761030_CARBAFIN

Carbohydrate-based fine chemicals: Development of a glycosylation platform cell factory and optimization of downstream processing for the sustainable production of glycosides.

H2020-NMBP-BIO-2017

Deliverable D7.1:
First data management plan

Date of Delivery	**Contractual**	30.06.18	**Actual**	29.06.18
Status	version 1			
Nature	ORDP (open research data pilot)			
Dissemination level	PU (public)			

Authors (Beneficiary)	Christiane Luley			
Responsible Author	**Name**	Christiane Luley	**E-mail**	christiane.luley@acib.at
	Beneficiary	acib	**Phone**	+43 316 873 8404
Number of pages	7			

Data management plan (DMP)

We used the Horizon 2020 FAIR Data Management Plan (DMP) template (Version: 26 July 2016) and the DMP online tool for preparation of our first data management plan. According to the general definition of DMP by the European Commission, participating in the Open Research Data (ORD) Pilot does not necessarily mean opening up all our research data. Rather, the ORD pilot follows the principle "as open as possible, as closed as necessary" and focuses on encouraging sound data management as an essential part of research best practice.

1 Data summary

The purpose of our data collection/generation is to provide underlying data of scientific publications with the appropriate metadata (and standards) and in the adequate format to increase reproducibility. Our CARBAFIN project is an Innovation Action and the chance is high that we exploit and protect our research results by patenting. Next to it we will also disseminate and share our research results in scientific publications. We have an innovation management team to decide on the best dissemination and exploitation measures.

CARBAFIN is developing a biocatalytic glycosylation platform technology with integrated downstream processing for the industrial production of functional glycosides. The project will produce data for biocatalyst preparation and characterization, for biocatalytic reaction and process engineering and for product isolation and characterization. Moreover data on dissemination activities including surveys will be made available via our CARBAFIN homepage (www.carbafin.eu).

An overview on types and formats of data (including standards) that could be generated within the course of the project is given in Table 1. We will follow the STRENDA convention in respect to standards for reporting enzyme data and biocatalytic reaction data [1-4].

The origin of the data will mainly come from spectrophotometric measurements [Absorbance unit] and chromatographic measurements with UV, RI and pulsed-amperometric detection [mAU/min, RIU/min, nC/min] but also from DNA sequencing [nucleotide sequence] and from statistical analyses. We will document the information first in a laboratory notebook and then copy the data into an Excel spreadsheet. The Excel spreadsheet will be saved as a comma separated value file (.CSV).

The expected size of each dataset will not exceed 50 GB which fits perfect with the guidelines from Zenodo.

Our data might be useful ('data utility') for other researchers/scientists in the field of biocatalysis.

Table 1. Examples for types and formats of data, including standards.

Operation	Type of data and standards	Format
Biocatalyst/enzyme variant	Gene/protein sequence Vector map Expression strain/plasmid	FASTA SCF Collection
Bioreactor cultivation	Cell dry mass (CDM) [g/L] Total protein [g/L] Activity [U/g_{CDM}]	CSV
Protein purification	UV chromatogram [mAU/min] Specific activity [U/mg protein]	CSV
Biocatalyst/enzyme characterization	Temperature optimum [°C] pH optimum [pH] Kinetic parameters: k_{cat} [s^{-1}], K_m [mM, µM, nM]	CSV
Reaction engineering	Product yield [%] Product turnover [µmol product/g biocatalyst]	CSV
Process engineering	Substrate consumption/product formation [mM or g/L] Space time yield [g/L/h]	CSV
Product isolation	Recovery yield [%] Purity [%]	CSV
Product characterization	1H and 13C-NMR [ppm] Efficacy test according to applications in Cometics, Food and Feed	CSV
Dissemination	Survey “Gender in biotech jobs” Consumer awareness	Homepage

2 FAIR data

2.1 Making data findable, including provisions for metadata

The underlying data of scientific publications produced in this project will be stored on the research data and publication repository Zenodo. Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citeable. Zenodo further supports harvesting of all content via the OAI-PMH protocol.

Appropriate search keywords that optimize possibilities for re-use will be provided in the context of the publication. The underlying dataset can be easily linked to the publication using tools provided by Zenodo.

We will provide clear version numbers (v1.0, v1.1, v1.2 …). DOI versioning from Zenodo allows us to edit/update the record's files after they have been published, to cite a specific version of our record or to cite all versions of our record. Note, that metadata of our upload can be edited without creating a new version.

We will follow the STRENDA convention in respect to standards for reporting enzyme data and biocatalytic reaction data [1-4]. Note that STRENDA does not dictate or limit the experimental techniques used in enzymology experiments. The emphasis is on providing useful and reliable information. We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a TXT document that will be stored with the data file. The metadata will fully describe the data files and the context of the measurements (see Table 2).

Table 2. Possible types of metadata.

Category	**Information on**
Setting	Instrument, (operator), date
Methodology	Assay method, type of assay, limit of detection, wavelength, column, mobile phase, flow rate, temperature, time, …
Assay/Reaction conditions	Substrate purity, substrate concentration, biocatalyst/enzyme concentration, buffer system, pH, temperature, pressure, coupled assay components, mixing, sampling, stopping, …
Data processing	Description of software, equations

2.2 Making data openly accessible

Underlying data of scientific publications produced in the project will be made openly available as the default. However, we will still keep the possibility to partially opt-out for the individual datasets. Datasets from life cycle assessment and economic analysis cannot be shared (or need to be shared under restrictions) as they have a major impact on development of business plans for our industrial beneficiaries.

The underlying data of scientific publications and associated metadata will be made accessible by deposition in the research data and publication repository Zenodo. Zenodo is a certified repository which supports open access but also enables closed access. Access to datasets shared under restriction will be discussed in more detail during the second data management plan. Zenodo accepts data under a variety of licenses in order to be inclusive.

Software tools that can read CSV files (spreadsheet) and SCF files (DNA sequence viewer) are needed to access our data. We follow the file format guide currently supported by the Sequence Read Archives (SRA) at NCBI, EBI, and DDBJ for gene and protein sequence format. Therefore, documentation about the software is not needed to access the data included.

2.3 Making data interoperable

The underlying data of scientific publications produced in our project will be interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. The data will adhere to standards for formats, as much as possible compliant with available (open) software applications.

According to the DCC homepage (http://www.dcc.ac.uk/resources/metadata-standards) we will follow data and metadata vocabularies, standards or methodologies from Biology (in particular from Synthetic Biology, Molecular Biology, Biochemistry, Biotechnology and Bioprocess engineering) to make our data interoperable. We will use the STRENDA Guidelines, registered in FAIRsharing.org, as reference for metadata and standards within our discipline (http://www.beilstein-institut.de/en/projects/strenda/guidelines). FAIRsharing.org is a web portal that collects interrelated data standards, databases, and policies in the life, environmental and biomedical sciences.

We will be using standard vocabularies for all data types present in our datasets, to allow inter-disciplinary interoperability. In case it is unavoidable that we use uncommon or generate project specific vocabularies, we will provide mappings to more commonly used ontologies.

2.4 Increase data re-use (through clarifying licences)

To permit the widest re-use possible, we will license the data by Creative Commons Attribution CC-BY 4.0. CC-BY 4.0 permits unrestricted use, distribution and reproduction in any medium provided that the original document is properly cited. It is a machine readable license available free of charge from creativecommons.org.

The underlying data of scientific publications will be made available for re-use once the publication is accepted. Zenodo offers the function “Reserve DOI”, so we can already use the right DOI when writing the publication. A text field will display the DOI that our record will have once it is published. This will not register the DOI yet, nor will it publish our record. Next to open access publications, Zenodo offers the possibility to upload embargoed, restricted or closed access publications. When we publish with “green” open access we will do self-archiving of the publication in Zenodo. We will consider embargo periods imposed by the Journal and link it to our publications.

According to Zenodo the data remains re-usable forever. Our data is stored in CERN Data Center. Both data files and metadata are kept in multiple online and independent replicas. CERN has considerable knowledge and experience in building and operating large scale digital repositories and a commitment to maintain this data centre to collect and store 100s of PBs of LHC data as it grows over the next 20 years. In the highly unlikely event that Zenodo will have to close operations, they guarantee that they will migrate all content to other suitable repositories, and since all uploads have DOIs, all citations and links to Zenodo resources (such as our data) will not be affected.

According to the statement above the underlying data of scientific publications produced in our project will be usable by third parties also after the end of our project using CC-BY 4.0

3 Allocation of resources

According to Zenodo`s Terms-of-Use, content may be uploaded free of charge by those without ready access to an organized data centre. As we do not have an organized data centre available within our consortium we assume that the costs for making data FAIR in our project are limited to the costs for open access publishing (gold open access).

Anyway, costs for open access publishing as well as costs related to open access to research data are eligible for reimbursement during the duration of the project as part of the Horizon 2020 grant.

The project manager together with the General Assembly members will be responsible for data management in our project.

The resources for long term preservation are going to be discussed for the second data management plan. Discussion will include questions on costs and potential value, who decides and how what data will be kept and for how long.

4 Data security

We will follow provisions of help.zenodo.org/features for data security (including data recovery as well as secure storage and transfer of sensitive data). At Zenodo the research output is stored safely for the future in the same cloud infrastructure as research data from CERN's Large Hadron Collider. They are using CERN's battle-tested repository software Invenio, which is used by some of the world's largest repositories such as INSPIRE HEP and CERN Document Server.

The underlying data of scientific publications as well as the publication itself will be safely stored in the certified research data repository Zenodo for long term preservation and curation.

5 Ethical aspects

Concerning underlying data of scientific publications we do not see any ethical or legal issues that can have an impact on data sharing. For ethics reviews see Deliverables D8.1 and D8.2.

If we do questionnaires dealing with personal data we will include an informed consent for data sharing and long term preservation.

6 Other issues

We do not make use of other national/funder/sectorial/departmental procedures for data management at the moment.

Literature

[1] Gardossi, L., Poulsen, P.B., Ballesteros, A., Hult, K., Švedas, V.K., Vasić-Rački, Đ., Carrea, G., Magnusson, A., Schmid, A., Wohlgemuth, R., and Halling, P.J. (2010) Guidelines for reporting of biocatalytic reactions. *Trends in Biotechnology*, 28 (4): 171-180.

[2] Tipton, K.F., Armstrong, R.N., Bakker, B.M., Bairoch, A., Cornish-Bowden, A., Halling, P.J., Hofmeyr, J.-H., Leyh, T.S., Kettner, C., Raushel, F.M., Rohwer, J., Schomburg, D., and Steinbeck, C. (2014) Standards for Reporting Enzyme Data: The STRENDA Consortium: What it aims to do and why it should be helpful. *Perspectives in Science*, 1 (1): 131-137.

[3] *STRENDA GUIDELINES - LIST LEVEL 1A. Data required for a complete Description of an Experiment.* 2016, doi:10.3762/strenda.17

[4] *STRENDA GUIDELINES LIST LEVEL 1B. Description of Enzyme Activity Data.* 2016, doi:10.3762/strenda.27