



NewFrac Training Network



Deliverable 1.4 - Data Management Plan

Marie Skłodowska-Curie Actions (MSCA)

Innovative Training Networks (ITN)

H2020-MSCA-ITN-2019

Grant Agreement n° 861061

Document Description and Properties

Document ID	H2020-MSCA-ITN-2019 Grant Agreement Nr. 861061
Document Title	D1.4 - Data Management Plan
Contractual date of delivery to REA	Month 6
Lead Beneficiary	US
Editor(s)	Vladislav Mantic José Reinoso Israel García
Work Package No.	1
Work Package Title	Management and Financing
Nature	Other
Number of Pages	9
Dissemination Level	Public
Contributors	US
Version Nr.	2

1. INTRODUCTION

This is the first version of the NewFrac Data Management Plan (“DMP”). We are now at the beginning of the project activities and so it will try to outline the datasets that will be produced within the project and the storing, accessing and documentation policies to be applied to them.

As the DMP is intended to be a living document, it will be updated during the project progress and when significant changes occur.

2. DATA SUMMARY

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The overarching objective of the NEWFRAC network is a high-level training of a new generation of creative, entrepreneurial and innovative early-stage researchers (ESRs) through the development and engineering applications of a new modelling framework focused on the prediction and analysis of multi-field fracture phenomena in heterogeneous engineering systems at different scales.

Trough individual research projects (IRPs), ESRs will develop new failure-predictive computational tools, entailing multi-field, multi-material and multi-scale fracture analyses, and apply them to relevant problems in strategic industrial sectors like Energy, Health and Transport.

To this aim, the NEWFRAC network will integrate two recent strategies of high impact for fracture modelling: Finite Fracture Mechanics (FFM) and a variational approach to fracture referred to as Phase Field (PF) approach, as well as novel methodologies for a consistent hybridization of strategies. These strategies will enable simulation of complex fracture phenomena in different engineering applications which was difficult to achieve using previous methods.

What types and formats of data will the project generate/collect?

The main data generated by the NewFrac project could be summarised to computational codes self-developed, calculations and experimental data.

The principal computations will be carried out using finite element codes, in particular FEniCS and Simulia Abaqus. These finite element codes are intensively used by the main researchers composing the original consortium. Abaqus user subroutines and FEniCS formulations will be generated and shared between the participants. These files are mainly code files, in different programming languages as C++, Fortran and Python, in addition, input files for models will be also shared, which will enhance the collaboration between the ESRs.

The computation of the models employed during the project is expected to generate a huge amount of results. The storing of this data for an adequate and working sharing will be one of the challenges of the project, but it will allow the ESRs to avoid the repetition of work.



The experimental data will be shared using the standard formats.

Will you re-use any existing data and how?

It is encouraged to make existing data available for research within NewFrac Consortium to be used as a start point. In order to ensure the scope of the investigation, it's very important to have a statement of previous tools for the project to define the start and the end of the data obtained.

What is the origin of the data?

The origin of the data which will be used as a start point of each ESR is self-developed codes and institutional repositories of the beneficiaries. The source of the data will be defined by each of the partners.

What is the expected size of the data?

The expected size of development codes is about 10 GB including version control and about 1 TB for numerical results. These file sizes are heavy and complex to manage, but common in our sector so we are familiar with its handling.

To whom might it be useful ('data utility')?

The value of this research will be useful for the scientific community in general and for the industrial sector in particular. Due to the nature of the research to be carried out by the network, it might have direct applications in industrial sectors such as Energy, Health and Transport, among others.

2. FAIR DATA

2.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

The data produced and/or used in the project will be uploaded to public repositories. This will only apply to data which is not protected by confidential rules. The project will use Zenodo and GitLab as main repository platforms. This repository assigns a persistent identifier. The beneficiary/partner responsible for the correct upload for each portion of data will include metadata to make easy to find and identify the data by other researchers.

Organization of the data will be arranged using two levels:

- A kernel repository containing the main developments of the project after a careful verification and validation purpose.
- A lower ESR-project data, where the specific developments for each ESR activity can be checked.

What naming conventions do you follow?

Datasets will be named as follows: NEWFRAC_DSN_DST_TT_PAR_vX_YZ_YYYYMMDD.EXT, where DSN is the name of the dataset, DST is the name of the datatype, TT is the task number, PAR is the partner



acronym, X is the version major number, YZ is the version minor number and YYYY, MM and DD are the year, month and number respectively. EXT is the file extension.

Will search keywords be provided that optimize possibilities for re-use?

Typical keywords in the discipline of fracture mechanics will be used in the context of the project.

Do you provide clear version numbers?

Yes, we will provide a major version number with a digit, and a minor version number with two digits. The major version number will be zero for data which are not consolidated. Once the data are consolidated and revised, the major version number will be 1 or higher.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

We will generate metadata for each dataset, in order to improve searchability. We will adopt the Dublin Core Metadata standard. This standard includes fifteen elements that introduce a vocabulary of concepts: 1) Title, 2) Creator, 3) Subject, 4) Description, 5) Publisher, 6) Contributor, 7) Date, 8) Type, 9) Format, 10) Identifier, 11) Source, 12) Language, 13) Relation, 14) Coverage, 15) Rights. Every research in the context of the project will include this metadata in the dataset made public.

2.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

In order to maximise the impact of NewFrac research data, data will be shared within the consortium. Selected data and results will be shared within the scientific community through publications in scientific journals and participation in international/national conferences as well as through open access data repositories as ZENODO and OpenAIRE (European Commission).

Pre-existing code libraries, codes already existing at the starting point of the project don't have to be shared openly. This addresses codes that have been developed in past internal projects or codes that have been implemented by external research partners exclusively for a beneficiary will not necessarily be shared.

The NewFrac Database is first stored and organized in a database by the data owner (personal computer or on the institutional secure server) and on the private project database (Access through the project website <https://www.newfrac.eu/>).

How will the data be made accessible (e.g. by deposition in a repository)?

Data will be made accessible by deposition in a private repository (GitLab and Teams) and institutional repositories of the beneficiaries (such as E-prints for IMT, IRIS for POLITO and idUS for US, in addition to arXiv). Finally, the data will be made available in Zenodo, which allows to fulfill the four FAIR requirements.



What methods or software tools are needed to access the data?

In general, all the data will be accessible using open formats. The only exception previewed is the usage of ODB files, which are the native output files of Simulia Abaqus, one of the main simulation software that will be used. Even in this case, we will try to export the results to VTK format, in order to make them accessible,

Is documentation about the software needed to access the data included?

In general, we will use simple formats, to make the data as accessible as possible. A README file will be included in those cases more complicated.

Is it possible to include the relevant software (e.g. in open source code)?

In some cases, it is possible. We will include them in those cases where it is possible.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

We will use Zenodo as the main and general repository because it fulfils the FAIR guidelines, and it is available.

Have you explored appropriate arrangements with the identified repository?

Zenodo is open for research projects

If there are restrictions on use, how will access be provided?

There is an administrator of the repository who will give access to the involved persons. Also, there is limit access to the private Microsoft Team Project managed by Coordinator.

Is there a need for a data access committee?

No need for a data access committee, all ESR have access and consequently all the beneficiaries.

Are there well described conditions for access (i.e. a machine-readable license)?

According to the information provided by Zenodo, the open access status and a license are embedded in all metadata formats (i.e. machine-readable).

How will the identity of the person accessing the data be ascertained?

In the private repository for sharing data between members of the consortium, an user account will be provided for each member. In the main public repository (Zenodo), the data is accessed without need of identification.

2.3 Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much



as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Data format is totally interoperable between members, allowing it to be exchange and re-used. The principal idea of sharing codes, results and data is being interoperable and comparable. Each participant will be familiar with the computation used and shared in the repositories.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

We will adopt the Dublic Core Metada standard. This standard includes fifteen elements that introduce a vocabulary of concepts: 1) Title, 2) Creator, 3) Subject, 4) Description, 5) Publisher, 6) Contributor, 7) Date, 8) Type, 9) Fromat, 10) Identifier, 11) Source, 12) Language, 13) Relation, 14) Coverage, 15) Rights. Every research in the context of the project will include this metadata in the dataset made public.

In the private context, we will use a common set of software to make interoperable the codes, computona and experimental results developed by the different ESRs.

Will you be using standard vocabularies for all data types present in your data set, to allow interdisciplinary interoperability?

We will use the vocabulary proposed by the the Dublic Core Metada standard.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

In this case, we will generate mappings to more commonly used ontologies and we will provide in the metadata a DOI link to a guide uploaded to a open repository.

2.4 Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The data will be licensed using the Open Database License (ODbL), except in those cases where due to confidential or commercial reasons, it is not allowed. In those cases, we will look for a licence as open as possible, and compatible with the restriction's nature of the dataset.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

In general, data will be available once the scientific publications or patents are made available.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

It will be.

How long is it intended that the data remains re-usable?

According to the CA, a request for Access Rights may be made up to three years after the end of the Project or, after the termination of the requesting Party's participation in the Project, provided however



that a request for Access Rights either from TAU or by TAU may be made up to one year after the end of the Project or after the termination of the requesting Party's participation in the Project.

Are data quality assurance processes described?

Each partner will be responsible for the data management they provide, assuring the quality and reliability of their contribution.

The NEWFRAC Technology Transfer Lab (TT Lab) will also constitute an ideal platform for the valorisation and exploitation of IP. The TT Lab will be equipped with the computational tools developed in the framework of the project and all case studies investigated will constitute demonstrators for the proposed technology. Being planned, the creation of one start-up after the IP protection of the results carried out by the TT Lab.

3. ALLOCATION OF RESOURCES

What are the costs for making data FAIR in your project?

There are no costs associated to make the database FAIR and long term preserved as planned at the moment, the repositories we will use are preferentially free access. However, depending on the size of the repository, alternatives plan for using such repositories will be chosen. Pricing for multiple teams in Gitlab for instance is of about 19 USD/user/month.

Who will be responsible for data management in your project?

Each of the partners will be responsible for the data management they provide, being Universidad de Sevilla the coordinator of the access and participation.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

The resources for long term preservation are not discussed yet. As indicated on section two of this deliverable, the data obtained on the research must be clearly defined to decide what is going to be preserve long term.

4. DATA SECURITY

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The following guidelines will be followed in order to ensure the security of the data:

- Store data in at least two separate locations to avoid loss of data.
- Limit the use of USB flash drives.
- Label files in a systematically structured way in order to ensure the coherence of the final dataset.

All project deliverables and data will be stored and shared in a private repository restricted to the project consortium. As an initial step, only the Consortium Partners will have access to the cloud storage where dataset and metadata are filed. Following, scientific publications and articles, the dataset deliverables and the final demonstrator research results will be shared through ZENODO, GitLab and OpenAIRE and other database to promote the data making FAIR.

Is the data safely stored in certified repositories for long term preservation and curation?

Zenodo and OpenAIRE are certified repositories which complies with a long-term preservation of the data.

5. ETHICAL ASPECTS

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

The NEWFRAC network objectives do not involve experimental work with human embryos/foetuses, humans, human cells, personal data or animals. This network will not be used for military or malevolent applications, not harm the environment and animals in any way, and not use elements that may cause harm to humans.

As a part of WP5 experiments will be conducted on cadaveric femurs at the experimental biomechanics lab. These femurs are purchased from NDRI (National Disease Research Interchange <http://ndriresource.org/>) in the USA, a company that is certified to provide such specimens. The Scientist-in-Charge, Prof. Yosibash has been performing such experiments for the past 15 years and he has a formal approval from the TAU Ethics Committee that approves the standards and methods used in these experiments.

Therefore, NEWFRAC does not implicate any research objective with ethical issues.

6. OTHER ISSUES

**Do you make use of other national/funder/sectorial/departmental procedures for data management?
If yes, which ones?**

Each of the partners will follow their national and institutional procedures for data management, in addition to this NewFrac DMP.

7. CONCLUSIONS

This deliverable describes the main principles and guidelines for the Data Management for the NewFrac project, which will be updated during the lifetime of the project when any new situation occurs.

NewFrac team is totally involved with FAIR data principles and they will be chased during the project duration with no exception.