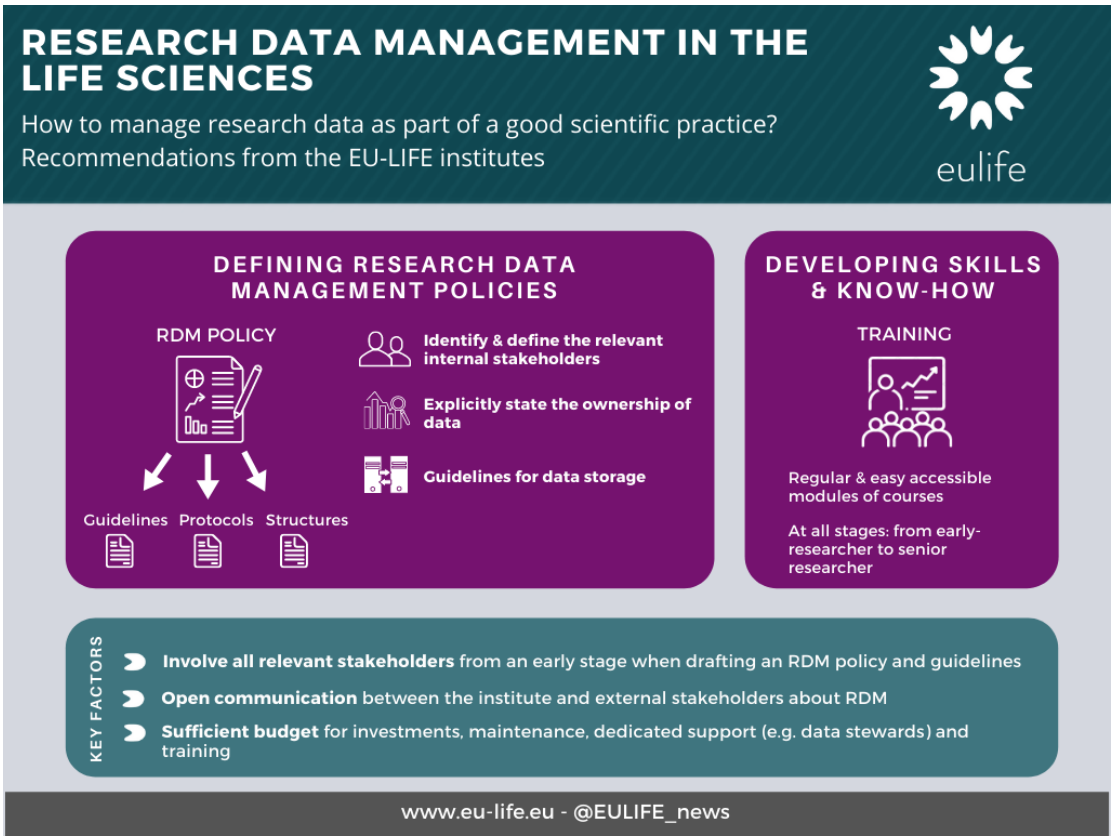
A hands-on guide for Research Data Management in the Life Sciences

Recommendations from the EU-LIFE institutes

May 2021

eulife

Table of contents



Background

Research Data Management (RDM) – including FAIR and Open Data – is an essential part of good scientific practice. It includes the whole process of handling research data, from its generation and entry into the research cycle to the dissemination, sharing, archiving, and destruction.

In order to get a better understanding of what the EU-LIFE research institutions are implementing and developing in the field of RDM, the alliance set up a dedicated task force and ran a survey within its member institutes (see Annex A for a summary of the survey). Results from the survey clearly pointed out that RDM is currently a priority and most EU-LIFE institutes are in the process of developing policies, guidelines and training. Subsequently, the task force organized a series of 3 workshops to discuss and gain further insight into the opportunities, obstacles and challenges regarding implementation of RDM workshops focused specifically on institutional RDM policies (workshop 1), RDM training (workshop 2) and the role of core facilities[1] in RDM (workshop 3). Furthermore, the workshops have been a very useful platform to share best practice among the EU-LIFE institutes.

This hands-on guide summarizes the key findings from the workshops and, building on them, it brings forward recommendations to implement RDM at institutional level. Since the organization and infrastructure is unique at each institute, there is not a single RDM model but a multitude of options from which institutes can choose. The goal of this exercise is to provide inspiration for good RDM practice.

[1] Core facilities: infrastructures and technologies within a research institutes that support research and are usually organized as units or services. Examples of core facilities are bioinformatics services, imaging facilities, genomics units, etc.

Introduction

Research institutes are faced with a rapidly increasing amount of digital data. Its efficient and (cost)-effective management has become a major challenge that is further complicated by the need for data to be safely stored over long periods while at the same time being accessible from multiple platforms.

In recent years, the concept of Open and FAIR data management (i.e. data that meet the principles of openness and reproducibility, as well as Findability, Accessibility, Interoperability, and Reusability) has gained much support and resulted in changes in the way we perform research. Open and FAIR data management is fundamental to ensure data sharing and re-use of data, including the use of data across research facilities, collaborators and different stakeholders.

Due to these developments, there has also been a growing focus on how to handle research data in the best possible way to adhere to the before mentioned principles while still maintaining a reasonable cost-level, in terms of infrastructure and human resources.

Defining and developing RDM policies

An RDM policy is a document describing the general principles and the overall process of RDM at the institute level. It includes the purpose and relevance of the policy and identifies the relevant stakeholders regarding RDM. It also provides an overview of pertinent frameworks - such as legal, GDPR (General Data Protection Regulation), FAIR, Open Science and IP (Intellectual Property) - and how the RDM policy relates to these different areas and other policies/regulations at the institute (see also Examples 1 and 2 describing RDM policies at two different EU-LIFE institutes). Building on the general RDM policy, institutes develop operational guidelines, protocols, forms and structures dealing with specific aspects of RDM. These procedures will need more frequent updating than the general RDM policy.

Different research areas produce very different types of data and thus need specific RDM guidelines. Such field specific guidelines or plans could include sections on the duration of data storage, type of data to be stored, trustworthy repositories, traceability, safety and classification of data.

Role of the different stakeholders

It is important that the policy defines the role of the different stakeholders in the RDM process within the research institute. Many EU-LIFE institutes are currently setting up steering committees on RDM. Some of these have a more technical focus and are involving experts from core facilities and IT, while others have a broader representation and include scientific coordinators, the institute direction, scientific affair officers, legal representatives, etc. The RDM committee can act as the driving force of defining and reviewing the policy and ensuring the involvement of the relevant stakeholders.

Most institutes have already established, or are in the process of establishing, dedicated professional roles on data management (as for example a data steward (see Example 1) or a data core facility, although these roles are often quite differently profiled and heterogeneous among institutes.

When defining and implementing the RDM policy, it is important to engage directly with all relevant internal stakeholders. Below we identify stakeholders at the institute level and their individual responsibility in the process of defining and implementing an RDM policy.

Board of Directors: the role of the directorship at the institute is to set the strategy and goal for implementing an RDM policy and highlighting the relevance of such a policy. They are also a crucial actor in the internal communication and enforcement of the policy, which includes promoting involvement and commitment from all internal stakeholders. Lack of adequate funding is a major obstacle for achieving the desired goals for RDM at the institutes. An essential role of the directorship is to make sure resources are available at the institute to implement the policy. This includes both funding and providing the necessary infrastructure and human resources.

Researchers: since the researchers will be working hands-on with RDM, they should play a part in defining and developing the policy and the guidelines. Furthermore, it is very important that researchers at all levels of their career receive proper training and support in order to benefit from, implement, follow and comply with the policy and the guidelines.

Core facilities: core facilities are typically a major producer of data and therefore they need to effectively manage the data that is produced and processed within their scope. It is central that core facilities play a role in the development and implementation of the RDM policy and guidelines. Furthermore, they are also key in supporting and guiding the researchers by making the data accessible and useable, and providing training.

IT department: the IT department is essential in providing the suitable infrastructure for RDM. It is vital in setting up and communicating the technical boundaries for storage and organization of data and to formulate the guidelines in particular regarding technical aspects. Often, the IT department is responsible for IT security and data integrity.

Library: some institutes have a library department. The library can function as a data repository and support the researchers in managing their data. They also possess the knowledge on open access publishing. At other institutes a dedicated data desk can function similarly to a library (see Example 4).

Legal/ethical department: this department plays a role in drafting and monitoring contracts and aligning the policy with other legal frameworks in place, including GDPR and grant agreements. The legal department also has expertise in international and national legislation and contractual obligations and can thus aid in defining responsibility and ownership of data.

Tech Transfer department: together with the legal department the tech transfer department has a fundamental role regarding IP of the data. In cases where patent applications are filed this affects level and timing of openness of data.

Department responsible for training: the department at the institute responsible for developing training and training programs (for some institutes this could involve the Human Resources department, Scientific Affairs department and/or Graduate Schools) have a responsibility for providing the necessary training for PhDs, established researchers and technical staff. The subject of support and training is further elaborated below.

eulife

Example 1: Research Data Management and Open Access Policy at the International Institute of Molecular and Cell Biology (IIMCB), Warsaw

Open Access and RDM policy

At IIMCB an Open Access Policy was introduced in 2019. The policy has been implemented to ensure access to and exchange of knowledge, facilitating broad cooperation and avoiding unnecessary repetition of work. The policy covers both instructions for open access to publications and open access to research data.

The Data Steward at the IIMCB

The Policy also specifies the appointment of a Data Steward and defines the role of the Data Steward as follows:

- *Informing about the policy*
- *Coordination of work related to the development and maintenance of the IIMCB open access infrastructure*
- *Advising on best practices according to the policy*
- *Educating the addressees of the policy within a scope of open access to publications and RDM*
- *Monitoring and reporting of the policy implementation results to the director of the IIMCB*

Furthermore, IIMCB is currently working on developing dedicated RDM policy guidelines, procedures and tools for data providers.

RDM plan

The IIMCB's researchers are typically using a template provided by the National Science Center, Poland when making a plan for RDM. They have also the possibility to use an online tool DMPonline (https://dmponline.dcc.ac.uk/). The template includes questions about the following areas:

- Data description
- Data collection
- Re-use of existing data
- Data documentation
- Legal requirements
- Data sharing
- Storage and back-up of data
- Long-term preservation of data
- Responsibility and resources of RDM

Ownership of data

An important aspect to consider is data ownership and this should be explicitly stated in the RDM policy and guidelines. Usually the data is owned by the institute, with the researcher being the immediate responsible for the management of data. An essential point to consider when it comes to ownership of the data is the difference between data produced at the institute for external users and data produced for internal users; these two cases might need to be addressed separately in guidelines. Here the legal departments at the institute might aid in defining the final responsibility.

Storage of data

It is important to include in the policy some guidelines for data storage at the institutes. This includes defining what data to store, where to store it, when to delete and importantly who has the responsibility for making the decisions regarding deletion. The IT department has a role in setting-up the proper storage space and infrastructure tailored to the need of researchers and core facilities. Considerations about documentation and naming of data can also be included in the RDM policy to allow for easy identification and re-use of data.

Implementing an ELN (Electronic Lab Notebook) either at institute or core facility level as part of data management, including storage, handling, and sharing, can be a useful tool in good RDM. It creates a uniform platform for data handling in the institute that is easily managed. An integrated ELN tool can provide a central secure environment for researchers and further encourage researchers to store data in the same place in a findable and retrievable way. A next step would be implementation of an automated data management tool for meta-data generation as well as storage rules. At the moment, two EU-LIFE institutes have an ELN in place that automatically links to services at the core facilities and stores metadata from scientific experiments.

Example 2: Policy and guidelines for Data Management at Instituto Gulbenkian de Ciência (IGC), Lisbon

At IGC an institutional policy for data management is currently under development. IGC is committed to open, rigorous, and reproducible science and the principles of FAIR (Findable, Accessible, Interoperable, and Reusable) data and RRI (Responsible Research and Innovation). In light of these principles IGC started preparing a policy for data management, trying to strike a balance between the ethical principle that research data should be made available with as few restrictions as possible, including considerations about file formats and annotation, and the intellectual property of its researchers. One guiding principle that was especially considered was unobtrusiveness to the research workflow, as the institute recognizes that this is the main obstacle for adoption of the policy. This policy aims to cover all stages of the research data's life-cycle.

The policy at IGC defines the following guidelines:

- *Ownership*: The institute is the custodian of all data produced by it is employees
- *Responsibility for RDM*: The principle investigator of a given research project is the main responsible for its RDM
- *Data management*: A data management plan shall be developed for each research project at the institute, setting out specific procedures for data management and importantly, what qualifies as significant data to be preserved. In order to alleviate burden on researchers, expert committees set out guidelines on proper annotation procedures and field-specific conventions for the most common data types.
- *Storage*: All data shall be stored at the institute's designated storage infrastructure to ensure proper backup and versioning.
- *Sharing*: Sharing of data should follow the FAIR principles. Specifically, the policy states that "*All significant data to be shared should be made publicly available within 2 years after the project's end. This does not apply to data supporting published findings, for which the data should be made available at the time of publication*". To facilitate this, the institute has created an open data repository, based on Zenodo. Furthermore, a CC-BY-NC license is recommended for all data sets.
- *Retention*: The policy states that significant data will be preserved by IGC for a period of at least 10 years

The policy for data management at IGC also states that "*Grant applications should budget for the associated costs of data storage and management for the duration of the project*" way to ensure proper funding for RDM.

Training and support

The development of skills and know-how is key to a successful implementation of RDM.

Introducing good practices for RDM early in a research career will help shape the mind-set of early-stage researchers to be more aware of the relevance of RDM in the context of Open Science and FAIR data. This can be achieved by introducing mandatory modules already as part of a master's program and transitioning into dedicated courses at the level of PhD students and postdocs. Also for senior researchers, it is central to have specific training on RDM

preferably adapted to the relevant research disciplines and data produced. For the senior researchers, the training could include how to develop a data management plan when starting a new project or funding application. Senior scientists have a responsibility to make sure that junior scientists receive the proper training. It could be of advantage to include in the RDM training a train-the-trainer principle to help facilitate transfer of knowledge between researchers. Bedsides scientists, the institutes should also offer training on RDM aimed at core facility staff, technicians and research managers when relevant. This training should be practical and discipline-specific.

To promote RDM training at the institutes, it is key to have regular and easy accessible modules or courses. This could for example be achieved by utilizing an e-learning platform.

The training of researchers in RDM, besides providing knowledge and practical tools, should underline the importance and relevance for researchers to implement the RDM guidelines. It is key that the implementation of RDM is not perceived as simply an additional workload for scientists, but instead as an integral part of good scientific practice that besides benefitting the researcher's work is also of benefit to the larger scientific community and society. This can be done by highlighting the researchers' own roles, advantages and responsibilities together with the relevance of RDM for Open Science and FAIR data. Furthermore, giving concrete and practical tools for data management that are specific for the research discipline, and indicating the practical benefit of RDM, can aid in the implementation of RDM. Finally, showing both best- and worst-case scenarios of RDM implementation can illustrate the relevance of this aspect for the researchers.

As for now, at EU-LIFE institutes (see Example 3) modules of RDM are typically imbedded in other courses such as courses in Bioinformatics, ELN and Statistics. As part of the training in RDM, most institutes have or are developing an information point or helpdesk on RDM to provide specific support for individual cases. Currently, three EU-LIFE institutes have a dedicated information point/helpdesk/webpage on research data management (see also Example 4) and five are currently developing such an information point.

Personal/patient data management is another important aspect that often requires additional support and information. Although not discussed in depth here, it is fully recognized as a complex issue that requires specific expertise.

Example 3: Training in RDM at Centre for Genomic Regulation (CRG), Barcelona

At CRG modules in RDM are embedded in other courses. E.g. courses on reproducibility in bioinformatics and wet-lab experiments which contains aspects of RDM are offered. Courses in ELN for new-comers and general statistics are mandatory for first-year PhD students and open to all researchers at CRG. Specialized courses in workflow pipelines for reproducible and automated data analysis in Linux are offered twice a year. Courses are published on the CRG website https://www.crg.eu/en/events.

Example 4: Data desks at Netherlands Cancer Institute (NKI), Amsterdam

At the NKI there are 2 different data desks; one for data of patients and one for data related to business operations. These data desks have access to a data warehouse in which data from different sources are combined and can be mined.

To access patient data a researcher or clinician first has to hand in a request via an online request tool. Subsequently, the Internal Review Board reviews whether the data requested are instrumental in answering the research question and whether the research question and the use of the data are justified (e.g. did the patient give a consent). They also check whether the data can be given anonymized or pseudo-anonymized. Ideally, data can be enriched through the data desk when researcher report back about the use of the data. Besides keeping track of patient data used for research, this protocol also helps monitoring who has access to the data. Furthermore, the use of patient data according to GDPR and other legislation is facilitated and it is easy for researcher to get access to the many different data sources available in the institution (clinical trial data, radiology data, pathology data, data from electronic patient records, etc.).

The development of the data desks and the data warehouse was a combined effort from the IT department, financial department, clinical departments and researchers. First the data warehouse was build, then the data desks were introduced. The request and review process was already in place for the use of patient material but extended to also include patient data.

Recommendations

RDM is an essential part of research and of good scientific practice. Funders, publishers and governments are making it mandatory for institutes and researchers to adhere to certain guidelines regarding RDM, Open Science, FAIR data, etc. The costs for storing data are rising and becoming limitative if not managed properly. Most importantly, we are convinced that the highest quality of research can only be achieved by managing data in the best possible way.

In general, RDM is currently in swift development among EU-LIFE institutes. We believe that the alliance can play an essential role in facilitating good practice sharing among the institutes and therefore further help implementation and development of RDM policies.

Based on the survey and internal discussions hold during the EU-LIFE RDM workshops, our main recommendations are:

- An institutional RDM policy should identify the relevant stakeholders and include the purpose and relevance of the policy. It should also provide an overview of the pertinent frameworks at place at the institute and how the policy relates to them.
- Building on this general RDM policy, specific operational guidelines and protocols should be developed.
- When drafting an RDM policy and RDM guidelines all relevant stakeholders should be involved from an early stage. This will promote engagement, improve the quality of the policy and support acceptance and implementation. The continued engagement of stakeholders also allows for adjustments of the policy and/or guideline when needed.
- The RDM policy should define who owns and who manages the data at all times; and which tasks and responsibilities come with it.
- Proper data management, including processing, storage and sharing requires appropriate infrastructure and software. Implementation of ELN software is an example of such infrastructure that could be beneficial.
- There should be an open communication between the institute and external stakeholders about RDM.
- The institute leadership should prioritize RDM financially. It is crucial that the policy is complemented with sufficient budget for investments, maintenance, dedicated support (e.g. data stewards) and training.

- Institutes should implement courses and training in RDM at all stages and in particular for young researchers to raise awareness and highlight the relevance of RDM, Open Science and FAIR data. They could include courses in RDM that are specific for the relevant scientific fields. This will further facilitate the proper use of RDM as an integrated part of doing research. Training is also essential to secure broad commitment to use the RDM policies and guidelines in place.

Acknowledgements

This report was written by the EU-LIFE Research Data Management Task Force consisting of Sofie K. Christensen (EU-LIFE), Michela Bertero (CRG), Henri Van Luenen (NKI), Marta Agostinho (EU-LIFE) upon the contribution and input of the participants in the EU-LIFE Research Data Management workshops (see Annex B).

eulife

ANNEX A: Summary of the EU-LIFE Survey on Research Data Management

Objective

Good Research Data Management (RDM) is an essential part of good scientific practice including FAIR data and Open Science. Many research institutions provide tools, infrastructure and support to facilitate RDM, although there is no uniform approach.

We ran the following survey in the first quartile of 2020 to gain insight in what EU-LIFE research institutions require, implement and what they are developing in the field of RDM. The survey aimed to provide a better understanding of the status and the potential joint actions EU-LIFE could take to provide better support to researchers in RDM. The results were also the basis for the joint workshops focused on RDM that were delivered in 2020.

Summary

11 EU-LIFE institutes (out of 14 members at that time) answered the survey. The results clearly highlight that RDM is an important priority for EU-LIFE institutes, with several actions and policies in development or already in place. Main observations are the following:

- *Policies*: Most institutes are in the process of developing a policy; among those who have it, the policy is relatively "straightforward".
- *Leadership*: Most institutes have or are setting up a steering committee on data management. In some institutes, the committee has a technical focus (IT, core facilities, etc.); in others, it has a broader representation (scientific coordination, direction, scientific affairs, legal representative, etc.). Interestingly, most institutes have or are establishing dedicated roles on data management, although these roles are often quite different and heterogeneous among institutes.
- *Core Facilities*: Most institutes´ Core Facilities have or are developing policies on data storage. Very few have a system for automatically linking ELNs to the services and data / metadata generated at Core Facilities.
- *Services and Support*: Most institutes have or are developing an information point/helpdesk/webpage on research data management. Personal data management is an important aspect that often requires more support and information. The technical infrastructure is in place in most institutes, but with a varying degree of services. Most institutes offer infrastructure for creating/collecting data, processing, archiving and storing; almost half of respondents for data sharing. Destroying or re-using data is a service considered only by 2 institutes (additional 3 in development).
- *Training*: Most institutes offer training in a scattered, not systematic way.
- *Cost model*: A few institutes offer some services free of charge and some on a paid model. Overall, most institutes have not yet performedan in-depth analysis of costs for research data management.

For further details on the survey findings contact the EU-LIFE office.

ANNEX B: EU-LIFE RDM workshops 2020

Topics

Workshop 1: RDM policies; 21st of September 2020; moderated by Henri Van Luenen (NKI)
Workshop 2: Training in RDM; 1st of October 2020; moderated by Henri Van Luenen (NKI)
Workshop 3: Core facilities and RDM; 10th of November; moderated by Michela Bertero (CRG)

Participants

Agnieszka Ziemka (IIMCB), Alexander Botzki (VIB), Anne Vognsen (BRIC), Arnaud Gerard Michel Ceol (IEO), Beatriz G. Fernandez (IGC), Bernhard Korn (FMI), Cheryl Smythe (Babraham), Damjana Kastelic (CRG), Danielle Hoyle (Babraham), Dörthe Nickel (IC), Eduard Sabido (CRG), Emmanuel Dialynas (IMBB), Emmy Verschuren (FIMM), Emyr James (CRG), Jan Stransky (CEITEC), Jiri Pavlicek (CEITEC), Jirka Novacek (CEITEC), Josef Houser (CEITEC), Julia Ponomarenko (CRG), Jutta Steinkotter (MDC), Katerina Hosmova (CEITEC), Katja Kivinen (FIMM), Krzysztof Skowronek (IIMCB), Laetitia Chanas (IC), Lavanya Premvardhan (IC), Monica Morales (CRG), Myriam Alcalay (IEO), Nicolas Favre (FMI), Nuno Moreno (IGC), Ondrej Hradil (CEITEC), Özlem Özkan (MDC), Panagiotis Alexiou (CEITEC), Pavel Mikulecky (CEITEC), Pavla Foltynova (CEITEC), Paweł Kobylarz (IIMCB), Rinaldo Beck (VIB), Robert Pyke (Babraham), Robert Vacha (CEITEC), Roberta Carbone (IEO), Sara El-Gebali (MDC), Sofia Foukaraki (IMBB), Sonja Reiland (CRG), Steffi Besselink (FMI), Tania Charnavets (CEITEC), Theodoros Kosteas (IMBB), Tiago Paixao (IGC), Urszula Wyrzykowska (IIMCB), Vasiliki Teod (IMBB), Vladir Sklenar (CEITEC), Vojtech Bystry (CEITEC)

eulife

About EU-LIFE

EU-LIFE is an alliance of research centres whose mission is to support and strengthen European research excellence (www.eu-life.eu). EU-LIFE members are leading research institutes in their countries and internationally renowned for producing excellent research, widely transferring knowledge and nurturing talent.

EU-LIFE Partners

Center for Genomic Regulation (**CRG, Spain**) | Central European Institute of Technology (**CEITEC, Czech Republic**) | European Institute of Oncology (**IEO, Italy**) | Flanders Institute For Biotechnology (**VIB, Belgium**) | Friedrich Miescher Institute for Biomedical Research (**FMI, Switzerland**) | Institut Curie (**IC, France**) | Institute for Molecular Medicine Finland (**FIMM, Finland)** | Instituto Gulbenkian de Ciência (**IGC, Portugal**) | International Institute of Molecular and Cell biology in Warsaw (**IIMCB, Poland**) I Max Delbrück Center for Molecular Medicine in the Helmholtz Association (**MDC, Germany**) | Research Center for Molecular Medicine of the Austrian Academy of Sciences (**CeMM, Austria**) | The Babraham Institute (**Babraham, United Kingdom)** | The Netherlands Cancer Institute (**NKI, The Netherlands**) | The University of Copenhagen Biotech Research & Innovation Centre (**BRIC, Denmark)** I Institute of Molecular Biology & Biotechnology (**IMBB FORTH, Greece,** Associate Partner)

For more information

Marta Agostinho, PhD, EU-LIFE Coordinator

Email: marta.agostinho@eu-life.eu I Mobile: +34619570820 I http://eu-life.eu/