REQUIREMENTS FOR LEVERAGING OPEN DATA – CASE STUDY IN CATALYSIS

Erwin Lam
ETH Zurich, ETHZ
SwissCAT+
Switzerland
elam@ethz.ch
ORCID: 0000-0002-8641-7928

Paco Laveille
ETH Zurich, ETHZ
SwissCAT+
Switzerland
plaveille@ethz.ch
ORCID: 0000-0002-2687-0093

***Abstract* – In a highly digitalized world, openly accessible scientific data becomes of utmost importance. It ensures that scientific results can be accessed without restriction to understand historical trends, gain further insights by utilizing new computational tools, and avoid repeating costly experiments. This article discusses the relevant components that should be included when scientific data are made open-access or archived for their future reutilization. The discussion will focus on the field of experimental catalysis as a case study where the challenge consists of setting up a data management system being able to handle multiple raw data files/formats from different laboratory instruments. The discussion addresses how data are handled and what actions are taken to ensure that its processing workflow remains transparent and interoperable. It highlights the importance of making each of the data management workflow component openly available including raw data, processed data, metadata, and the processing codes.**

***Keywords* – Data Management, Open Research Data, Data Science, Digitalization, Catalysis.**

I. INTRODUCTION

Data represents a very valuable asset and is used to understand phenomena or predict future outcomes.[1], [2] Currently, those data are often not digitally available (e.g. handwritten notes) and in the context of academic research data, they are often not or only partially shared, to the community. Having data openly accessible contributes to maximizing the value that data can bring to society.[3] Applications can range from health care, finance or science and technology, where data are used to guide decision making, discovery, problem solving and optimization (Fig. 1).[1]

With the increase in device digitalization, ever more data and complex data structures are generated across different fields.[4] Digital data are often generated from various individual instruments/devices. They capture information and store them in specific formats that are machine and/or human readable. The data format is often tailored for the specific information that the device captures, for example text, pictures or videos.[4] Moreover, instruments capturing the same information can originate from different suppliers adding further complexity on data structures/formats even when providing the same information (e.g. computer operating systems, smartphone manufacturer).[5], [6]

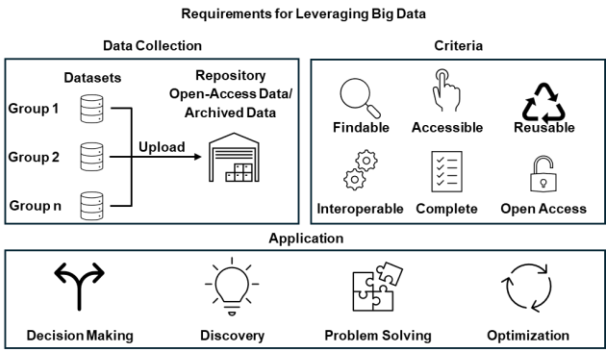

Fig.1. Scheme showing the basis of ORD and FAIR principles criteria.

To extract relevant information, datasets generated by each device usually need to be further processed to perform field-specific calculation and/or combined to relate information obtained by different equipment. Thus, complex data processing workflows are needed to merge, process and normalize datasets. Without documenting and reporting this complete processing activity, it is often not possible to understand the origin of the processed data or to replicate the data processing workflow.[7]

SRDD 2024: Swiss Research Data Day, Chur (Switzerland).

DOI: 10.nnnn/nnnnnnn.nnnnnnn

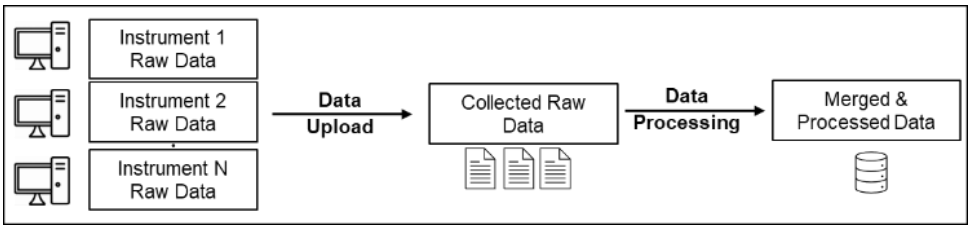

Fig. 2. Schematic data management workflow showing data generation, data collection and data processing as a requirement to obtain a complete FAIR and ORD datasets.

Data requires to be compatible with the Findable, Accessible, Interoperable and Reusable (FAIR) principles (Fig. 1)[7] to be considered open research data (ORD).[8], [9] Only if the whole data management pipeline, including raw data and processed data with their processing codes/workflows are available, data are FAIR and ORD compatible (Fig. 2). These features should also be considered when data is archived, allowing to understand historical context of the data handling, when accessed in several years or decades from now.

In scientific research, data represents an asset which often are costly to generate involving capital, operational and personnel costs. It highlights the importance of FAIR and ORD principles allowing to find and reuse data by anyone at any point of time. Adhering widely to those principles would avoid duplicating expensive experiments/resources to obtain the same information. Additionally, data from different experiments, or even different research groups, can be combined to perform statistical or machine learning based analysis to further gain knowledge and accelerate discovery.[10] This however preludes that datasets from different sources/facilities are normalized the same way to allow comparisons amongst them. This is often not the case, requiring re-processing of the source/raw data. It highlights the importance of having transparent and reproducible data management workflows.

In the next section, the challenges and requirements are discussed to ensure that overall catalysis related data are made completely accessible for the community across facilities and laboratories.

II. Data in Catalysis

Catalysis is the field in natural science studying substances and processes speeding up chemical reactions. It is estimated that >90% of industrial processes requires a catalyst.[11] This includes the production of fuels, plastics, fertilizers or pharmaceuticals, showcasing the impact catalysis has on society. Therefore, catalysis is a ubiquitous field of research where large amount of data is generated daily in academia and industry.[12]

In research and development within the field of catalysis, scientists operate several instruments to generate data for a research project/question. The data are generally occupying low storage space and are less complex but may come as multiple individual files. The main challenge arises from combining all these data (formats) originating from different instruments into a unified dataset.

In a typical workflow, a sample (catalyst) is investigated by conducting experiments and generating data about their fabrication recipe, physico-chemical properties and performances for a specific or a set of reactions. All these datasets consist predominantly of multiple files of tabular data (e.g. csv files, Excel sheets) that are structured differently and contain different information about the sample. Therefore, establishing a highly generalizable and transparent data management workflow within a facility is a major challenge.

Besides tabular data, other data types such as sequential, multi-dimensional or imaging data (e.g. spectra, chromatograms, isotherms, microscopy images) exist. This is mostly the case when data about a sample's physico-chemical properties are generated (catalyst characterization activities). Further data processing is required to extract information/descriptors from these data types. Examples of such data processing include integration of peaks, location of minima/maxima or curve fittings. Finally, metadata information describing attributes of the data files themselves, are also important. They capture for example the type of instrument and analytical method, details about the sample/project, and information about data processing performed on the specific software of the equipment.

To combine for example tabular data into a unified dataset, a straightforward way is through a relational database structure, i.e., combining individual tables through common column entries as unique identifiers. It is critical to set and maintain this relation among different data files (e.g. the sample name/barcode as identifier). One step to comply with FAIR and ORD principles is to have a transparent documentation on how those data files are connected. Especially, since each dataset has their unique data structure, highly dependent on the instrument's manufacturer.

Having collected and combined all individual raw data (tabular, sequential, metadata etc.) into one dataset, the next important step is to perform "field-specific data processing" such as data normalization. This step allows to properly compare samples (catalysts), typically catalyst characterization or performance information. Comparing catalysts is critical to drive the decision-making process such as identifying the sample best suited for further investigation or scale up for industrial application. These are generally the costliest steps emphasizing the utmost importance of having a proper data processing strategy in this field.

SRDD 2024: Swiss Research Data Day, Chur (Switzerland).

DOI: 10.nnnn/nnnnnnn.nnnnnnn

Thus, beyond the experiment and the generation of data, further complex data manipulations are performed on the data. These data manipulation steps should be reported in a transparent way to allow anyone reproducing them. If data are openly available or have been archived, it should be possible to reconstruct the whole data management pipeline at any point of time. These data manipulation/processing steps can be performed on dedicated software, on Excel sheets or through codes. Processing data through codes offers the opportunity to easily streamline and track this activity. However, codes are often written for a very specific dataset. It is therefore challenging to use the code for processing other datasets. Thus, many replicas of codes with slight modifications are created to process individual datasets. It leads to the issue where for each dataset, different codes need to be understood to process them and ultimately limits the adoption of FAIR data principles.

Ideally data processing codes should be written as general as possible allowing for their utilization on several sources of data or by different research groups, without having to modify the code directly. One way to achieve this, is to make use of input/configuration files, providing instructions on the raw data structure and how it should be processed. This instruction file also further acts as metadata for the data processing activity itself. Instruction files are often structured in YAML or JSON formats which are both human and machine-readable.

The benefits of having the combination of raw data, processing codes and processing instructions allows to transparently understand how data were processed promoting reproducibility. However, a holistic data management workflow in catalysis does not exist to this day or are only adopted on very domain specific topics within the field of catalysis only used by individual groups or a person. In the next section we describe how such workflow could be compiled and what are the different packages of data/content that should be archived.

III. CATALYSIS DATA MANAGEMENT

The example describes the data management approach performed at ETHZ SwissCAT+, a research and development facility that uses data-driven high-throughput and automated experimentation to accelerate catalyst discovery and optimization.[13] The facility operates as a service provider for academic and private users, performing the complete experimental activity including design of experiment, catalyst synthesis, characterization, performance evaluation, and data analysis and visualization. As such, ETHZ SwissCAT+ generates and aggregates large amounts of data, requiring the implementation of a suitable data management strategy.

Within the facility, multiple highly automated and digitized instruments are used to capture data related to specific experiments.[14] The established data management system is structured based on user's individual request. For each request, a "project folder", subdivided into "tasks folders" are created, where data files captured by different instruments for a given project's task are grouped together (Fig. 3).[15] The projects are individually tailored towards the needs of the project's requester. Therefore, ETHZ SwissCAT+ generates diverse data structures covering a broad range of applications in catalysis.

As such, the implementation of a holistic data management strategy is critical to:

(I) Seamlessly merge/process data from multiple instruments based on project and task.
(II) Comply to FAIR and ORD principles to emphasize the quality of the results.
(III) Apply advanced computation tools such as AI/Ml to analyse those data and accelerate optimization and discovery.
(IV) Share data to the project requester in a transparent way.

To implement such a data management system, various key features were set-up:

(I) Centralized database to collect and store data.
(II) Systematic approach to represent data.
(III) General and transparent workflow to merge and process data.

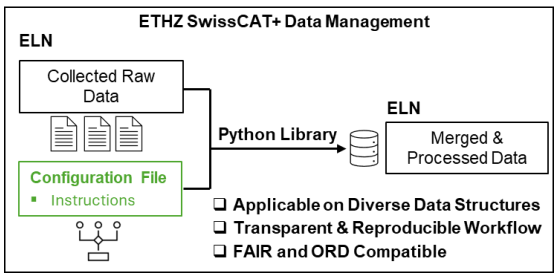

Fig. 3. Example of a general and transparent data management process at ETHZ SwissCAT+. The workflow includes raw data collection to a centralized database and using code libraries combined with an instruction file to perform data processing steps.

These features allow to collect raw data generated by individual instruments, merge raw data into a unified database (e.g. through a relational database structure), and finally perform post-processing (normalization) on the data.

Within the data management strategy at ETHZ SwissCAT+, raw data are collected in an electronic laboratory notebook (ELN) that allows to track the progress of experimental work and store data in a structured way (e.g. grouped by projects and tasks).[16], [17], [18] ELNs are widely used within catalysis research and many options are currently available. The choice of the right ELN often depends on the available features relevant to the research group and the cost of acquiring such a software.

Features such as an application programming interface (API) within the ELN allows to automate and streamline data upload/download and processing. Through the API, individual raw data files within a task are fetched, merged and processed. The processing workflow involves many steps such as cleaning the data, providing instructions on how to connect individual data together and the type of normalization and calculation to be performed. At ETHZ

SwissCAT+ a Python library has been developed to handle these workflows in an automated and traceable way.

This Python library (pycatdat),[15] allow to access the ELN, controls the user's credential, downloads the required raw data files, merge raw data through a relational database structure, and perform standard normalization and field-specific calculation (e.g. reaction rate, selectivity). A YAML configuration file contains the minimum required information to run the Python library and perform the whole data management workflow. This allows to easily record the processing parameters and to reproduce them by providing the Python library, raw data files and configuration file. This approach facilitates the application of FAIR and ORD principle in such facilities generating large amount of data daily.

IV. Challenges and Further Improvements

Data management and processing should be made as automated as possible. This avoids the introduction of human errors during data handling such as manual data naming or variation in individual's processing workflow. Automated processing allows to implement traceability and data quality control. One bottleneck is that some laboratory instruments do not provide their data in a readable format requiring a dedicated software to either pre-process or convert the data into a more common format (e.g. CSV), with sometimes a loss of information. This approach hampers the transparency and FAIR use of data. Furthermore, even if some instruments/devices are highly automated and digitized, some often lack the basic function to automatically export the data in a chosen location and format upon the completion of an experiment; It requires a person to actively open the manufacturer-specific software and manually export data, therefore introduces the possibility of human error and hampers the full automation of the data management workflow.

Processing of sequential, multi-dimensional or imaging data in a general and automated way is also very challenging. It involves sophisticated mathematical functions such as peak integration, minima/maxima detection etc. Each instrument that generates such data usually perform processing and calculation according to a specific physical theory, requiring expert domain knowledge. Often, data needs to be processed manually, which can be very tedious and untransparent. Therefore, further approaches to automate such data analysis fostering open-access data usage and adopting FAIR principles is needed. The next step toward this goal is to implement standardized methods for performing complex data processing steps that acts as guidelines for the field-specific community.[19]

To establish a community-wide accepted open-access data policy with the consideration of FAIR principles is a major challenge. It requires sensibilisation of the community on the importance and benefits of those approaches. It can only be achieved if all the involved parties see the benefit of such approaches, and that resources and datasets are openly available, understandable, and easy to use. Further investment and incentives to develop such tools and for educating the scientific generations of all ages to use them is therefore required. We are probably in a stage where the benefit of such tools, in terms of improving the quality of scientific results and accelerating the discovery processes have not been demonstrated enough.

V. Conclusion

Extracting the maximum value of data, requires making them openly available especially in the context of academic research. Openly accessible data should be structured to adhere to the FAIR data principles. Within the field of catalysis, multiple individual data files are created stemming from different instruments. These data are merged and processed to extract relevant information. The general workflow consists of taking raw data, combining them in a unified dataset and performing post-processing normalization and calculation. Open-access data should therefore consist of all raw and processed data files as well as the instructions/codes used to merge and processed them. An example is presented where raw data are collected in a centralised database in the form of an electronic laboratory notebook. Python codes are then used to extract data and process them in a transparent way by providing a configuration file containing human and machine-readable instructions on how the data was processed. Data processing in the field of natural science is highly complex and more standardized procedures needs to be established. Lastly, to fully extract the maximum value of data, the community needs to recognize the value data can have and work towards the goal of having open-access data that can be effectively reused to accelerate the scientific discovery process.

References

[1] J. Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big Data Soc*, vol. 6, no. 1, p. 2053951718820549, Jan. 2019, doi: 10.1177/2053951718820549.

[2] N. Purtova and G. van Maanen, "Data as an economic good, data as a commons, and data governance," *Law Innov Technol*, vol. 16, no. 1, pp. 1–42, Jan. 2024, doi: 10.1080/17579961.2023.2265270.

[3] "Making Open Science a Reality," Oct. 2015. doi: 10.1787/5jrs2f963zs1-en.

[4] J. H. Nord, A. Koohang, and J. Paliszkiewicz, "The Internet of Things: Review and theoretical framework," *Expert Syst Appl*, vol. 133, pp. 97–108, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.014.

[5] V. Lapatas, M. Stefanidakis, R. C. Jimenez, A. Via, and M. V. Schneider, "Data integration in biological research: an overview," *Journal of Biological Research-Thessaloniki*, vol. 22, no. 1, p. 9, 2015, doi: 10.1186/s40709-015-0032-5.

[6] J. W. Sakshaug and R. C. Steorts, "Recent Advances in Data Integration," *J Surv Stat Methodol*, vol. 11,

no. 3, pp. 513–517, Jun. 2023, doi: 10.1093/jssam/smad009.

[7] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, p. 160018, 2016, doi: 10.1038/sdata.2016.18.

[8] C. B. M. H. P. Araujo, "Recognising Open Research Data in Research Assessment: Overview of Practices and Challenges," *Zenodo*, May 2024, doi: 10.5281/zenodo.11060207.

[9] P. H. P. Jati, Y. Lin, S. Nodehi, D. B. Cahyono, and M. van Reisen, "FAIR Versus Open Data: A Comparison of Objectives and Principles," *Data Intell*, vol. 4, no. 4, pp. 867–881, Oct. 2022, doi: 10.1162/dint_a_00176.

[10] A. Trunschke, "Prospects and challenges for autonomous catalyst discovery viewed from an experimental perspective," *Catal Sci Technol*, vol. 12, no. 11, pp. 3650–3669, 2022, doi: 10.1039/D2CY00275B.

[11] R. Stevenson, Science, doi: 10.1126/article.46237.

[12] M. Suvarna and J. Pérez-Ramírez, "Embracing data science in catalysis research," *Nat Catal*, vol. 7, no. 6, pp. 624–635, 2024, doi: 10.1038/s41929-024-01150-3.

[13] P. Laveille *et al.*, "Swiss CAT+, a Data-driven Infrastructure for Accelerated Catalysts Discovery and Optimization," *Chimia (Aarau)*, vol. 77, no. 3, p. 154, Mar. 2023, doi: 10.2533/chimia.2023.154.

[14] A. Ramirez *et al.*, "Accelerated exploration of heterogeneous CO2 hydrogenation catalysts by Bayesian-optimized high-throughput and automated experimentation," *Chem Catalysis*, p. 100888, 2024, doi: https://doi.org/10.1016/j.checat.2023.100888.

[15] E. Lam *et al.*, "General Data Management Workflow to Process Tabular Data in Automated and High throughput Heterogeneous Catalysis Research," *Digital Discovery*, 2025. doi: 10.1039/D4DD00350K.

[16] C. Barillari, D. S. M. Ottoz, J. M. Fuentes-Serna, C. Ramakrishnan, B. Rinn, and F. Rudolf, "openBIS ELN-LIMS: an open-source database for academic laboratories," *Bioinformatics*, vol. 32, no. 4, pp. 638–640, Feb. 2016, doi: 10.1093/bioinformatics/btv606.

[17] A. Bauch *et al.*, "openBIS: a flexible framework for managing and analyzing complex data in biology research," *BMC Bioinformatics*, vol. 12, no. 1, p. 468, 2011, doi: 10.1186/1471-2105-12-468.

[18] S. G. Higgins, A. A. Nogiwa-Valdez, and M. M. Stevens, "Considerations for implementing electronic laboratory notebooks in an academic research environment," *Nat Protoc*, vol. 17, no. 2, pp. 179–189, 2022, doi: 10.1038/s41596-021-00645-8.

[19] A. Trunschke *et al.*, "Towards Experimental Handbooks in Catalysis," *Top Catal*, vol. 63, no. 19, pp. 1683–1699, 2020, doi: 10.1007/s11244-020-01380-2.

ACKNOWLEDGMENTS

This work was funded by the ETH Domain through the Forschungsinfrastrukturen Program. This work was supported by the Open Research Data Program of the ETH Board and the Swiss Universities Open Science Program

CONTRIBUTIONS

Erwin Lam: conceptualization, funding acquisition, project administration, visualization, writing – original draft, writing – review & editing

Paco Laveille: conceptualization, funding acquisition, investigation, project administration, writing – review & editing