Horizon 2020 Work programme
Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy
Call
H2020-FNR-2020: Food and Natural Resources
Topic name
FNR-16-2020: ENZYMES FOR MORE ENVIRONMENT-FRIENDLY CONSUMER PRODUCTS

FuturEnzyme:
Technologies of the Future for Low-Cost Enzymes for Environment-Friendly Products
Final ID: 101000327

26/11/2021

FUTURENZYME REPOSITORY FOR DATA STORAGE AND MANAGEMENT

D8.5

VÍCTOR GUALLAR
BSC
Calle Jordi Girona 31, Barcelona, 08034,Spain

Document information sheet

Work package:	WP8, Communication, Dissemination and Exploitation
Authors:	BSC (Víctor Guallar), CSIC (Manuel Ferrer, Patricia Molina)
Document version:	1
Date:	26/11/2021
Starting date:	01/06/2021
Duration:	48 months
Lead beneficiary:	BSC
Participant(s):	ALL
Dissemination Level:	Confidential, only for Consortium members (including the Commission Services)
Type	Websites, patents filling, etc.
Due date (months)	6
Contact details:	Victor Guallar, victor.guallar@bsc.es

Summary

FUTURENZYME REPOSITORY FOR DATA STORAGE AND MANAGEMENT

1. Scope of Deliverable

This deliverable consists in an internal repository/database created ad hoc in the Consortium with free access to Consortium members, and accessible through the project webpage (www.futurenzyme.eu) and BSC/CSIC (ssh (or scp) routes: username@dt01.bsc.es), for storing and managing all raw data and research datasets that are the result of the activities developed in the project. The datasets will be stored in appropriated formats allowing readability and accessibility; for that, all partners will agree on preparing the data according to the guidelines, described in deliverable 8.4 (Data Management Plan). Those raw data and research datasets associated to scientific publications will be transferred to public repositories (mainly FuturEnzyme's Community at Zenodo) once the publications become publicly available; datasets related to unpublished work will remain in the internal repository, unless authors decide to include them in Zenodo.

2. Decision-taken strategy to publish and to manage the data/metadata

When a whole-sense pool of data is obtained by one/several partner/s, the decision to disseminate and communicate or protect the results has to be taken (**Figure 1**). If publication is chosen, open access (OA) has to be guaranteed (preferably gold, green is possible) in order to accomplish Open Research Data (ORD), following the principle of "as open as possible, as closed as necessary". In this case, the article will always be uploaded to the corresponding repository (see below for further details), even when gold access is selected.

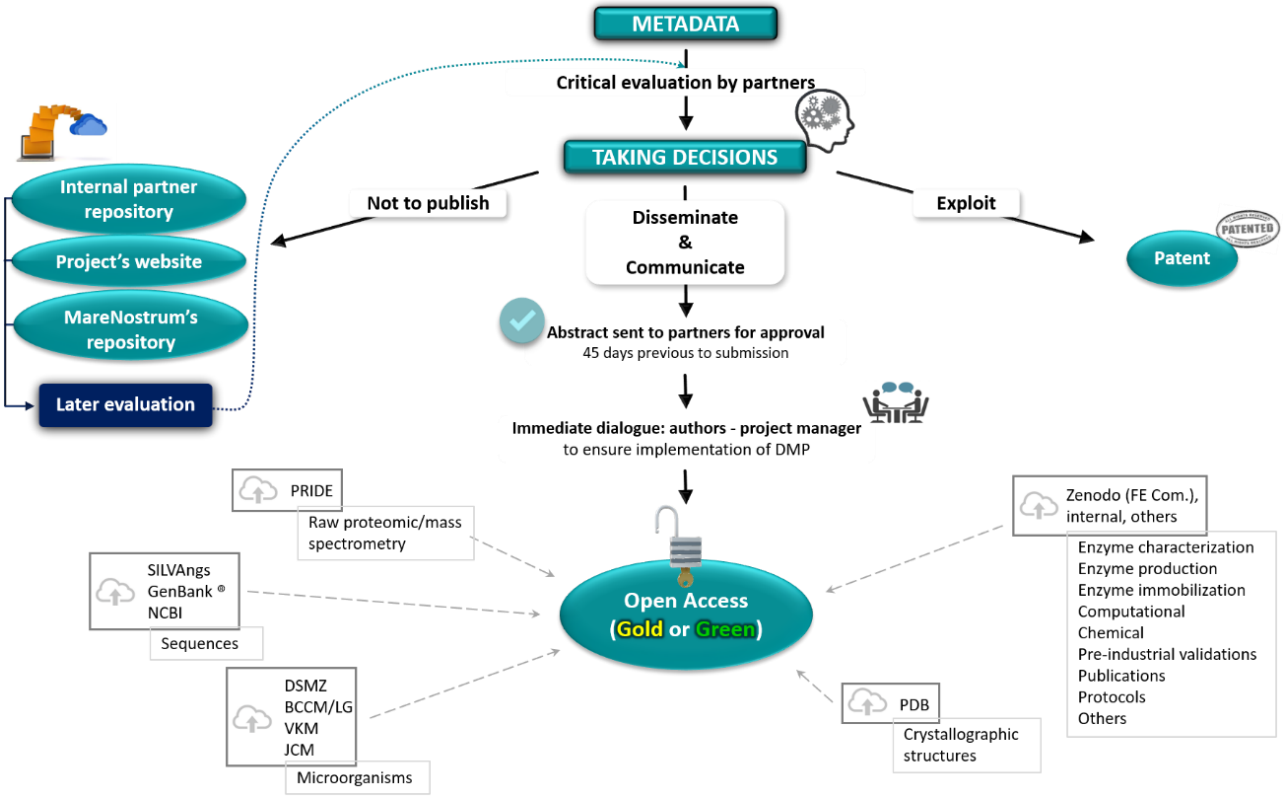

Figure 1. Diagram followed after whole-sense research and data are achieved to decide the end of the acquired know-how. Adapted from H2020 online manual, and submitted Deliverable D8.4.

The data generated from experimental and computational work will show the performance of the enzymes/microorganisms in terms of identification and characterisation following the templates created for the project (see Annex). These documents will be filled with as much information as available and be as complete as possible, always including the researcher's identification in charge of the analyses or assessments. This information plus the results from the Life Cycle Assessments (LCA), bibliographic (academic

and patent), consumer and market evaluations, etc. produced in the frame of FuturEnzyme will also be made available in the appropriate Consortium's repository depending on the needs and decisions taken by the authors (see section 2).

In all cases, every time an entry is uploaded to any repository, the name of the action, acronym and grant number, as well as the DOIs and permanent identifiers, have to be added, so they can be easily sent to the national and European agencies upon request at any time.

All partners are free (and encouraged) to include publications and data also in public repositories from their institutions (e.g. Digital CSIC, https://digital.csic.es/).

Regarding data that is not meant to be published nor protected, it is the researcher's decision whether to upload it in the repositories or not. It is strongly recommended to do so in order to achieve FAIR and RRI principles.

With this strategy, the project results will meet Open Access and Open Science status.

3. Repositories for FuturEnzyme

The information, knowledge and data that is going to be generated in the frame of the project has to be handled in order to accomplish FAIR (findable, accessible, interoperable and reusable) and RRI (Responsible Research and Innovation) principles by the procedures described in the Deliverable D8.4 (Data Management Plan).

There will be 3 main repositories/databases for managing FuturEnzyme's data, metadata and information:

- The private area of the project's website.
- The repository created in MareNostrum 5 Supercomputer (BSC) for FuturEnzyme (ongoing).
- The Zenodo Community FuturEnzyme.

These repositories are detailed below.

3.1. Private area of the project's website

To access this intranet, a user and password only delivered to Consortium members are needed. After login with these credentials at www.futurenzyme.eu, the page shown in **Figure 2** appears. In here, files will be uploaded via the Project Manager (Patricia Molina) and the website developer, with whom a Non-disclosure agreement (NDA) has been signed. A section inside this private area (*Private data storage*) has been set up to include (via the Project Manager) those files containing information and data that the partners want to upload but that they decide not to make visible for other Consortium members (e.g. not fully characterised enzyme, preliminary results, etc.). Other partners can see that a file has been added, but cannot see the contained information (a password will protect the file). Once the authors decide to share such files with all the Consortium members, they will be transferred to the *Shared data section*.

In the *Shared material* section, files such as videos for protocols, visual identity images, project's logo, etc. will be made available for the Consortium members to download. Any member is free to upload material to this section (via the Project Manager), considering that it is relevant and appropriated for the project.

The *Shared templates* section includes those templates created ad hoc for the project to produce reports, deliverables, presentations, etc.

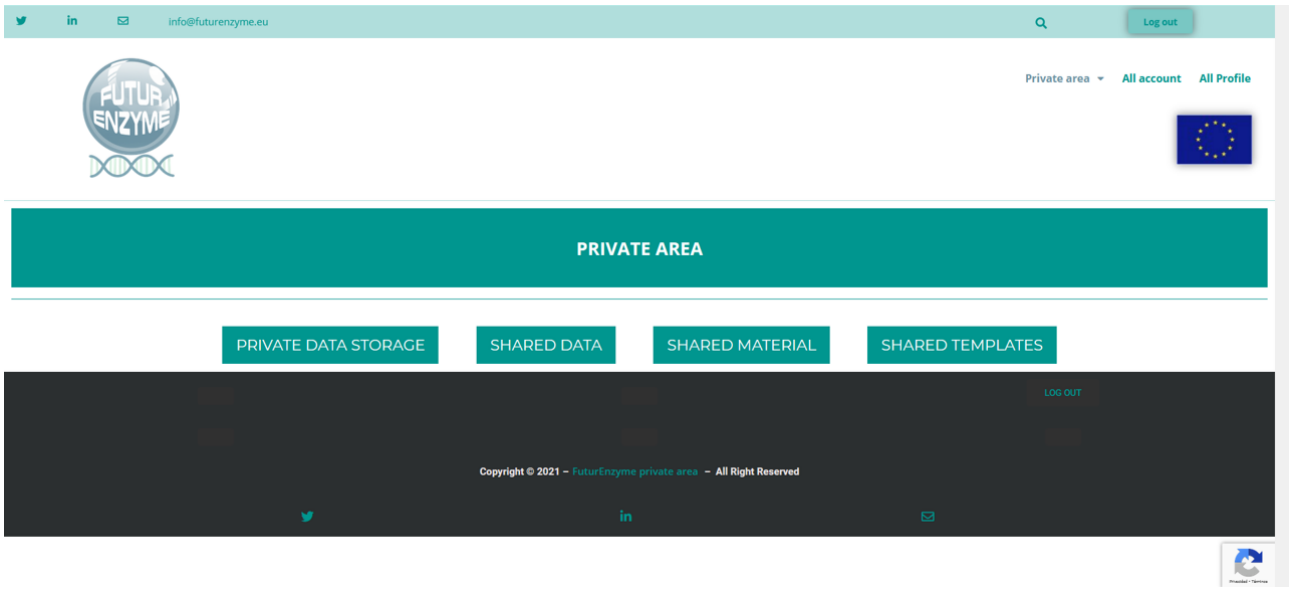

Figure 2. Private area of the FuturEnzyme's webpage. Four sections have been created to organise the material to be uploaded. In the menu, pages containing information on the partners' uploading have been included. The access is exclusive for FuturEnzyme members.

3.2. Repository created in MareNostrum 5 Supercomputer (BSC) for FuturEnzyme (ongoing)

The idea of this repository is double. First, to serve as backup for all the information to be deposited in the private area of the project's website. Second, to have a space to store information related to the project. The type of information/data would be sequencing, computational, biochemical, structure, experiment, etc. Each partner will have at its disposal 1-2 TB, with a total of 15 TB for the full Consortium.

The Backup data is hosted in Agora, the storage system deployed by BSC in 2020. BSC storage is composed of two different media types: hard drive disks (high cost, low latency) and tapes (low cost, high latency) for medium/long-term storage. The file management is provided by a Hierarchical Storage Management (HSM) software that automatically moves data between high-cost and low-cost storage media. The HSM system is configured for a transparent migration between storage layers. Secure Shell (ssh) enables secure transfer and access to the backup data only to authorized users.

The storage space is granted free of charge offered by the European Commission, provided by the EOSC project DICE https://www.dice-eosc.eu/

In order to minimise the number of interlocutors, only three persons will have access to Backup secure transfer and access: Victor Guallar (victor.guallar@bsc.es, BSC), Ana Robles (ana.robles@bsc.es, BSC), and Patricia Molina (patricia.molina@icp.csic.es, Project Manager FuturEnzyme, CSIC). These three persons will receive from partners the datasets and information to be stored in the "storage system". This information may include any type of datasets/information detailed in the Deliverable D8.4 (Data Management Plan).

Connections will be made with the following ssh (or scp) routes: username@dt01.bsc.es, where username will correspond to any of the three authorized users

3.3. Zenodo Community FuturEnzyme

Publications that are the results of the project activities and files containing information that supports such publications will be made available in the Zenodo's public repository (www.zenodo.org), precisely in the FuturEnzyme's Community (https://zenodo.org/communities/futurenzyme/; identifier: futurenzyme; **Figure 3**).

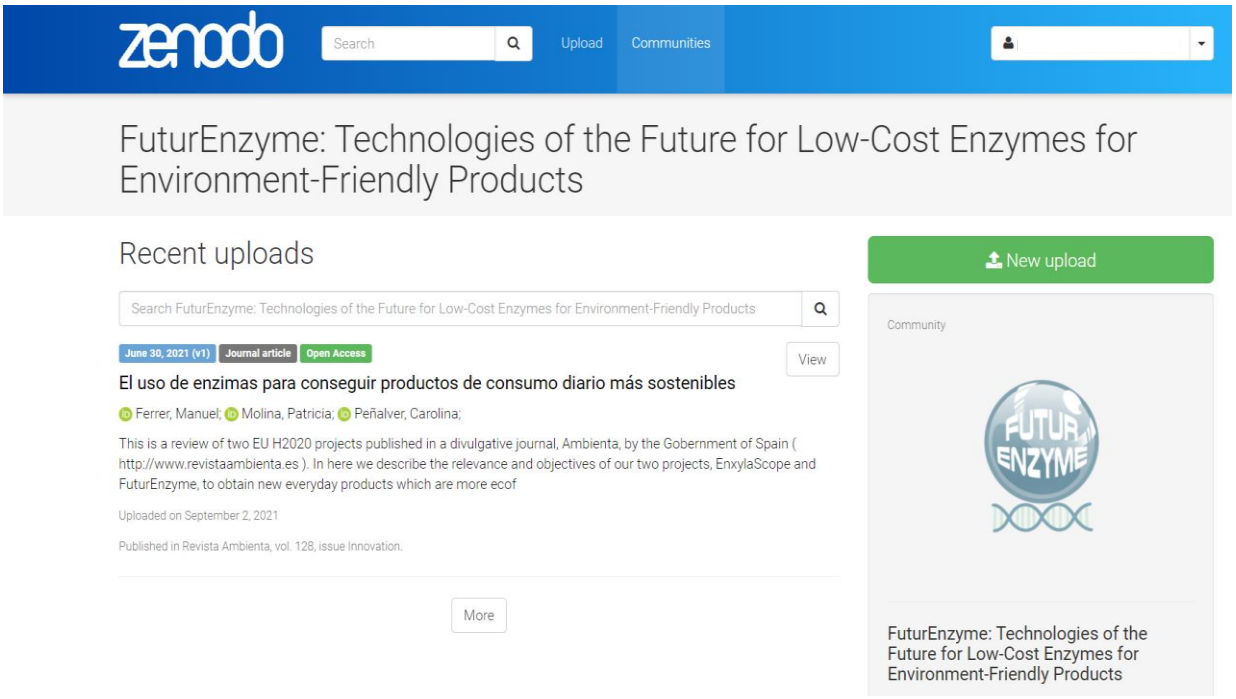

Figure 3. FuturEnzyme's Community page at Zenodo.

Every partner responsible for the publication and files will upload the entry (https://zenodo.org/deposit/new?c=futurenzyme), selecting the type of data, title, authors, grant number, type of access, and all the information requested that is available. This makes it easy and without charge for anyone to access the whole pool of information, if/when it is embargo free. The FuturEnzyme's Project Manager will revise and accept/cure/ask for modifications to the partners who upload material. Other categories that can be uploaded to Zenodo when appropriate are: poster, presentation, dataset, image, video, software, lesson, and other. The access rights that can be chosen to match the different needs of the researchers who upload the material at any moment are: open access, closed access (access only by owner), embargoed, and restricted. All uploaded data is given a digital object identifier (DOI) number, and the researcher has to include keywords to optimise possibilities for reusing the information. Besides, Zenodo allows for Creative Commons license. In addition to the functionalities that Zenodo represents as public repository, this platform is integrated into reporting lines for research funded by the European Commission via OpenAIRE (www.openaire.eu) in collaboration with CERN ("Conseil Européen pour la Recherche Nucléaire", or European Council for Nuclear Research, Geneva, Switzerland) with the aim to support, boost and measure the correct application of the European policies regarding open access to scientific publications and data.

4. Data storage and management

4.1. File extensions to be used for data and metadata storage in FuturEnzyme's repositories

Scientific datasets from measurement tests/computational analysis will be stored in the original format of the file and in spreadsheet (*.xlsx), word processor (*.docx, *.txt) or a similar format of common use.

Posters, leaflets, brochures, reports, presentations or any other similar material will be stored produced in Portable Document Format file (*.pdf).

Videos will be created as .mp4 file.

Related files that need to be used together will be compressed (*.zip), including a Readme.txt describing in English the content and how to retrieve the information.

4.2. Standardisation of file names

When creating a file for the work done/to be done, there is always a discrepancy in how to name it, moreover when people from different groups and even countries are involved. For this reason, all the files produced by a member of the Consortium in the context of FuturEnzyme activities that will be uploaded to any of the repositories will be named following this pattern:

FE_Partner name_Deliverable/Task number_Type of material/data_Name_Subname_Version number

Example 1: FE_CSIC_D1.1_Enzyme_EH25_mutb_v1

Example 2: FE_ITB_T4.1_LCA_Detergent benchmarking_sample 1_v2

This procedure standardises and unifies the designation of the files, helping to find, compare and track any information required.

4.3. Quality controls

Quality control of the deposited datasets will be guaranteed. For example, in case of protein structures deposited, before they can be deposited, the PDB system performs an exhaustive analysis of the quality of the experiment, with technical criteria of crystallographic analysis and control of the protein geometry. Small mismatches are retouched or corrected and resubmitted. When everything is correct, they are assigned a PDB code, but they can remain inaccessible for a maximum of 1 year or until they are published. Note that the PDB itself analyses the structure and provides a document called "Validation Report" with all the analysis.

When data/publications are uploaded to Zenodo, a curation is required before they are published, which will be carried out by the Project Manager. With this we ensure that not anyone can upload whatever to FuturEnzyme's community, and that everything that appears within it is legitimate to the project.

4.4. Data management after the end of the project's lifetime

All the data uploaded to FuturEnzyme's Zenodo Community will be permanently accessible and managed by the Project Manager at least until the end of the project. The material uploaded to the project's website and MareNostrum 5 repository (see Section 3.2) for our project will be available and managed as far as the sites are maintained (at least five more years after the end of the project for FuturEnzyme's website, permanently for MareNostrum 5). After the end of the project, a copy of the material will be kept at CSIC's facilities by the Coordinator partner (Manuel Ferrer, Patricia Molina).

5. Allocation of resources for data and metadata open access

The FuturEnzyme's website is an eligible cost included un the Grant Agreement budget and considers the creation and maintenance of the Private area.

As previously mentioned, the storage space at BSC is granted free of charge offered by the European Commission, provided by the EOSC project DICE https://www.dice-eosc.eu/.

Zenodo repository (including Creative Commons) and other institutional repositories chosen by the partners are free of charge (as detailed above, in Section 3). On the other hand, publication in green or gold open access peer-review journals has Article Processing Charges (APC). This costs that are eligible in Horizon 2020 programme as stated in the Grant Agreement Article 6, specifically in section 2.D.3: "Costs of other goods and services … are eligible …include, for instance, consumables and supplies, dissemination (including open access), protection of results, certificates on the financial statements (if they are required by the Agreement), certificates on the methodology, translations and publications." Such costs will be covered by the partner responsible of the information to be published. Table 3.4b of the Grant Agreement specifies the budget destined for open access publications and articles in non-scientific periodicals magazines dedicated to the dissemination of knowledge for scientific and non-scientific community. Whenever multiple entities are

involved in the same publication, the costs can be distributed as all the parties convene. In this table, it is also specified the cost intended for the project's website creation and maintenance is also specified.

6. Data security

6.1. Confidentiality

In order to preserve the confidentiality of the results obtained in the project, every partner will decide when to make them public (following the process depicted in **Figure 1**). The documents can be uploaded without being made public both to the private area of the project's web (via the Project Manager and the website developer, the latter under non-disclosure agreement) and to the FuturEnzyme's Zenodo Community. In the private area of the FuturEnzyme's web, a section for private storage is created as previously described to allow to every partner to upload their material and protect it with a password, that will be transferred when needed to the shared data section. Zenodo also allows for different types of access: open, closed, embargoed or restricted. Regarding MareNostrum's repository, it will only be accessible to the Consortium members, and the Project Manager will upload/manage the documents.

Since QR codes will also be generated for the datasets, lists of bioresources, etc., those documents that need to be protected will be locked by a password, only known by the project manager and the responsible of the material or the whole Consortium, depending on the specific needs.

6.2. Security copies

The Consortium members responsible for each file, dataset, etc. will store them in their own devices and regularly made security copies.

Besides, the Project Manager will keep a copy of all the documents generated by the partners that will be uploaded to the project's repositories mentioned in here. This copy will be physically stored in a hard drive, and security copies will be made regularly in two more devices.