Catalysis Communications 162 (2022) 106384

Contents lists available at ScienceDirect

Catalysis Communications

journal homepage: www.elsevier.com/locate/catcom

Short communication

Designing a data infrastructure for catalysis science aligned to FAIR data principles

Abraham Nieva de la Hidalga [a,b,*], Josephine Goodall [a,b], Corinne Anyika [a,b], Brian Matthews [c], C. Richard A. Catlow [a,b,d]

[a] UK Catalysis Hub, Research Complex at Harwell, Rutherford Appleton Laboratory, R92 Harwell, Oxfordshire OX11 0FA, United Kingdom
[b] School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff, CF10 3AT, United Kingdom
[c] Scientific Computing Department STFC, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX, United Kingdom
[d] Department of Chemistry, University College London, 20 Gordon Street, London WC1E 6BT, United Kingdom

ARTICLE INFO

Keywords:
Research data management
Prototyping
Catalysis research data
FAIR data principles

ABSTRACT

The UK Catalysis Hub (UKCH) is designing and implementing an infrastructure to facilitate the management of research data produced by researchers, the Catalysis Data Infrastructure (CDI). The CDI is proposed to encompass the presentation of research outputs (publications and data) in a digital repository that brings together an array of heterogeneous data types. The CDI is designed to hold references to research outputs, maintains links between them and promotes publishing and sharing of data. The proposal is to create persistent relationships between the different types of data and publications complying with FAIR data principles (findability, accessibility, interoperability, and reuse). In this paper, we will discuss how the elicited requirements for data management are being incorporated in the design of the CDI. The prototype has been used in discussion with researchers and in presentations to the UKCH community, generating increased interest and providing ideas for further development. Additionally, the CDI prototype and its code are publicly available for further analysis.

1. Introduction

Experimental and computational simulation techniques developed to understand the nature of materials and their practical applications in catalysis research rely on the use of data for building and validating complex models. The UK Catalysis Hub (UKCH) enables cutting-edge research in catalytic science, by facilitating access to state-of-the-art resources and expertise. UKCH provides access to well-equipped laboratories, sponsors access to facilities provided by the Science and Technology Facilities Council (STFC) and offers expert advice for processing and analysis of the data produced from experiments and theoretical models.

UKCH researchers use advanced processing and analysis software such as Mantid [1], DAWN [2], Larch [24], and Demeter [27] to handle the data produced by their research projects. These tools allow scientists to process and analyze data interactively. Additionally, each scientist has a choice of analysis software such as MATLAB, R, and Excel, to further analyze data and to format results for publishing. STFC facilities (Central Laser Facility [6], Diamond Light Source [13], and ISIS Neutron and Muon Source [8,16]) operate 24 h a day and have the capacity to perform thousands of readings producing large datasets. In this scenario, the amounts of data generated by each experiment is constantly growing, and so is the time employed in data management. Moreover, new experiment proposals [32] aiming to collect even larger quantities of data highlight the importance of a strategy for managing research outputs.

Having in mind the current and future requirements for producing, curating, and preserving increasing data volumes, the UK Catalysis Hub has proposed implementing a portal for facilitating research data management, the Catalysis Data Infrastructure (CDI). The CDI is designed to facilitate the cataloguing of research outputs in an easy to access digital repository. At the start of the project, publications were the only consistently identifiable outputs of UKCH research, having been collected and catalogued by the UKCH administrative team. The CDI is intended to complement the publications catalogue with links to an array of heterogeneous data sets which are also valuable research results. In this case, the CDI holds references to research outputs, maintains links between them and promotes publishing and sharing of data.

* Corresponding author at: UK Catalysis Hub, Research Complex at Harwell, Rutherford Appleton Laboratory, R92 Harwell, Oxfordshire OX11 0FA, United Kingdom.
E-mail address: nievadelahidalgaa@cardiff.ac.uk (A. Nieva de la Hidalga).

https://doi.org/10.1016/j.catcom.2021.106384
Received 31 October 2021; Received in revised form 13 December 2021; Accepted 20 December 2021
Available online 21 December 2021
1566-7367/

In recent years, complying with FAIR data principles (findability, accessibility, interoperability, and reuse [33]) has gathered interest because it eases data integration, which is needed for cross-disciplinary linking and combination of data from different domains. Alignment to FAIR Data Principles make data more valuable as it is easier to find through unique identifiers and easier to combine and integrate thanks to the formal shared knowledge representation. Consequently, we propose making the CDI FAIR compliant by design. The CDI will support findability and accessibility of data assets by cataloguing the assets and providing links and identifiers. Interoperability and reusability will be supported by the provision of context, this is by linking the datasets to publications which specify the types of software resources used to produce and exploit research data objects. Until now, researchers have managed their research data in alignment with the data policies defined by universities, research institutions, and publishers using different infrastructures. This practice has resulted in data fragmentation and informal publishing practices [31,34]. As a result, there is an imbalance in the way data is published, with some papers citing all the data required to replicate and validate their results while others only point to some of the data [22,28].

The wide availability of existing infrastructures for publishing data, and the existing expertise of researchers in their use suggests that there is no need for yet another data repository. However, a catalogue linking data and research outputs can fill the existing gaps identified in previous research [22,28,31,34]. As a result, the CDI is designed as a catalogue linking data, publications, and authors. In this form, the CDI would make research data more visible and highlight areas which need attention.

2. Related work

The use of software prototypes is an established software engineering practice [5,18,25,29,30]. A functional prototype can be used in proof-of-concept studies to support the illustration of complex design proposals to a wide range of system stakeholders. Publishing the prototype and letting users freely interact with it allows stakeholders to better understand the design and provide useful feedback. There are various cases in which prototypes have been used successfully to present implementation proposals and to refine and prioritize user requirements. Prototyping has been used for multiple purposes such as the description of architechtural decisions, discussion of interface design, and presentation of new functionalities. Davis et al. use a Web service-based e-science demonstrator to explain the architectural design for a text mining platform [7]. Klampanos et al. describe the implementation of an information registry prototype to demonstrate how it can enable collaboration and ensure consistency across the distributed infrastructure for Dispel and dispel4py [19]. Leong et al. present the implementation of three use cases to demonstrate the feasibility and benefits of applying a cloud driven approach to supercomputing ecosystems [20] for large scale experimental facilities.

In the scientific data management domain, Goble et.al used a demonstrator to present the design principles and functionality of the myGrid middleware suite, to facilitate sharing bioinformatics workflows [10]. Nieva and Hardisty describe an interface to bring together different Natural History data repositories [25]. Additionally, Hardisty et al. [12] have also proposed a dashboard interface as an integration gateway for heterogeneous collections of natural history data repositories. Following the examples above, a prototype of the CDI was designed and deployed to facilitate the demonstration and evaluation of the proposal. This strategy is used as an alternative to presenting complex design diagrams or mock-ups. The approach facilitates the participation of different kinds of stakeholders during the design and implementation phase.

3. Problem formulation

A high-level view of the processes needed to produce catalysis research data is presented in Fig. 1.[1] The image shows two parallel workflows which interact assyncrounously at various stages. The top workflow corresponds to large-scale research facilities such as the Central Laser Facility (CLF [6]) Diamond Light Source (Diamond [13]), and ISIS Muon and Neutron Source (ISIS [8,16]). These facilities have an operational framework founded on their Data Management Policies. In these large-scale facilities, the operational framework is commonly enacted by Laboratory Information Management (LIM) systems and the Data Management System (DMS). The main commonality of these facilities is that they use ICAT, an advanced catalogue system that combines LIM and DMS functionalities [9]. ICAT is developed by the Scientific Computing Department of the Science and Technology Facilities Council (SCD-STFC) and other institutions. The ICAT system contains complementary data for each experiment like proposal, PI, Experimenter, Grant(s), device(s), experiment metadata and experiment results. The lower workflow corresponds to the tasks performed by research institutions, universities, industry, and publishers. These entities will likely have their own data management practices, providing guidelines, repositories, and tools to facilitate data management. Scientists who have been awarded experimental time at the Facilities are the key stakeholders bringing together the resources at their disposal to generate the required research data. These scientists are the target users of the CDI as the amount of data produced from processing and analysing can rapidly grow and become more complex.

The data management activities are related to the entire workflow shown in Fig. 1, as metadata from the design and commissioning phases is carried through and linked to the final products. However, the tasks highlighted in red are the source of the diverse and expanding datasets which the CDI aims to catalogue. These tasks are the ones which require further support, as researchers report that processing and analysing data after the experiment requires substantial amount of time and processing resources. The research facilities provide software for collecting and formatting the data generated (for instance Mantid [1] and DAWN [2]); however, the researchers still need to handle the data and combine it with other data according to their objectives.

Researchers rely on a combination of data and software resources (own and shared) in their daily work. In this context, there are several issues that the researcher needs to handle, such as mastering the use of several types of analysis tools including lab equipment, processing software and databases; converting data so that it can be used at different stages; and ensure the reproducibility of the results by tracking equipment and software used, entry parameters, intermediate results, and versions of completed runs. The classic research workflow (Fig. 1) is well understood by all researchers. It covers activities which start with the design of an experiment and conclude with the publication of results. However, this workflow also requires the inclusion of data management, to provide a full picture. Data management is an essential activity within the research workflow. This workflow is not linear, and it does not occur all in one place. Multiple entities may participate and provide resources. From the researcher's viewpoint this is a continuous process in which s/he is the main operator. As such, the research and data management workflow may look like a linear pipeline in which researchers marshal resources devices, software and databases provided by different entities.

4. The proposed approach

Researchers are already familiar with the management of valuable research data. For this they use various repositories depending on the types of data assets they need to preserve. These include institutional

[1] Adpated from talk at the FAIR publishing of chemistry research data objects, 17 RDA Plenary, 2021/04.

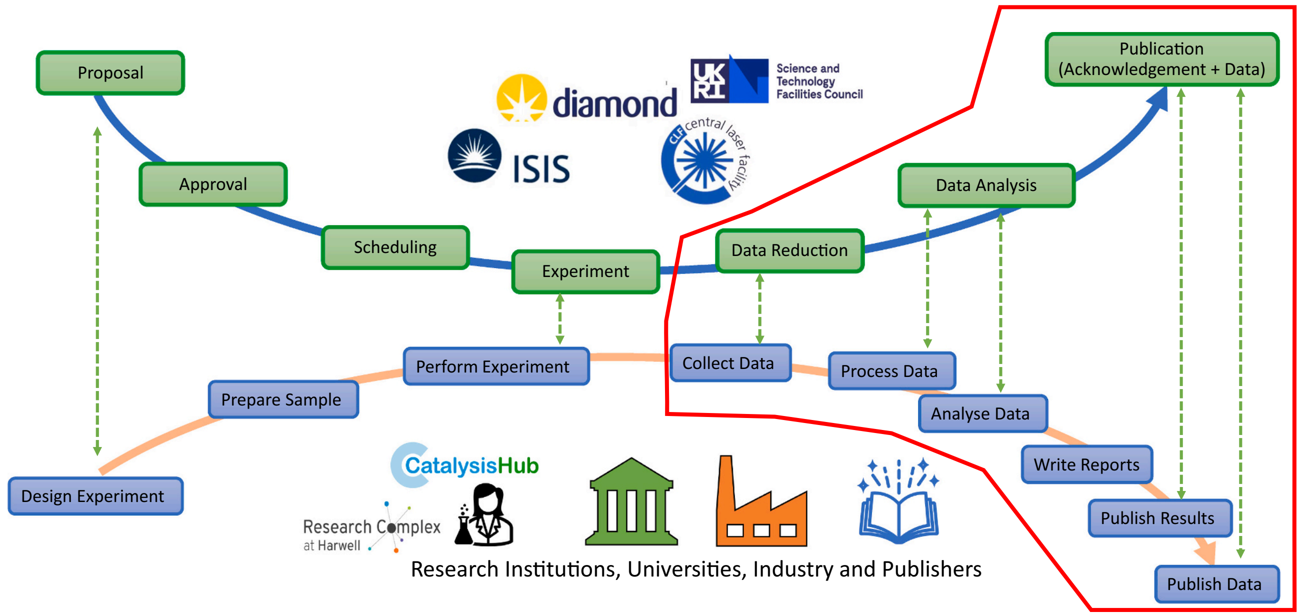

Fig. 1. The diagram shows two parallel workflows executed during research collaboration. The upper workflow is common to STFC facilities and the lower one is the workflow of institutions accessing/collaborating with STFC facilities. The CDI tracks data and publications produced by tasks highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

repositories, specialized databases, publisher repositories and research data portals. Instead of launching yet another repository, the CDI is designed as a specialized catalogue indexing and linking various research data objects, publications, researchers, and institutions. While researchers have various tools for organizing and storing references to their published works such as Mendeley [14] or ORCID [11], there is a need for a similar service for indexing their published data objects. The CDI can serve to bridge this gap.

The constant increase in size and complexity of data objects further justifies the CDI proposal. For instance, UKCH researchers have designed and commissioned a high throughput XAS reactor system for operando spectroscopy studies. On discussions about the operation of this reactor one of the recurring questions is how to manage the resulting higher data throughputs [32]. A similar reactor was developed by Kammert et al. [17], this reactor allows for the simultaneous performance of up to four independent experiments in a beamline, also increasing the types of data and the sizes produced by an experiment.

To illustrate the value of publishing data objects, in addition to publications (articles, theses, books), the interface for the catalogue was designed following the dashboard pattern [23,26]. The dashboard interface presents research production indicators in the form of tables and graphs. Fig. 2 shows the landing page of the CDI with the main research production indicators at the time of the final revision of this article (December 2021). In addition to indexing the publications and data objects, the inclusion of author, institution and themes are intended to support the search for data objects and publications in alternative presentation and categorization formats.

4.1. Linking data and publications

Unlike publications, research data objects are not commonly included as indicators of research production, while publication counts, and citations are commonly used to measure a researchers' performance. Despite repeated efforts to encourage publishing research data, it continues to be an activity which is only performed to fulfill publishers and funders requirements at the end of the research process. Moreover, there are limited numbers of examples which point to the reuse of data or that cite data effectively. In principle, the CDI aims to address this imbalance by bringing research data objects to the forefront. Attribution, research specialisation and collaboration are also seen as important indicators of successful research. For this reason, authors, institutions, and research themes were also considered as first-class entities. This results in five main entities being tracked in the CDI: publications, authors, institutions, themes, and datasets. Each of these entities in turn have their own landing page which allows searching through the entities which are al interlinked with each other. This enables the user of the CDI to search for publications and data objects by theme, author, and institution. The result is an interface which allows drilling down on the details of each entity to discover more detailed facts, providing a richer context for searching for data and publications.

4.2. Gathering and categorising data

The UK Catalysis Hub tracks researchers' publications as they are the most visible products of research. The initial list of publications produced in line with the UKCH research themes represented the base for the building of the CDI database. Publications are also a source of further information, for instance researchers coauthoring papers, the researchers' institutions, and the data objects produced by the research. For this reason, the primary goal for collecting data focused on identifying the publications acknowledging the support of the UKCH and then extracting data from those publications about related entities (authors, institutions, research themes, and data objects).

Authors, institutions, and research themes can be commoly found in the main metadata of publications, whereas data objects are not normally consistently linked. Data is either referenced in the article's main text, as part of the data statement required by some publishers or as part of the supplementary materials. The repositories used by researchers are provided by research institutions, universities, or publishers. Research data repositories can be categorized as general-purpose, institutional, publishers, and specialized databases. The array of repositories goes from general purpose repositories (e. g. Zenodo) which are interinstitutional and accept data assets from any field to specialized databases which catalogue only specific types of data (e. g. CCDC). Table 1 shows some examples of the repositories used to store different types of research data objects.

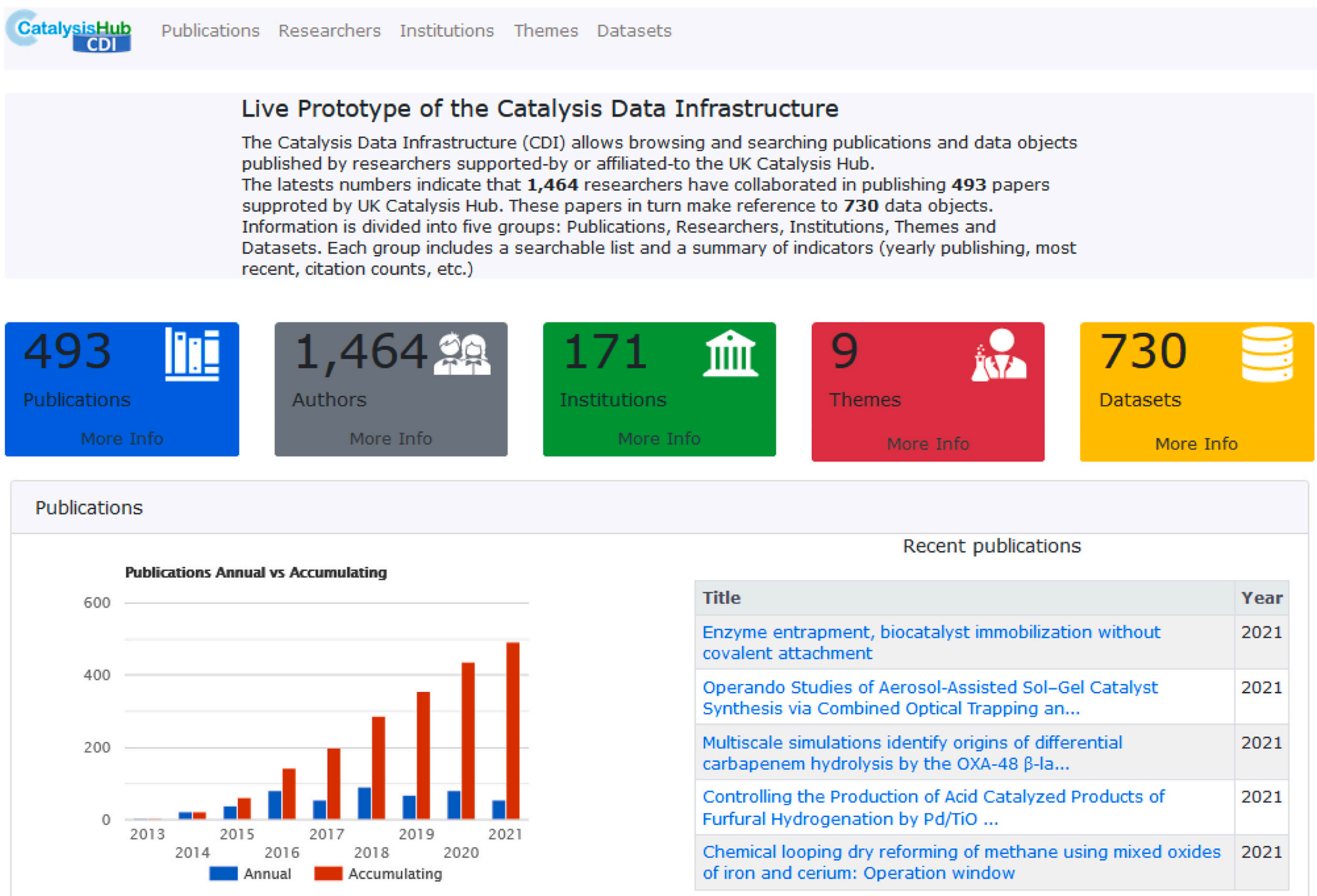

Fig. 2. A view of the Prototype Catalysis Data Infrastructure Home Page.

Table 1
Example of research data repositories.

Type	Name		Provider
General-purpose	Figshare Research Repository https://figshare.com		Digital Science
	Zenodo Research Data Repository https://zenodo.org		CERN
	Dryad Digital Repository https://datadryad.org		Dryad
Institutional	Research Facilities	eData: the STFC Research Data Repository https://edata.stfc.ac.uk/	Science and Technology Facilities Council
		ISIS Data Catalogue https://data.isis.stfc.ac.uk/	ISIS muon and neutron source
	Universities	Cardiff University Research Portal https://research.cardiff.ac.uk/	Cardiff University
		Research Data Oxford https://researchdata.ox.ac.uk/	University of Oxford
Publishers	Supplementary information RSC http://www.rsc.org/suppdata		Royal Society of Chemistry
	Supplementary information ACS https://pubs.acs.org/doi/suppl		American Chemical Society
Specialized Databases	The Cambridge Crystallographic Data Centre (CCDC) https://www.ccdc.cam.ac.uk/		University of Cambridge
	SUNCAT Catalysis Hub https://www.catalysis-hub.org		Stanford University

4.3. *Implementing the prototype*

After gathering the relevant data, a working prototype was seen as the most effective way to present the design proposal to all stakeholders, to encourage actual discussion and to facilitate the illustration of design alternatives. The prototype has been demonstrated in different internal and external forums, allowing a wide variety of stakeholders to participate in the discussions about the design of the prototype. The prototype evolved from a publications database, which was used to extract research output indicators, presented in an online application which can be accessed over the internet. The current version of the prototype was developed using Ruby on Rails and a SQL database in the backend, the latest version of this prototype is published online and can be accessed from Catalysis Hub github page for the project https://github.com/UK-Catalysis-Hub/ukcathubapp.

5. Demonstrations and analysis

This section discusses the gathering and classification of data presented in the prototype as part of the experimental setup. These data are valuable and will be preserved and migrated to the actual CDI implementation. Additionally, we discuss the evolution of the prototype in line with the different demonstrations and conclude by discussing the alignment with fair data principles.

5.1. *Datasets and experimental setup*

As described previously, the basis for the database was the initial list of UKCH publications. This list contained 270 publications mostly from the first phase of the UKCH (2013–2018). Because of the lack of a system and methodology for gathering publication data, the publications list was gathered from web searches[2] and filtering for acknowledgement of the UKCH support in publications. These searches returned a large quantity of results which then had to be manually curated before adding them to the publications' list. At the start of this project, this search was switched to using the CrossRef API. Before the implementation of the prototype the searches were performed using python [3]. After the creation of the prototype, ruby has been used to retrieve and parse the data for publications [4]. The search can be made for any of the grant numbers associated with UKCH funding, and by the affiliations of authors. The results of these searches are manually validated and catalogued before being added to the CDI database, to ensure that they are actual products of UKCH research.

The search for datasets is more complicated as publications do not normally include related data objects in their metadata, and the policies for citing and referencing data produced varies between journals. Some journals will require the inclusion of a data statement, indicating how to get access to the publications' data source, others require a complementary data section/annex. Consequently, it is necessary to analyze the full article text to find references to data objects. During the search and gathering process, we discovered that some articles include the references in the pdf version, while others indicate that references to associated data are provided in the online version. For this reason, locating data objects required retrieving and parsing both the pdf and html versions of the articles. The analysis was performed by applying data mining tools (pdfminer[3] and ChemDataExtractor[4]) to the articles full text.

The initial list of publications has been continuously updated during the analysis and design phase. The numbers at the time of writhing indicate that 493 papers have been published and that these papers in turn refer to 730 data objects. Information is divided into five groups: publications, researchers, institutions, themes, and datasets. Each group includes a searchable list and a summary of indicators (yearly publishing, most recent, citation counts, etc.)

5.2. Demonstrations

The aim of the prototyping activities is to rapidly present and gather feedback from UKCH stakeholders and the wider research community. For this, three versions of the CDI design proposal and its prototype have been presented. These include periodic project reports to the UKCH steering group (bimonthly and yearly reports), presentations at UKCH workshops and symposiums (2019–2021), UK catalysis conference (2020 and 2021), UKCH conference (2020), and the 17th RDA plenary (2021). Before the publishing of the prototype interface in April 2021, the demonstrations concentrated on presenting the advances in the collection and classification of data, as well as walktroughs of mockups and screenshots of the application in development. In contrast, after publishing the prototype we have been able to reach a wider audience by exposing the proposal to more user groups. The size of the groups has varied from about a dozen members of the steering group to larger groups at conferences and symposiums.

The showcasing and discussion of the prototype with key stakeholders have provided valuable feedback, suggestions for improvement and future developments. The prototype has been well accepted, and researchers have found and suggested interesting uses for it. For instance, users have been actively reviewing the publications list and pointing out to overlooked publications. Users also suggested using, in addition to themes, keywords to group research and highlight the areas of expertise and the types of compounds covered. Users have also pointed out to highlighting datasets which use open standards and the types of applications which can be used to exploit them. Colleagues from other research groups have also made recommendations about the structuring of the data using ontologies and recommending alternative presentation strategies.

5.3. Data objects and alignment to the FAIR data principles

The CDI is designed to align to FAIR data principles, however, because of the nature of the data and repositories used, the alignment is not complete in all cases. The following paragraphs explain how well the CDI aligns with FAIR data principles.

5.3.1. Findability

The findability principle is supported by making the locations and types of data assest readily available.

5.3.2. Availability

The availability principle can be tested by trying to recover the indexed data objects. If the objects can be recovered, even when the recovery implies contacting the repositories or filling a form, we can affirm that the Availability principle has been preserved.

5.3.3. Interoperability

The interoperability principle of can be assessed in relation to the structuring of the data objects. Structure data objects are highly organized and easily decipherable by machine algorithms which facilitate input, search, and manipulation of those data [15]. Unstructured data, typically categorized as qualitative data, cannot be processed, and analyzed via conventional data tools and methods, it does not have a predefined data model [15]. With the latest numbers from the CDI, 44% of the data objects are structured data objects and 56% are unstructured (documents, images, videos, or archive files). While structured data can still be processed, for instance by data mining algorithms, they are less interoperable than structured data objects. The CDI supports interoperability by specifying the data object types upfront, so even when interoperability is reduced because of unstructured nature of some data objects it alerts potential users of the factors to consider for handling those objects.

5.3.4. Reusability

The reusability principle is linked to the accessibility and interoperability principles as a data object which can be retrieved and interpreted has some degree of usability. In this way, by supporting the accessibility and interoperability principles, the CDI supports reusability. However, in addition to these attributes, reusability also refers to the fact that access to data is allowed, in this authorization for use needs also to be acknowledged. Currently, the CDI does not track authorization or licensing, so this principle is not completely supported.

Most data objects published by UKCH researchers are unstructured data objects. Fig. 3 may explain this. The chart shows a correlation in the type of structuring of the data and the role of the data objects. This trend in which most of published data accompanying chemistry articles are supplementary and unstructured has been observed by other researchers [28,31,34]. This may be interpreted as supplementary data is less FAIR because they are harder intreroperate and reuse with none or minimal human intervention. To improve this, researchers should strive to include not only unestructured supplementary data but also structured supporting data.

6. Conclusion and future work

The exercise of collecting data references indicates that authors

[2] Using Publish or Perish, available from: https://harzing.com/resources/publish-or-perish

[3] Applied pdfminer python parser, available from https://github.com/euske/pdfminer

[4] Applied ChemDataExtractor, available from: http://www.chemdataextractor2.org/

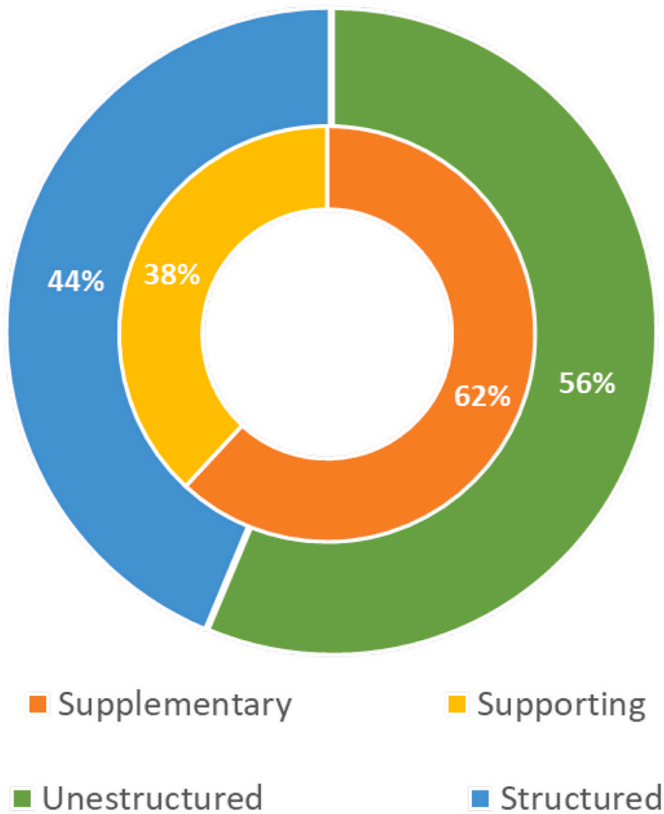

Fig. 3. Distribution of data objects in UKCH publications indexed by the CDI. The inner ring of the chart classifies the datasets as supplementary or supporting. The outher ring of the chart classifies those same objects as structured and unestructured.[a]
[a]According to data from the CDI at the time of the final revision of this article (December 2021).

regularly publish some form of FAIR compliant data. However, the main types of published data also indicates that there are large quantities and types of data which remain unpublished, or if published they are not consistently linked to publications.

Asking researchers to increase the amount and types of data they publish needs to consider that the size of the supporting datasets may be considerable. For instance, the paper by Messinis et al. [21] referrers to 39 different data objects: one a supplementary data document, 37 cristalography information files, and one large dataset (1.5GB of data including spectra, images, and intermediate results). This can indicate that further enhancement of repositories may be required to accommodate publishing larger datasets, effective linking to publications, and a review of the practices for managing datasets (licensing, indexing, citation and others). The development of the CDI provides a unique opportunity for promoting this approach while verifying how well the UKCH community is doing in improving data publishing practices.

At the moment, the design of the CDI is being revised to beter structure the data stored by alingnitn to ontologies, such as the provenance ontology (PROV-O), the data cataloguing vocabulary (DCAT) for the description of data objects, the semantic publishing and referencing ontology (SPAR), and the scholarly link exchange framework (SCHOLIX) for linking data objects and publications. In addition to facilitating the maintenance of the repository, incorporating these ontologies in the design will pave the way for introduction of more sophisticated retrieval and exploitation tools such as workflows, and artificial intelligence applications.

In terms of monitoring alignment to FAIR principles, future versions will assign a FAIRness score to each data object.

The research presented in this paper is relevant to the research catalysis community, but its findings can also help other chemistry research areas as an example for designing similar indexes for their data assets.

Although, the UK Catalysis Hub is a British institution, it fosters and facilitates collaborations with institutions in Europe and the rest of the world. The latests numbers indicate that 1464 researchers from 171 institutions in 34 countries have collaborated in research publications supproted by UK Catalysis Hub. These papers in turn refer to 730 data objects.

Declaration of Competing Interest

None.

Acknowledgements

UK Catalysis Hub is kindly thanked for resources and support provided via our membership of the UK Catalysis Hub Consortium and funded by EPSRC grant: EP/R026939/1, EP/R026815/1, EP/R026645/1, EP/R027129/1 or EP/M013219/1(biocatalysis)). We also thank the the reviewers for their thoughtful comments and efforts towards improving this paper.

References

[1] O. Arnold, J.C. Bilheux, J.M. Borreguero, A. Buts, S.I. Campbell, L. Chapon, V. E. Lynch, Mantid—Data analysis and visualization package for neutron scattering and μ SR experiments, Nuclear Inst. Methods Phys. Res. Sect. A. 764 (2014) 156–166.
[2] M. Basham, J. Filik, M.T. Wharmby, P.C.Y. Chang, B. El Kassaby, M. Gerring, J. Aishima, K. Levik, B.C.A. Pulford, I. Sikharulidze, et al., Data analysis WorkbeNch (DAWN), J. Synchrotron Radiat. 22 (2015), https://doi.org/10.1107/S1600577515002283.
[3] S. Chamberlain, Habanero Python Client Api for Crossref, available from, https://github.com/sckott/habanero, 2021.
[4] S. Chamberlain, Serrano Ruby Client Api for Crossref, available from, https://github.com/sckott/serrano, 2021.
[5] P.C. Clements, Active Reviews for Intermediate Designs (CMU/SEI-2000-TN-009). Software Engineering Institute, Carnegie Mellon University, 2000, pp. 1–25. http://www.sei.cmu.edu/library/abstracts/reports/00tn009.cfm.
[6] CLF, Central Laser Facility, retrieved on 2020-08-15, from: https://www.clf.stfc.ac.uk/Pages/home.aspx, 2020.
[7] N. Davis, G. Demetriou, R. Gaizauskas, Y. Guo, I. Roberts, Web service architectures for text mining: an exploration of the issues via an e-science demonstrator, Int. J. Web Serv. Res. 3 (4) (2006) 95–112.
[8] D.J.S. Findlay, ISIS-pulsed neutron and muon source, in: 2007 IEEE Particle Accelerator Conference (PAC), IEEE, 2007, June, pp. 695–699.
[9] D. Flannery, B. Matthews, T. Griffin, J. Bicarregui, M. Gleaves, L. Lerusse, R. Downing, A. Ashton, S. Sufi, G. Drinkwater, K. Kleese, ICAT: integrating data infrastructure for facilities-based science, in: 2009 Fifth IEEE International Conference on e-Science, IEEE, 2009, December, pp. 201–207.
[10] C. Goble, C. Wroe, R. Stevens, myGrid Consortium, The myGrid project: services, architecture and demonstrator, in: Proceedings of the UK e-Science Programme All Hands Conference, Engineering and Physical Sciences Research Council, 2003, pp. 595–603.
[11] L.L. Haak, M. Fenner, L. Paglione, E. Pentz, H. Ratner, ORCID: a system to uniquely identify researchers, Learn. Pub. 25 (4) (2012) 259–264.
[12] A. Hardisty, H. Saarenmaa, A. Casino, M. Dillen, K. Gödderz, Q. Groom, H. Hardy, D. Koureas, A. Nieva de la Hidalga, D.L. Paul, V. Runnel, X. Vermeersch, M. van Walsum, L. Willemse, Conceptual design blueprint for the DiSSCo digitization infrastructure – DELIVERABLE D8.1, Res. Ideas Outcomes 6 (2020), e54280, https://doi.org/10.3897/rio.6.e54280.
[13] C. Helmers, H.G. Overman, My precious! The location and diffusion of scientific research: evidence from the synchrotron diamond light source, Econ. J. 127 (604) (2017) 2006–2040.
[14] Holt Zaugg, Richard E. West, Isaku Tateishi, Daniel L. Randall, Mendeley: Creating Communities of Scholarly Inquiry through Research Collaboration, 2011.
[15] IBM Cloud Education, Structured vs. Unstructured Data: What's the Difference?, Accessed: 2021-12-08 Available online from: https://www.ibm.com/cloud/blog/structured-vs-unstructured-data, 2016.
[16] ISIS, ISIS Neutron and Muon Source, retrieved on 2020-08-15, from: https://www.isis.stfc.ac.uk/Pages/home.aspx, 2020.
[17] J.D. Kammert, G. Brezicki, R. Acevedo-Esteves, E. Stavitski, R.J. Davis, High-throughput operando-ready X-ray absorption spectroscopy flow reactor cell for powder samples, Rev. Sci. Instrum. 91 (1) (2020), 013107.
[18] M. Käpyaho, M. Kauppinen, Agile requirements engineering with prototyping: a case study, in: 2015 IEEE 23rd International requirements engineering conference (RE), IEEE, 2015, August, pp. 334–343.

[19] I.A. Klampanos, P. Martin, M. Atkinson, Consistency and Collaboration for Fine-Grained Scientific Workflow Development: The dispel4py Information Registry, Tech. Rep, 2019.
[20] S.H. Leong, H.C. Stadler, M.C. Chang, J.P. Dorsch, T. Aliaga, A.W. Ashton, SELVEDAS: A data and compute as a service workflow demonstrator targeting supercomputing ecosystems, in: 2020 IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud), IEEE, 2020, November, pp. 7–13.
[21] A.M. Messinis, S.L. Luckham, P.P. Wells, D. Gianolio, E.K. Gibson, H.M. O'Brien, R. B. Bedford, The highly surprising behaviour of diphosphine ligands in iron-catalysed Negishi cross-coupling, Nat. Catal. 2 (2) (2019) 123–133.
[22] Nature Research, The State of Open Data 2018: Global Attitudes towards Open Data, Springer Nature, 2018. https://figshare.com/articles/report/The_State_of_Open_Data_Report_2018/7195058.
[23] Theresa Neil, Mobile Design Pattern Gallery: UI Patterns for Smartphone Apps, O'Reilly Media, Inc, 2014, 2014.
[24] M. Newville, Larch: an analysis package for XAFS and related spectroscopies, Int. J. Phys. Conf. Ser. 430 (2013, April) 012007.
[25] A. Nieva de la Hidalga, A. Hardisty, Making heterogeneous specimen data 'FAIR': implementing a digital specimen repository, Biodiv. Inform. Sci. Stand. 3 (2019), e37163, https://doi.org/10.3897/biss.3.37163.
[26] L. Pappas, L. Whitman, Riding the technology wave: effective dashboard data visualization, in: M.J. Smith, G. Salvendy (Eds.), Human Interface and the Management of Information. Interacting with Information. Human Interface 2011. Lecture Notes in Computer Science vol 6771, Springer, Berlin, Heidelberg, 2011, https://doi.org/10.1007/978-3-642-21793-7_29.
[27] B. Ravel, M. Newville, (2005) ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT, J. Synchrotron Radiat. 12 (2005) 537–541, https://doi.org/10.1107/S0909049505012719.
[28] V. Scalfani, (2017) RDA CRDIG Open Meeting, ACS Spring Meeting, San Francisco, March 2017.
[29] A.G. Sutcliffe, User-Centered Requirements Engineering, Springer, London, 2002.
[30] L. Teixeira, V. Saavedra, C. Ferreira, J. Simões, B.S. Santos, Requirements engineering using mockups and prototyping tools: developing a healthcare web-application, in: In International Conference on Human Interface and the Management of Information, Springer, Cham, 2014, June, pp. 652–663.
[31] J. Thielen, Y. Li, Profiling common types of research data and methods published by organic synthesis chemists at the University of Michigan. Paper presented at the SLA 2015 Annual Conference & Info Expo, Boston, MA. http://hdl.handle.net/2027.42/111832, 2015.
[32] Shaojun Xu, Giannantonio Cibin, Diego Gianolio, Veronica Celorrio, Stephen Parry, Emma K. Gibson, C. Richard, A. Catlow, High throughput XAS reactor system for operando spectroscopy study, in: Oral Presentation, UK Catalysis Conference 2021, 2021-01-08, 2021.
[33] M.D. Wilkinson, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016). URL: https://www.nature.com/articles/sdata201618.
[34] R.P. Womack, Research data in core journals in biology, chemistry, mathematics, and physics, PLoS One 10 (12) (2015), e0143460.