21-CFR part 11. An **audit trail** is produced with the data and contains the hash of the result data.

- The **metadata** generated are a combination between the metadata automatically **generated by the instrument** and the metadata **collected by the user,** as described above. **Known standards** are used as much as possible, and new formats are generated when needed, with an emphasis on documentation, understandability and useability. In any case, new formats should follow common standards in similar disciplines.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

- **All the data generated** within the project framework are meant to be **openly available.** However some **restraints** might be applied (for instance in case of embargo delay) and **exceptions** may occur (for example when data privacy is required or in case private companies are involved). All restriction cases are detailed in the Grant Agreement (GA) under a specific **licence agreement**. The re-use of data should respect the licence specificities.
- The data are available through a network **access using a login** for consultation and download, respecting authorized rights.
- To access data, **access tools** are developed, including a platform gathering data access, download and processing**.** An open documentation is provided.

- The **raw data, result data**, associated **metadata** and documentation are stored on the developed EU_FT-ICR_MS infrastructure **platform** - or on specific **repositories** dedicated to it.
  *The documentation might also be available with the relevant open-source part of **code** associated on - versioned - public repositories like github.*
- The login id is used to determine the access rights of the user and the privacy of the data produced by this user on the plateform. The **restrictions for access** are regulated by an **authentification system.** This is directly included in the data management tool built by Casc4de based on **EGI** models.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

- Interoperability is the **main objective** of the project. To ensure it, the data and metadata formats used are **standardized**. Metadata are stored in standard formats, and the vocabulary is defined either using a DTD (Document Type Definition, which remains open for adding new fields but has a static structure), or using the document provided by the instrument constructor, related to the parameters acquisition. Basic language is english.
- **No complete ontology** is available for the FT_ICR_MS domain. Existing ontologies specific to several dedicated domains are thus used. For instance:
  - General scientific Vocabularies : http://www.ontobee.org/ontology/ERO
  - Controlled Vocabularies from Proteomics Standards Initiative : http://www.psidev.info/groups/controlled-vocabularies
  - Mass Spectrometry Ontology : http://www.ontobee.org/ontology/MS
- In any other cases, a vocabulary can be created bottom-up, from the requisite of the intervening users. **No mapping** is planned at the moment in order to allow inter-disciplinary interoperability.

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

Most information below were extracted from the Grant Agreement:

- Regarding the digital research data generated in the action (‘data’), the beneficiaries must:

1. **deposit in the data repository** and take measures to make it possible for third parties to **access, mine, exploit, reproduce and disseminate** — free of charge for any user — the following:
   (i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
   (ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan' (see G.A. Annex 1);
2. provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — when possible — provide the tools and instruments themselves).

- Unless the Commission requests or agrees otherwise or unless it is impossible, **any dissemination of results (in any form, including electronic) must:**
  1. display the EU emblem
  2. include the following text: *“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731077”.*
  3. When displayed together with another logo, the EU emblem must have appropriate prominence.
- All project members will follow the commission recommendations and the **Code of Practice** regarding the **intellectual property** in knowledge transfer activity. This recommendations are uniformed to make sure all members follow the same rules concerning **open access policy and licenses**.
- Softwares and database will be protected using licenses and **author rights**. The **Creative Common Attribution** CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/) license will be applied to the database, as all the data created within the EU FT-ICR network during the four years of the project have to be accessible. This license fits the Grant Agreement requirements.
- The produced data will be in free access between the different project members and a knowledge transfer will be performed on regular basis regarding their exploitation (used tools, scientific results and publications).
- Each beneficiary **may grant licences** to its results (or otherwise give the right to exploit them), if this does not impede the access rights under Article 31.
- In addition, **exclusive licences** for results may be granted only if all the other beneficiaries concerned have waived their access rights (see Article 31.1).
- Requests for access may be made — unless agreed otherwise — **up to one year after the duration action period** set out in Article 3 — the duration of the action will be 48 months as of 1 January 2018 (‘starting date of the action’).
- The beneficiaries must give each other access — on a royalty-free basis — to results needed for implementing their own tasks under the action.
  The beneficiaries must give each other — under fair and reasonable conditions (see Article 25.3) — access to results needed for exploiting their own results**.**
  "*The consortium will also push toward having users allowing access to all the raw data, even those not leading to publications, acquired under the TransNational Access provision after an embargo period."* - *extracted from p.111/308 of the Grant Agreement*.
- To ensure the **quality of data** several processes are used.
  1. to avoid losing data, a **backup** system is set up - both on-site and off-site.
  2. to provide high-confidence data, they are **hashed** -SHA or MD5 - and **signed** - to prove their source. All data are eventually encrypted.
  3. for processing results there, an **audit trail** system is provided to insure **tamper-resistance**.
  4. for **reproducibility** purposes, all conditions, including experiment settings, process parameters and processing software are stored alongside the raw and result data.
- There is no limited period to reuse the data, but a limited period to access them. *(See above)*

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

- The **costs** associated to the DMP are managed either at the **global level** by EU, or on a **local level**, by the PIs of each sites.
  For this reason, they are not detailled on this DMP.
- The company **CASC4DE**, part of the project consortium, **is in charge of the elaboration** of the DMP by coordinating the general requirements of the network, as well as the specific requirements on each site, and by managing the redaction of the DMP.
  Each site has to suggest a **local contact** in charge of the implementation of this DMP. This person also has to check that the local requirements are correctly transposed into the DMP.
- A **long term preservation** of the data and of the associated programs allows users to access their data even if modifications in the different laboratory set-up take place, or after a laboratory delocation, insuring a long term conservation of the scientific results. The long term preservation over the whole network also allows to gather information across the different scientific fields covered by the network, and ables the construction of a global FT-ICR data-base, from which additional results might be mined.
  This requires specfic funding for maintaining the **staff** and the **material**. The main cost arises from hiring **people** to manage the data preservation and doing the **maintenance**.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

- All data are **electronically signed** when created with a hash key such as SHA or others, allowing to detect any tampering of the information.
- The **data recovery** is ensure by a **multiple backup system**. Data samples are duplicated in the laboratory where they have been generated: on an **on-site backup** (preventing the data loss in case of deterioration of the first dataset), and on an **off-site backup**, using the file system that will be created for the project. This prevents risks of losing data when a technical problem arises in the laboratory collecting the data.
- For security reasons, data transfers follow the **specific transfer protocols** - *e.g: x509 certificate* - required by the EU.
- The data access is also secured thanks to the use of a VPN network: the users will have to connect first to the VPN network in order to access and manage the data. Doing so, the network is crypted and isolated. Each member can connect to this

network using a certificate, specific to his/her machine, and provided by CASC4DE. The network itself will be protected by a global certificate provided either internally or from an independent authority (Let’s encrypt for instance). This last option is the preferred one but the compatibility with the VPN network still need to be checked. Besides, different access contents will be granted depending on the certificate. The VPN network will also allow to check and regulate the members activity using the log files. Indeed, every time a machine connects to the VPN network, the information is kept in a log file, through the related machine certificate.

Cf. *www.egi.eu*

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

- The data handled in the project **do not imply any ethical problem** to consider except in particular cases such as **patient data** or **cultural heritage**.
  In these specific cases, the user producing the data will have to clear all ethical problems with the appropriate method determined for each scientific domain (for instance **anonymization** of patient data) **before uploading and making available** for open access these data. These curated data can then be uploaded with no particular treatment as long as the curation is adequately mentionned in the metadata.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Question sans réponse.