A portal for indexing distributed FAIR digital objects for catalysis research

Abraham Nieva de la Hidalga[‡,§], Josephine Goodall[‡,§], Richard A. Catlow[|,‡,§], Corinne Anyka[¶,§,#], Brian Matthews[¤]

‡ School of Chemistry Cardiff University, Cardiff, United Kingdom
§ UK Catalysis Hub, Harwell, United Kingdom
| University College London, London, United Kingdom
¶ School of Chemistry Cardiff University, Caridff, United Kingdom
# Finden Ltd., Harwell, United Kingdom
¤ Scientific Computing Department, Science and Technology Facilities Council, Harwell, United Kingdom

Corresponding author: Abraham Nieva de la Hidalga (nievadelahidalgaa@cardiff.ac.uk)

Abstract

A research object (RO) is defined as a semantically rich aggregation of (potentially distributed) resources that provide a layer of structure on top of information delivered as linked data (Bechhofer et al. 2013, Soiland-Reyes et al. 2022). A RO provides a container for the aggregation of resources, produced, and consumed by common services and shareable within and across organisational boundaries. This work sees research digital objects as composites which may consist of objects hosted in different repositories.

In catalysis research, the characterisation of a sample may require analysing experimental data obtained from an instrument, data from a computer model, and/or comparing to data from a specialized database. Additionally, data may need to be reduced and cleaned before analysis, resulting in intermediate data. In this scenario the composite research object is integrated by all these data objects and their corresponding metadata. UK Catalysis Hub (UKCH) researchers perform these tasks as part of their day-to-day work. However, most of the time they need to manually collect, catalogue, and preserve all these data assets.

The UKCH aims to support researchers with tools and services for the management and processing of data, through the development of the Catalysis Data Infrastructure (CDI Nieva de la Hidalga et al. 2022a) and the Catalysis Research Workbench (CRW Nieva de la Hidalga et al. 2022b). This work is integrated in the context of the Physical Sciences Data Infrastructure (PSDI Coles and Knight 2022). The PSDI aims to provide a layer that enables transparent access to existing resources whilst ensuring that they remain dedicated to its specific application. The intention is to explore the concept of the composite research digital object and the services required to facilitate both human and programmatic interactions with those objects to browse, review, retrieve, and use digital objects in the context of the research produced by UKCH scientists. The CDI will act as a

thematic portal presenting data managed through the PSDI and serve as an example for the development of similar portals targeting specific research domains.

In this case, the CDI is in the process of being redesigned with a sematic metadata model. The basic ontologies being considered for this model are: DCAT (Albertoni et al. 2022) will encode the metadata of digital objects; PROV-O (Belhajjame et al. 2013) will track the generation of digital objects. SPAR (Peroni and Shotton 2018) to encode publications data; SCHOLIX to encode the links between publications and data objects (Burton et al. 2017); FOAF (Brickley and Miller 2014) to encode researcher information; the Organization Ontology (ORG Reynolds 2014) to encode institution information; EXPO (Soldatova et al. 2006) to encode experiment information; and various domain specific ontologies for adding metadata about experiments, for instance CHEBI (Hastings et al. 2011), CHEMINF (Hastings et al. 2015), and FIX (Chebi-Administrators 2005).

The implementation of the CDI using these ontologies will provide a roadmap for the integration of FAIR data object repositories with a service infrastructure which supports reproducibility, reuse of data, reuse of processing tools and implementation of advanced processing tools.

The integration of the CDI and CRW with existing and new infrastructures will further support the work of catalysis scientists. In this context, a researcher can access the CDI to look for publications, see if there are data objects linked to them, and then look for processing tools which can be used to reproduce the results. An experiment for an early use case demonstrated the feasibility of reproducing published results using data and metadata linked to existing publications (Nieva de la Hidalga et al. 2022b). In the experiment, papers citing processable data were used to retrieve, process, and reproduce published results with no need for contacting the authors. Fig. 1 presents a view of the experiment performed.

The current practices of publishing catalysis research data can be seen as aligned to the FAIR data principles, for instance Fig. 1 above can be also seen as Fig. 2

Reproducing results required several human-centered activities, partly due to the encoding of the metadata as text documents. The challenge is to accelerate and automate these processes. It is important to highlight the role of cataloguing interfaces, such as the CDI, containing DO crates with only metadata and links to the different data assests that constitute the composite digital objects. The users of these interfaces will in turn rely on transparent services which do not require them to manually track the location and formats of the data assets they want to retrieve and use.

Keywords

Research data management, Catalysis research data, FAIR data principles, Composite research object

Presenting author

Abraham Nieva de la Hidalga

Presented at

First International Conference on FAIR Digital Objects, presentation

Funding program

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))

Hosting institution

UK Catalysis Hub Research Complex at Harwell Rutherford Appleton Laboratory, R92 Harwell Oxford Oxfordshire OX11 0FA

Author contributions

ANH edited the abstract, JG and CA were responsible of curating the publications metadata, RC and BM reviewed the content and motivated the research project. All authors reviewed the contents of the abstract and agreed to its submission.

Conflicts of interest

References

- Albertoni R, Browning D, Cox S, Gonzalez Beltran A, Perego A, Winstanley P (2022) Data Catalog Vocabulary (DCAT) - Version 3. (W3C Recommendation). World Wide Web Consortium Version 3[In English]. URL: https://www.w3.org/TR/vocab-dcat-3/
- Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I, Gamble M, Michaelides D, Owen S, Newman D, Sufi S, Goble C (2013) Why linked data is not enough for scientists. Future Generation Computer Systems 29 (2): 599-611. https://doi.org/10.1016/j.future.2011.08.004

- Belhajjame K, Cheney J, Corsar D, Garijo D, Soiland-Reyes S, Zednik S, Zhao J, et al. (2013) PROV-O: The PROV Ontology. W3C Recommendation[In English]. URL: https://www.w3.org/TR/2013/REC-prov-o-20130430/
- Brickley D, Miller L (2014) FOAF Vocabulary Specification 0.99. XMLNS URL: https://xmlns.com/foaf/spec/
- Burton A, Koers H, Manghi P, Socker M, Fenner M, Aryani A, La Bruzo S, Diepenbroek M, Schindler U (2017) The Scholix framework for interoperability in data-literature information exchange. D-Lib Magazine 23 (1/2). URL: http://mirror.dlib.org/dlib/january17/burton/01burton.html
- Chebi-Administrators (2005) Physico-chemical methods and properties: FIX ontology. Chebi URL: http://purl.obolibrary.org/obo/fix.owl
- Coles S, Knight N (2022) AI3SD video: physical sciences data infrastructure: shaping the physical sciences roadmap. University of Southampton. https://doi.org/10.5258/SOTON/AI3SD0208
- Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M (2011) The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. PLoS ONE 6 (10). https://doi.org/10.1371/journal.pone.0025513
- Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2015) ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 44 https://doi.org/10.1093/nar/gkv1031
- Nieva de la Hidalga A, Goodall J, Anyika C, Matthews B, Catlow CRA (2022a) Designing a data infrastructure for catalysis science aligned to FAIR data principles. Catalysis Communications 162 https://doi.org/10.1016/j.catcom.2021.106384
- Nieva de la Hidalga A, Decarolis D, Xu S, Matam S, Enciso WYH, Goodall J, Matthews B, Catlow CRA (2022b) A Workflow Demonstrator for Processing Catalysis Research Data. Data Intelligence 4 (2): 455-470. https://doi.org/10.1162/dint_a_00143
- Peroni S, Shotton D (2018) The SPAR Ontologies. Lecture Notes in Computer Science119-136. https://doi.org/10.1007/978-3-030-00668-6_8
- Reynolds D (2014) The Organization Ontology (W3C Recommendation. World Wide Web Consortium URL: https://www.w3.org/TR/vocab-org/
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science1-42. https://doi.org/10.3233/ds-210053
- Soldatova L, King R, Clare A (2006) Ontology of scientific experiments (EXPO). sourcegforge URL: http://expo.sourceforge.net/

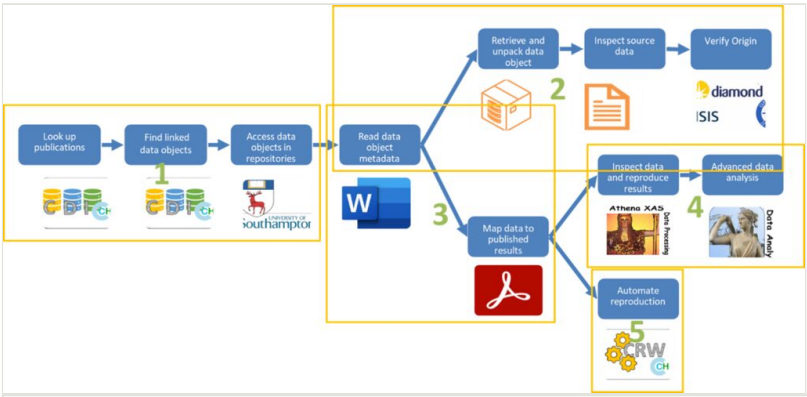

Figure 1.

Reproduction Experiment. 1 Find Linked Data Objects. 2. Track data provenance (verify where published data comes from). 3 Map published data to results. 4 Reproduce Results (learn how results were obtained). 5 Automate Reproduction and Use alternative software.

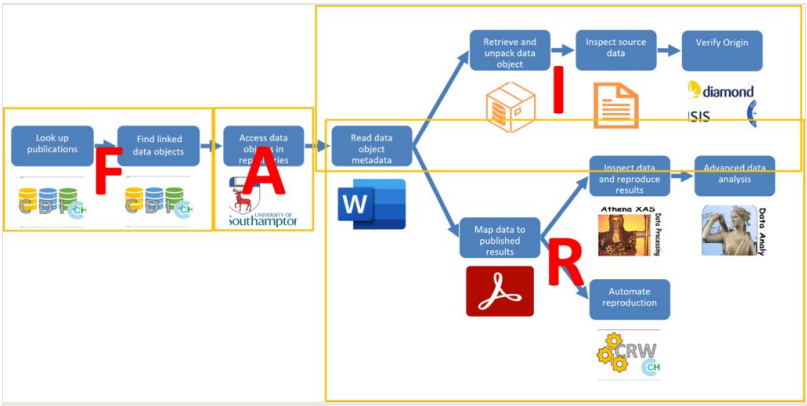

Figure 2.

Alignment of published catalysis research data to the FAIR data principles.