Hes·so

REPORT

OPEN SCIENCE

FINAL REPORT – APPEL À PROJETS OPEN DATA HES-SO 2021

NCCR CATALYSIS

Dr. Roger Marti, Virginie Sandmeier
Institute ChemTech
HEIA-FR

Table des matières

Abbreviations 3
I/ SUMMARY OF RESEARCH 4
II/ DATA DESCRIPTION 4
III/ STATE OF THE ART 5
IV/ METHOD OF IMPLEMENTATION OF FAIR PRINCIPLES 8
V/ LEGAL AND ETHICAL ISSUES 9
Annex 11

ABBREVIATIONS

NMR	:	Nuclear Magnetic Resonance
HPLC	:	High Performance Liquid Chromatography
GC	:	Gas Chromatography
ILC	:	Ionic Liquid Chromatography
ICP	:	Inductively Coupled Plasma
AA	:	Atomic Absorption
MP-AES	:	Microwave Plasma Atomic Emission Spectroscopy
UV-VIS	:	Ultraviolet-Visible Spectroscopy
CL	:	Chemi-Luminescence
SEM-EDX	:	Scanning Electron Microscopy, Energy Dispersive X-Ray
DLS	:	Dynamic Light Scattering
NIR	:	Near-Infrared Spectroscopy
FTIR	:	Fourier Transform InfraRed
TGA	:	Thermo-Gravimetric Analysis
DSC	:	Differential Scanning Calorimetry
DMP	:	Data Management Plan

I/ SUMMARY OF RESEARCH

The National Centre of Competence in Research Catalysis (NCCR) creates the scientific and technological bases to make the chemical industry more sustainable, efficient, and CO_2-neutral. The NCCR Catalysis will reform the chemical production by discovery of new catalytic processes, make use of sustainable raw materials to build up a sustainable chemistry industry. This interdisciplinary project brings together groups from chemistry, engineering, material & computer science.

At the HEIA-FR and more particularly within the institute ChemTech, our chemistry research group is involved in catalysis research which implies synthesis, characterisation, and analysis of chemical products and materials. Our research is funded by the SNF which required to be shared in OpenAccess and OpenData.

This project is about reviewing all the possible data we produce in the laboratory, how to deal with it and enable compliance with FAIR and Open Data principles.

II/ DATA DESCRIPTION

Generated data:

Two types of data will be generated:

1. Analytical data : ca. 2 GB per project

Analytical data includes all data generated by equipment during reaction monitoring, characterization of products formed or any other laboratory manipulation (depending on each project).

The generated data will first be saved in the proprietary format of the equipment used for analysis. In a second step, the data to be published/archived/preserved will be saved in an open format (Table 1, Annex 1).

Table 1 : Summary of saving formats by equipment

Analytical Equipment	Proprietary format y/n	Open format
NMR Spectroscopy (Bruker 300)	Yes, .fid	.pdf, .txt, .csv
HPLC/GC/ILC	Yes, .cmbx	.pdf, .txt, .csv
ICP	Yes, .WVQ	.pdf, .txt, .csv
AA	No	.csv
MP-AES	Yes, .mpws	.pdf, .csv
UV-VIS	Yes, .iwbk	.xml, .csv, .txt
ACL	No	.csv, .txt
SEM-EDX	Yes, .ipj, .PAME	.csv, .txt
Microscope	No, .jpg, .avi	.tif, .avi

Rheometer	.Yes, xps, .orx	.csv, .txt
DLS	Yes, .sop, .dts, .del	.csv, .pdf
NIR	Yes, .panormama	.csv, .txt
Fluorimeter	Yes, .opj	.txt, .pdf
Raman spectroscopy	Yes, .l6s	.txt, .xml
FTIR	Yes, .mat	.xml
TGA/DSC	No	.txt, .tif, .pdf

2. Laboratory data : ca. 100 MB per project

Laboratory data includes all information on reaction or analysis conditions, observations, interpretations of results, laboratory manipulations as well as (intermediate and final) project reports (non-exhaustive list, depending on the documents generated during the project). This information is usually presented in the form of reports and/or laboratory protocols.

Laboratory data (protocol, report, etc.) are taken by hand or via an electronic lab notebook (ELN). As ELN we are using Mbook from Mestrelab Research. The data will be saved as a .pdf or .txt file.

All additional data collected will be recorded in a FAIR format.

Reused data:

The data reused are laboratory data obtained from various publications (experimental part). The references are fully cited in the project report as well as in the various publications of the project.

III/ STATE OF THE ART[1]

There is no known favorable/specialized platform for data repository for organic chemistry or catalysis research. The choice of the database was made considering the recommendations of the SNF as well as the research previously carried out for another project by O. Vorlet within the ChemTech institute.

The main criteria observed were the following:

- ***License:*** publication under CC BY license
- ***Generation of DOI identifiers:*** Several scientific publications and communication have already been made or are planned for this project. To link each communication to the right version of the dataset, it is important to be able to generate DOI identifiers.
- ***Price:*** Free service is a bonus
- ***Platform:***
  For OpenScience, a non-commercial platform is indicated.

[1] According to the final report of the OpenCE project by Olivier Vorlet

The volume of data (Set <1Go) and durability of storage (perenniality of about 10 years).

- For equivalent performance, ease of use and the quality of the online documentation are the most important criteria

The criteria of the 6 platforms were evaluated with the following ratings:

✓ fully meets expectations
≈ partially meets expectations or conditions not clear
✗ does not meet expectations

Dryad

Website	https://datadryad.org/	
Owner	Oxford University (but not exclusively?), non-commercial	≈
Price	120 $ per data set	≈
Volume	20 Go per data set	✓
License	CC0 only	✗
Services	DOI, APIs, Versioning, OrcID.	≈

Figshare

Website	https://figshare.com/	
Owner	Private company, Springer Nature	✗
Price	Free	✓
Volume	Unlimited for public data, 20 GB for private data.	✓
License	CC By 4.0, CC0 public domain, GPL, MIT, Apache 2.0.	✗
Services	DOI, API, Versioning, OrcID.	≈

Harvard Dataverse

Website	https://dataverse.harvard.edu/	
Owner	Harvard University	✓
Price	Free	✓
Volume	2.5 Go per data set	✓
License	CC0 "public domain dedication", or custom licence parameters	✓
Services	DOI, OrcID, API, Versioning	≈

Open Science Framework (OSF)

Website	https://osf.io/	
Owner	Center for Open Science, non-commercial	✓
Price	Free	✓
Volume	50 Go by public project	✓
License	CC By 4.0, CC0 public domain, GPL, MIT, Apache 2.0.	≈
Services	DOI, API, Dropbox, Github, GitLab, Google Drive, OneDrive, Versionning	✓

Yareta

Website	https://yareta.unige.ch/	
Owner	Universitée de Genève, funded by l'Etat de Genève and SwissUniversities.	✓
Price	Free up to 50 GB, limited for Geneva institutions ?	≈
Volume	Unlimited.	✓
License	Creative Commons 4.0 (CC BY, CC BY SA, CC BY ND, CC BY NC, CC BY NC ND)	✓
Services	DOI, OrcID.	≈

Zenodo

Website	https://zenodo.org/	
Owner	Hosted by CERN in Geneva, non-commercial platform	✓
Price	Free	✓
Volume	50 Go per data set. Multiple data sets allowed	✓
License	Creative Commons 4.0 (CC By, CC By-SA, CC By-ND, CC By-NC, CC By-NC-ND), ou licence personnalisée	✓
Services	DOI, OrcID, Synchronisation GitHub, Versioning	✓

Based on the list, we eliminated some:

- **Dryad** because as it costs
- **Figshare** because it is hosted by private company (Springer Nature) and we afraid of conflict of interest (Springer Nature is a big player in the publishing business)
- **Harvard Dataverse** because of the limited volume of only 2.6 Go per data set, which can be limit for our work

From the remaining 3 repository, the **Zenodo** database was chosen for the project data repository, which is also in alignment with the recommendations by the NCCR Catalysis project office. This is also due as Zenodo offers a "*community feature*" that allows to group and organize different data sets under an "umbrella". We consider to generate such a community for our research institute ChemTech, which allows better management and publicity of our work.

For a first joint publication (*Green Chem.*, **2022**, *24*, 6879; the article is open access) with the research groups of Professor Pérez-Ramírez and Professor Guillén-Gosálbez from ETH Zürich we uploaded all experimental data on Zenodo:

- https://zenodo.org/record/6135989#.Y6KnY3bMKUk

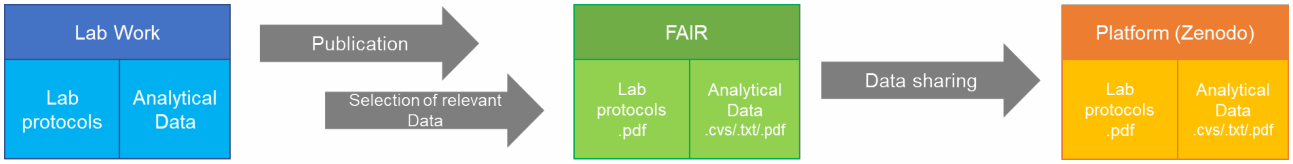
Our first experience by using Zenodo was very positive, it was easy to use and allows good organization of the data. We therefor will continue using Zenodo.

IV/ METHOD OF IMPLEMENTATION OF FAIR PRINCIPLES

At the end of the project or just after the publication, the essential and relevant data of the project will be published in open format (type: .pdf, .txt, .csv, etc.) on the Zenodo platform (see Schema 1).

For this purpose, handwritten data will be scanned and saved in .pdf format and data from MBook saved in .pdf format.

The data generated by the equipment will be saved in open format according to the tables in the appendix (Annex 1).



Schema 1 : Process flow

FAIR principles will be followed and applied as follows:

Findable
- DOI link for publications
- Data named according to a pre-established convention (see "Project Management for R&D Projects (MTI Group)" in Annex 3 and Readme.txt)
- Metadata filled in as much as possible (Readme.txt)

Accessible
- Publication on the Zenodo database
- Publication in Open Access with CC BY license

Interoperable
- Open file formats
- Use of versions to distinguish future data sets or modifications

Reusable
- Publication of data in open formats
- The dataset will be accompanied by a README file that describes the directory hierarchy. This should allow the data to be understood and add contextual value to the dataset if it is reused in the future.

The project has allowed us to deepen and better understand the interest and concept of OpenData. It allowed us to raise awareness among staff (research group, chemistry department, etc.), through training, on the interest, impact, and procedures for sharing and using FAIR data. As a bonus, a concrete model of a DMP has been set up to best assist the various project applicants within the Institute ChemTech (see Annex 3). Training is also provided to the research team and all members of the Filière Chimie to assist them in this process.

V/ LEGAL AND ETHICAL ISSUES

During the research the data is treated confidentially, following the intellectual properties agreement of the NCCR Catalysis and all data will be stored on local servers at the respective schools. Before releasing the data, we will seek a dialogue with the local TT offices (e.g. Céline Schelker for HEIA-FR) and the NCCR Catalysis TT office (Lauren Gamp, ETHZ), and once this is done we envision fully open access to all project data via Zenodo

Choice of license:

In order to keep the intellectual property while getting as close as possible to the concept of OpenScience (OpenAccess, OpenData, FAIR, etc), the data will be published under the CC BY license. This Creative Common license allows anyone to copy and adapt the data, but obliges to credit the source. It is a free copyleft license with an international scope.

BIBLIOGRAPHY

- Michael Keller, Céline Saudou and Nathalie Vogel, «Webinar Ra&D: Data Management Plan (DMP) – HES-SO 2022»
- Swiss National Science Foundation (FNS), Examples of data repositories, 2017, https://www.snf.ch/api/media/en/k64VoMUfwKoUMMY5/FAIR_data_repositories_examples.pdf
- Swissuniversities, National Open Research Data Strategy Analysis Report based on Survey and Workshop Panels, 2020, https://api.swiss-academies.ch/site/assets/files/34548/20201123_ord_grundlagenbericht_final_swu.pdf
- EPFL, Open Research Data and Open Support , Services for EPFL researchers and PhD students – Library - EPFL
- ETHZ : Research Data at ETH Zürich, Research Data at ETH Zurich – Staffnet | ETH Zurich
- HEI-VS : Gestion des données de la recherche, Gestion des données de la recherche | HES-SO Valais-Wallis (hevs.ch)
- UNIL : l'Open Science à l'UNIL, Les données de recherche - UNIL Open Science
- MIT : Data management, Find a data repository | Data management (mit.edu)

ANNEX

1. Information on equipment and file formats

Table 2: Analytical equipment

	Type	numéro de série	info	Connecté internet	Connexté réseau (au disque O)	Programme	extension de fichier de base	Enregistrable en (Open format) .CSV	.TXT	.PDF
GC-FID	Thermo TRACE 1300	714100774	GC-05	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	Thermo TRACE 1300	714100809 7 420140467	GC-04	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	Trace GC Ultra / FID / TRI PLUS AS	20058414 / 20058427	GC-06	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
GC-MS	GC Trace 1300 / AS1310 / ISQ LT	716100992 / 420160843 / ISQ1608555	GC-01	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes Attention donnée MS non prises - a exporter en plus en format .TXT	Création d'un rapport possible
	GC Trace / AS-3000 / MS-ISQ	267649 / 420102183 / 101228	GC-03	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes Attention donnée MS non prises - a exporter en plus en format .TXT	Création d'un rapport possible
GC-MS/HS	GC 1300 / MS ISQ / Triplus RSH	715000801 / 150502 / 342289	GC-07	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes Attention donnée MS non prises - a exporter en plus en format .TXT	Création d'un rapport possible
GC-HS	*En cours d'installation*		GC-02							
HPLC	Ultimate 3000 / Surveyor MSQ	50359200 / 0015 / 5722 / 5074 / 10114	HPLC-02	oui	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	Waters UPLC SQD2	E12QSM116A - E12SD110G - E12TUV274A - LCA270	HPLC-03	oui	non	MassLynx	--> prise d'image	non	non	non
	UHPLC Vanquish	Pump VF-P20-A : 8302616 Autosampler VF-A10-A : 8302628 Column Compartment VH-C10-A : 6300949 Photometer VF-D40-A : 8302599 Detector VF-D20-A : 8302489	HPLC-05	oui	non (connection via compte personnel)	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	Ultimate 300 LPG / ACC / MWD	5040.0031 / 5830.0010 / 5082.0030	HPLC-11	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	Ultimate 3000 LPG / ACC / PDA 100 - -sedex 85	8080598 / 8080878 / 03030817 / 07038037813	HPLC-01	non	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
	LPG3400SD / WPS3000TSL / TCC3000SD / DAD3000	5040.0031 / 5822.0020 / 5730.0010 / 5082.0010	HPLC-04	oui	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
(GPC)	Waters APC	L13AIS071N - L13ASM079N - J13URI114N - J13AZC052N	HPLC.06	non	non	Empower/QuickStart	??	non	sauver en .arw puis en .txt	Création d'un rapport possible
ILC	ICS-1100	074854-01	ILC-02	oui	non	Chroméléon	.cmbx --> ouvrable uniquement par chroméléon	non	info + données brutes des chromatogrammes	Création d'un rapport possible
ICP	Agilent ICP-OES 720	MY13440002	ICP-01	non	non	ICPExpert	.WVQ	oui	oui	oui
AA	Thermo ICE 3300 / 9499 400 30011	AA01132709	AA-01	oui	non	Thermo SOLAAR	.CSV	oui	oui	non
MP-AES	Agilent Technologie	AU12090106 68000A	4100 MP-AES	non	non	MP Expert	.mpws	oui	non	oui
UV-VIS	Thermo EVOLUTION 201	SA3S317009	UV-VIS-08	oui	non	Thermo INSIGHT	.XML	oui	non	non
	Thermo EVOLUTION 220	SA1P 3270002	UV-VIS-06	non connecté à un PC	non connecté à un PC	depuis l'équipement	.csv	oui	non	non
	PE Lambda 12	45240	UV-VIS-04	non connecté à un PC	non connecté à un PC	depuis l'équipement	fichier ASCII	non	non	non
	Thermo EVOLUTION 201	SA4U 342004	UV-VIS-07	oui	non	Thermo INSIGHT	.xml	oui	non	non
KF	Mettler Toledo C10S	B178052890	KF-01	non connecté à un PC	non connecté à un PC	-	relevé manuel	-	-	-
Densimètre	DMA 501			non connecté à un PC	non connecté à un PC	-	relevé manuel	-	-	-
Indice de réfraction	Anton Paar Abbemat 500	80938897	RI-01	non connecté à un PC	non connecté à un PC	-	relevé manuel	-	-	-
RMN	Bruker Ultra Shield 300	BZH 35 / 300 2708	RMN - 01	oui	non	Icon NMR / MestReNova	.mnova	oui (depuis Mnova mais pas depuis le PC de la RMN)	oui	oui

Table 3 : Chemical-physical equipment

	Type	numéro de série	info	Connecté internet	Connexté réseau (au disque O)	Programme	extension de fichier de base	Enregistrable en (Open format) .CSV	.TXT	.PDF
ACL	ACL Instruments	0109-1002-0001	CP/CL-01	non	non	ACL		oui	oui	non
SEM-EDX	Hitachi TM1000	0611-09	CP/SEM-EDX-02	oui	non	swiftED - TM	.ipj	non	Possibilité d'enregistrer en word -> pas installé sur ce PC	non
	Phenom ProX	MVE0227121208	CP/SEM-EDX-01	oui	non	Phenom Pro-suite	.PAME	oui	?	non
Microscope	Microscope Leica DML	561073200440	CP/Mikroskop-01	non	non	Motic Images Plus 2.0 ML	.jpg / .avi	non	non	non
Rhéomètre	Anton Paar MCR 302	818524420	CP/RHEO-02	oui	non	Anton Paar RheoCompass1.22	.xps	(copy past les données brute "Table" dans excel puis sauver en .csv)	(copy past les données brute "Table" dans word puis sauver en .txt)	
	Paar Physica MCR 300	426503	CP/RHEO-01	non	non	Start Rheoplus	.orx	(copy past les données brute dans excel)	(copy past les données brute dans word)	
DLS	Zetasizer Nano ZS	MAL1211099	CP/DLS-02	oui	non mais connection avec compte personnel	Zetasizer Software	.sop .dts .del	non	oui	oui
	Nanolab 3D	2014-02-002	CP/DLS-01	oui	non	LS Lab	.data .csv	oui	non	non
NIR	NBB MB 3600	1426036-001	CP/NIR-01	oui	non	Horizon mb	.panorama	oui	oui	non
UV-VIS	Thermo Scientific Evolution 220	SA2V144006	CP/UV-VIS-02	oui	non	Thermo INSIGHT	.iwbk	oui	non	oui
Fluorimètre	FluoroMax-4P	1085D-0413-FM	CP/FLUO-02	oui	non	FluorEssence	.opj	ASCII	oui	oui
RAMAN	LabRAM HR 800	L/5/849	CP/RAMAN-02	non	non	LabSpec6	.l6s	non	oui	non
FTIR	Brücker ALPHA	100304	CP/FTIR-02	non	non	OPUS	.mat		fichier	
TGA / DSC	Mettler TGA-DSC-01	B204640617	CP/TGA-DSC-01	non	non	STARe Software	??	non	oui	oui
	Mettler DSC2	B525087339	CP/DSC-03	non	non	STARe Software	??	non	oui	oui
	Mettler TGA SDTA 851^e	5116120921	CP/TGA-01	non	non	STARe Software	??	non	oui	oui