Light assisted solar fuel production by artificial CO_2 Reduction and water Oxidation

Deliverable D7.3

Data Management Plan (DMP)

Lead Beneficiary:	ICIQ – Institute of Chemical Research of Catalonia
Delivery date:	25 February 2021
Dissemination level:	Public
Version:	1.0

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 951843

Document Information

Grant Agreement Number	951843
Acronym	LICROX
Start date of project (Duration)	01/09/2020 (36 months)
Document due date	28/02/2021
Submission date	25/02/2021
Authors	ICIQ
Deliverable number	D7.3
Deliverable name	Data Management Plan (DMP)
WP	WP7 – Project Management

Version	Date	Author	Description
v 0.1	18/01/2021	Laura López (ICIQ)	Creation first draft
v 0.2	02/02/2021	Raffaella Buonsanti (EPFL)	Addition data types and software to be used, input on data size and procedures research misconduct procedures.
v 0.3	08/02/2021	Ian Sharp (TUM)	Addition data types and software to be used, input on data size, code of conduct and storage procedures.
v 0.4	15/02/2021	Alexandru Morosanu (Hysytech)	Addition data types and software to be used.
v 0.5	22/02/2021	Jordi Martorell (ICFO)	Addition data types
v 1.0	25/02/2021	Laura López (ICIQ)	Final document following final revision and approval by the Project Management Board

EXECUTIVE SUMMARY

This document, D7.3 Data Management Plan (DMP) presents the main elements of the data management policy used by the consortium during the **LICROX** project, which is funded by the European Union's H2020 Programme under Grant Agreement No.951843. It describes the types of data, how they will be preserved and if they are shared or confidential. The DMP provides an overview of the datasets. Initially data management will build on the existing infrastructure within the partners. Data produced by the consortium will be made available through the LICROX website and on the LICROX Teams Working Group, that serves as the internal information sharing tool for the consortium, maintained by the project management board. The publications and data repository for the project outcomes will be gradually shifted to the public repository Zenodo hosted at the Conseil Européen pour la Recherche Nucléaire (CERN) by partner ICIQ. In addition, each institution will also use their regular public repositories. This as a living document, subject to formal change control and a mid-term update is already planned by month 18 (D7.4).

List of acronyms and abbreviations

CA: Consortium Agreement

CGAB: Consortium General Assembly Board

DMP: Data Management Plan

DOI: Digital Object Identifier

D&IT: Dissemination and Innovation Team

GA: Grant Agreement

GDPR: General Data Protection Regulation

IP: Intellectual Property

IUPAC: International Union of Pure and Applied Chemistry

POPD: Processing Of Personal Data

WP: Work Package

Table of Contents

1. INTRODUCTION

As of the Work Programme 2017, the Open Research Data pilot has been extended to cover all thematic areas of Horizon 2020 per default. Participation in the Pilot implies that a Data Management Plan will have to be submitted as a deliverable during the implementation of the action. The purpose of the Data Management Plan (DMP) is to provide an overview of the main elements of the data management policy that is used by the consortium concerning the project data. This document provides a general overview of the research data that will be collected and generated within the project, including experimental and characterisation data sets. This plan outlines how these data will be handled during the project and after its completion. This first version of the DMP serves as starting point and guideline for the researchers in the LICROX project.

A DMP is not a fixed document, but it evolves during the lifespan of projects. New versions are created whenever significant changes to the project occur due to inclusion of new data sets, changes in consortium policies or other external factors. The DMP covers the complete research data life cycle. It describes the types of research data that will be collected, processed, or generated during the project, how the research data will be preserved and what parts of the datasets will be shared or kept confidential.

This document includes an overview of the datasets to be produced by the project and the specific conditions that are attached to them. An updated version is already scheduled by month 18.

2. LICROX STRATEGY

Any dissemination data linked to exploitable results will not be put into the open domain if this compromises its commercialization prospects or has inadequate protection. Intellectual Property (IP) issues and results protection will be in accordance with the Innovation Plans and agreements developed through the project (WP6, Task 6.3). Therefore, the decision to be taken by the project on how to publish its documents and data sets will come after the more general decision on whether to proceed with public dissemination directly or to seek first protection by registering the developed IP.

Treatment of personal data will be done following the General Data Protection Regulation (GDPR)[1] and will not be made public without written consent.

Open Access must be granted to all scientific publications resulting from Horizon 2020 actions. This will be done following the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (15 February 2016)[2] and any other following Open Access policy within Horizon Europe, as already described in D6.1 Dissemination and Communication Plan.

The categories of outputs to which LICROX will give Open Access (free of charge) will have to be agreed upon and approved by the consortium following the prior notice procedure for dissemination actions described in the CA (Section 8). Initial identified outputs include:

- Project reports and publications (publications, conference proceedings, workshops and survey results, etc.);

- Research data (characterisation sets, computational data, etc.);

- Dissemination and Outreach material (leaflets, roll-ups, infographics, videos, etc.);

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. This text includes the corrigendum published in the OJEU of 23 May 2018. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02016R0679-20160504&from=EN

[2] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, 15 February 2016.Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hioa-pilot-guide_en.pdf

- Deliverables (public ones).

Concerning publications, the consortium will provide Open Access to the scientific publications coming out of the action. A copy of the publication will be deposited in a public repository, OpenAIRE[3], Zenodo[4] or those provided by the consortium institutions, and available for downloading from the LICROX webpage.

Concerning research data, the main obligations of participating in the Open Research Data Pilot are:

1. To make it possible for third parties to access, mine, exploit, reproduce and disseminate -free of charge for any user -the following: (i) the published data, including associated metadata, needed to validate the results presented in scientific publications, as soon as possible (ii) other data, including raw data and associated metadata, as specified and within the deadlines laid down in the data management plan; and

2. To provide information about tools and instruments at the disposal of the beneficiaries and necessary for validating the results.

Concerning dissemination and outreach material, as well as public deliverables, those will be available through the project's website and uploaded to public repositories.

Zenodo will be used as the central scientific publication and data repository for the project outcomes. Released documents and data will progressively be added to the database. This repository has been designed to help researchers based at institutions of all sizes to share results in a wide variety of formats across all fields of science. The online repository has been created through the European Commission's Open AIRE plus project and is hosted at CERN.

Zenodo enables users to:

- easily share the long tail of small data sets in a wide variety of formats, including text, spreadsheets, audio, video, and images across all fields of science

- display and curate research results, get credited by making the research results citable and integrate them into existing reporting lines to funding agencies like the European Commission

- easily access and reuse shared research results

- define the different licenses and access levels that will be provided.

Furthermore, Zenodo assigns a Digital Object Identifier (DOI) to all publicly available uploads, to make content easily and uniquely citable. A new community within Zenodo has been created for LICROX (https://zenodo.org/communities/licrox/?page=1&size=20) where the public outcomes will be uploaded and linked with the European Commission grant reference: LICROX - Light assisted solar fuel production by artificial CO2 Reduction and water Oxidation (951843).

3. DATA COLLECTION

3.1. Data types, origin and formats

Several data origins are expected for the different data sets to be used and created within the project's lifespan.

Existing data will be used for the creation of new documents such as deliverables and publications from the project, the responsible people in charge will collect the corresponding bibliography documents needed to compare and contrast relevant information. Previous publications, and reports from partners, supporters and

[3] https://www.openaire.eu/
[4] https://zenodo.org/

other actors in the field will be taken into account, accessing several databases containing the reference documents.

New data will be created as the outcome of the research activities within the project, for instance chemical reactions verified by chemical and physicochemical analysis methods, analyses of physical and chemical properties of the materials generated, etc. The research data that will be collected and generated within the project are mainly experimental data (*e.g.* experimental procedures) collected in hard-copy lab notebooks and written lab notebooks, characterisation data sets (*e.g.* outcome of chemical analyses) collected as digital documents and reports.

- Experimental procedures: reaction schemes, reagent description, reaction conditions, experimental methodology, standards and safety issues. The dominant types of data will be text-based data, tables, drawings, plots and computational simulations. Data will be collected in hard-copy lab notebooks and written lab notebooks, as well as digital documents, such as Microsoft Office and ChemOffice (e.g. using ChemDraw for structure drawing) files in native file formats.

- HTE experiments: native UPLC/MS, UPC2 and GC/MS files for data acquisition and Analytical Studio (Virscidian) for data processing.

- NMR spectra: native TOPSPIN and ICON-NMR files for data acquisition (compressed into a .zip archive). Native TOPSPIN and MestReNova files for data processing.

- GC and HPLC chromatograms: Detailed separation data using GC or HPLC will be stored as digital format using Agilent ChemStation as well as printed hardcopies.

- X-ray crystallography (XRC): SCXRD Data collection frames stored in Bruker Apex II and CrysAlisPro formats. Processed data stored in SAINT, CrysAlisPro and SHELXL formats. PXRD raw data in Bruker DIFFRAC Plus format.

- MALDI-TOF-MS: native ASCII files for data acquisition. Native Origin files for data processing.

- TEM images: raw files are mostly in .dm3 format, they are converted in tiff or jpeg after analysis. Published data will be available as a pdf.

- SEM and AFM images: exported from the instrument in tiff format and published as pdf.

- IR, UV-Vis, Raman spectra, Spectroscopic Ellipsometry, X-ray Diffraction, Photothermal Deflection, X-ray photoelectron, energy dispersive X-ray: exported from instruments and stored as text files.

- Experiments performed at the synchrotron (X-Ray Absorption) are converted always in text files and then available at the very least as pdf format.

- Electrochemical data (CV, LSV, EIS) will be always converted and saved in text files.

- Engineering design documentation like mass and energy balance, concept diagrams, P&ID, mechanical design drawings and models will be saved electronically as .dwg, .sldprt and shared with the consortium in a .pdf format.

- Computational calculations: Datasets from Transfer matrix models and Comsol multiphysics calculations.

The consortium also plans a wide interaction with different stakeholders from different sectors. Input will be collected through the organisation of workshops and surveys, providing several documents and reports.

Finally, relevant data will be generated through the dissemination, exploitation and communication activities:

- Databases containing contact information of newsletter subscribers and workshop participants will also be generated: Excel sheets and Mailchimp documents (newsletter marketing platform used).

- Dissemination and communication materials: presentations, posters, flyers, videos, newsletters, press releases, website, social media, etc. Different formats will be used, such as Microsoft Office/Mac word and power point documents, pdf documents, audio-visual documents: mp4, jpg, png, etc. and html documents.

- Innovation and exploitation reports, such as IP searches, technology reports, list of companies in the field, etc as text and pdf files.

3.2. Data size and utility

All the research data and documentation files arising from LICROX will be stored at the responsible partner's internal fileservers, in addition, storage tools at the service of the consortium like the public website (www.licrox.eu) and the password-protected Teams working group used as the project's intranet will also be used to store project related documents and serve as sharing information tools among the consortium members and the different audiences. The size of the project generated data, including all raw data (experimental procedures, acquired data from different hardware equipment) as well as processed and structured data (documents, data analysis, results) and documents from engagement activities and dissemination and communication actions is expected to reach around 500 GB.

The data collected during the project will be useful to the research and innovation community to understand, verify and reproduce the project's results as well as provide sufficient information to establish the identity of each new compound. Accordingly, for sharing and verification purposes, all documentation and data (raw and processed) that allow replication of the research will be provided in the supplementary information section of validated peer-reviewed publications.

4. FAIR DATA

4. 1. Making data findable, including provisions for metadata

This project will follow the Guidelines on Data Management in Horizon 2020.[5] As mentioned in the previous points, the data generated by LICROX will be collected in scientific publications (peer-reviewed articles, conference proceedings, workshops, patents), dissemination and outreach material (presentations and posters at conferences, flyers, videos, press releases, etc.), deliverables and reports, research and computational data sets.

Researchers will comply with the different institutional guidelines for research integrity, code of conduct and good scientific practice available for each partner. General rules include protocols for documentation of experiments in laboratory notebooks to ensure that results obtained from the primary data can be reproduced completely and easily accessed due to clear information organisation including indexing and references.

[5] European Commission, Guidelines on Data Management in Horizon 2020: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Electronic data will be stored in folder structures to group files using clear and descriptive names. Files will be uniquely named and versioned: dataset name, project name, and date.

Any confidential reports, deliverables and research data will be kept at the partner's internal fileservers, provided with appropriate backup and security options and at the Teams LICROX Working Group. These data will be correctly labelled including the project acronym and grant number:

Format title for reports/research data: Title document_LICROX_grant agreement No. 951843

Format title for deliverables: D.X.Y Title of the Deliverable_ LICROX_grant agreement No.951843

For electronic-recording of experimental procedures and characterisation data sets, metadata will be automatically assigned during file saving following the same code.

Digital Object Identifiers, DOI, will be used as persistent identifiers to unequivocally identify journal articles, reports and the rest of the project's publications. This identifier will be assigned by the publisher in the case of scientific publications in journals. In the case of other publications or documents (deliverables, dissemination and communication materials) uploaded into the ZENODO repository, an automatic DOI is created by the system.

4.2. Making data openly accessible: data access and sharing

In order to discuss the public availability of data (experimental data and characterisation datasets), as outlined above, it is convenient to distinguish two different types of data sets within project:

- Sensitive data:

The Industrial Property Units or Technology Transfer Offices for each partner will help in designing and implementing in close collaboration with the researchers involved, a protection strategy for the results of research of commercial and industrial relevance. Patent protection will be sought for those results of research that present commercial and/or industrial potential together with a plan for the development and commercialization of the technology. Consortium partners should follow the protocol established in the CA (section 8), regarding the prior notice procedure (45 calendar days notice) for transfer of ownership and for dissemination actions. Enough time should be given to the other partners to present any objections (within 30 days since prior notice).

All data related to personal information, mainly person contact names from partners and supporters and e-mail addresses, is kept under the intranet site, only reachable for invited contributors within the consortium. Additionally, contact details could be stored in personal folders from the partner institutions, as long as these data are not made public, only shared with authorised personnel from the consortium and not used for commercial purposes. The amount of personal data collected will be kept to a minimum, in most cases only includes names, employing institutions and e-mail addresses. Whenever this kind of information is requested, explicit consent will be required from users, for instance including a mandatory option to be filled in formularies and registration forms, a specific deliverable giving answers to the ethical requirements for the treatment of personal data will be generated at month 21 (see ethical aspects point below).

General Data Protection Regulation (GDPR) policies will also be followed for obtained data from the communication tools, for instance, cookies use, and data protection procedures are given in the project's website, run by ICIQ partner, under the Terms and conditions: http://licrox.eu/terms-and-conditions/.

- Non-sensitive data:

After following the prior notice procedures, research data considered to be publicly available will be included in the Supporting Information data set that will be published together with the publications of the project's results. These relevant research data will serve to supplement the main scientific conclusions, validate the results and allow other scientists in the field to understand and use them. This important, ancillary information, which is relevant to the main article but does not appear in the print (or online) version of the journal, will comprise additional tables, datasets, figures, 3D structures, and other related nonessential multimedia files. For specialist software, the software and version number used will be provided.

Data quality will be assured through repeated and comparative experimental tests and measurements, compliance to standards for data recording, the use of standard terminology and through the validation of the data collected. Other quality assurance processes will include monitoring during CGAB and other project's meetings and the peer-review of publications based on the data.

Unpublished experimental results will be archived in the partner's laboratories (e.g. notebooks, reports, printed copies of compound characterisation analyses) and fileserver (electronic files of experimental results and compound characterisation) so that they will be accessible to other researchers in the same group or institution or shared with other partners, as convenient.

As explained in previous sections, public free-of-charge access to scientific articles, public deliverables and documents generated through the project is done through:

-The LICROX website. If the document is subsequently updated, the original version will be replaced by the latest version.

-The Zenodo repository and other public repositories regularly used by the partners.

-The CORDIS website, will host all public deliverables as submitted to the European Commission.

Any dissemination of results (in any form, including electronic) must follow article 29 of the GA:

(a) display the EU emblem and

(b) include the following text:

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951843".

The data will be accessible using the following software: MSOffice, ChemOffice, TOPSPIN, MestReNova, Virscidian Analytical Studio, Mercury, Agilent ChemStation, Gatan, J.A. Woollam CompleteEASE, Rigaku SmartLab, CasaXPS, SolidWorks, AutoCAD. The majority of the software programmes are available as commercial products.

Crystal structures data will be deposited in a public crystallographic database such as the Cambridge Structural Database.[6] The free Access Structures service offers downloads in a CIF file format.

4.3. Making data interoperable

The data produced in the project will be interoperable as the datasets will adhere to standardised formats: ASCII, txt, csv, xml, tiff, cdx, opj, etc. The datasets will be interoperable as Zenodo's basic metadata requirement (i.e.

[6] https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/

1) description, 2) creator / ownership, 3) access, 4) lifecycle, 5) persistent identifiers) is compliant with the recommended standards used by OpenAIRE. Zenodo provides well described conditions for access.[7]

Standard vocabularies will be used for all data sets to ensure inter-disciplinary interoperability and re-use. The data generated will comply with subject-specific vocabularies, conventions and standards of practice for presenting data in a specific field recommended by IUPAC (International Union of Pure and Applied Chemistry).

4.4. Data re-use, preservation and archiving

The data used in scientific publications, posters and oral communications will be made available for re-use as soon as is reasonably possible. Depending on the publisher's embargo policy, Article Processing Charges (APCs) will be covered by the project funds ("Gold" open access) to facilitate immediate free-access to scientific articles upon publication, maximum delay is 6 months since publication date. The re-use of data that does not relate to peer-reviewed publications will be made available on a case-by-case basis.

Open Access, through the LICROX public website will be maintained for at least 2 years after the project ends.

All public data, documents and deliverables will be archived and preserved on Zenodo and will be retained for the lifetime of the repository, which is currently the lifetime of the host laboratory CERN (at least for the next 20 years).

Internal information kept in the intranet site LICROX Teams Working Group will be available to consortium partners for at least 12 months after the project finishes, then important information will be kept in an internal server and a hard copy by the project coordinator in ICIQ.

Other research data will be stored at partner's internal fileservers, subject to the institutional data security measures, which in most cases include frequent backups, storage of copies on local drives and physical copies (stored for 5 years minimum).

5. RESPONSIBILITY FOR THE IMPLEMENTATION OF THE DMP

Data sharing in open access mode can be restricted as a legitimate reason to protect results expected to be commercially or industrially exploited. Approaches to limit such restrictions will include agreeing on a limited embargo period or publishing selected (non-confidential) data according to the prior notice procedure described in the CA (Section 8).

Each partner of the consortium is responsible for the storage and backup of the data produced in their respective host institutions. Furthermore, each partner is responsible for uploading all the relevant data produced during the project to the LICROX Teams Working Group (restricted to the members of the consortium).

ICIQ as coordinator and leader of the Task 6.4 Dissemination and communication of the results (WP6) is in charge of ensuring that guidelines within this document are correctly pursued and of collecting all the public documents and upload them in the LICROX public website and Zenodo.

[7] http://about.zenodo.org/policies/

6. ETHICAL ASPECTS

LICROX project complies with ethical principles, including the highest standards of research integrity — as set out, for instance, in the European Code of Conduct for Research Integrity — and, in particular, avoiding fabrication, falsification, plagiarism or another research misconduct.

Regarding the protection of personal data, a deliverable (D1.1. POPD -Requirement No. 5) will be prepared at M21 (ICIQ as responsible). This deliverable has been automatically added as part of the ethics and security assessment, and it will include the following aspects:

- Justification for the processing of sensitive personal data;

- Explanation on how all the data intended to process is relevant and limited to the purposes of the research project (in accordance with the 'data minimisation 'principle);

- A description of the technical and organisational measures that will be implemented to safeguard the rights and freedoms of the data subjects/research participants;

- Description of the anonymysation/pseudonymisation techniques that will be implemented.

In addition, the deliverable will include the specific consent/assent forms prepared for the project.