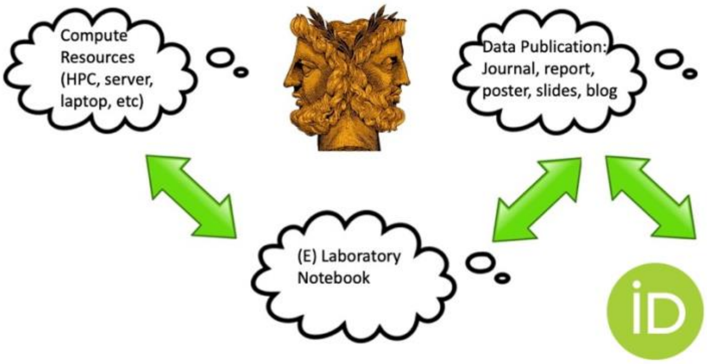
Teaching FAIR in Computational Chemistry: Managing and publishing data using the twin tools of Compute Portals and Repositories.

Henry S. Rzepa

Department of Chemistry, Molecular Sciences Research Hub, Imperial College London,
White City Campus, Wood Lane, London W12 OBZ, UK. ORCID: 0000-0002-8635-8390
Email: rzepa@imperial.ac.uk

Abstract: The history of the development of two tools for managing research resources and the data produced from them is summarised. These tools are a portal or electronic laboratory notebook for computational chemistry interfaced in one direction to a high-performance computing resource and in the other direction to a modern research data repository. The essential features of both these tools are described over two generations of each, with examples of student work cited as examples using persistent identifiers or PIDs, better known as DOIs. Underpinning this is metadata describing the data being processed. The evolution of managing data in this manner over almost two decades and its progress towards what can now be summarised by the acronym FAIR data is outlined.

Graphical Abstract:


Compute Resources (HPC, server, laptop, etc)
Data Publication: Journal, report, poster, slides, blog
(E) Laboratory Notebook
iD

Introduction

The evolution of computational chemistry in all its diverse forms in the modern taught university laboratory curriculum these past two decades has been remarkable. It is also morphing from a stand-alone exercise into one increasingly twinned with the more traditional laboratory "wet chemistry" experiments of molecule preparation and measurement of their properties. All this activity is associated with the collection of large amounts of digital data, which underpins the student interpretation and analysis of the experiment and is becoming part of the evaluation of the student experiment. Teaching the management this data is also evolving, in part by absorbing the changes in how research data itself is nowadays handled. This article[1] focuses on this last aspect by introducing how two software tools have been introduced to revolutionise data management by researchers and students alike in the chemistry department at Imperial College over these last two decades, taking the form of what is now increasingly referred to as the curation of research data in FAIR (Findable, Accessible, Interoperable and Reusable) form and its publication and recognition as a formal research object.

1. The first generation of portal and data repository

In 2005, we started two, initially separate, projects. The first involved deployment of an instance of the Dspace[2] data repository system, which following its first release in 2002 had become reasonably mature software technology. This was chosen as our solution for a data curation project funded by a UK library initiative. Such repositories were starting to become a necessary part of the UK university national research assessment framework and were being used for managing research objects such as (p)reprints and PhD dissertations. The existing instance for this purpose at Imperial College was known as Spiral.[3] Our focus was instead on the capture of primary research data in chemistry, on which the articles deposited into Spiral had been based. This project was called SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data).[4] The initial targets included experimental data relating to NMR, IR and UV-vis spectroscopies, together with computational data describing molecules and their properties and reactions. We also realised we had to develop procedures for deriving metadata describing the essential properties of the deposited data, which would in turn allow the data to be **F**ound by formulating suitable searches of the repository. The innovations included workflows designed to generate this metadata as automatically as possible and to identify crucial chemical metadata describing the molecular object the data related to. One example of this was to use a user-submitted molecular connection table (as a Molfile or ChemDraw file) to generate an InChI identifier[5] as metadata and incorporating this into a workflow associated with each deposition using the OpenBabel toolkit.[6]

In thinking how we might produce similar workflows for computational chemistry outputs, we soon realised that this task could be combined with another aim of making access to centralised high-performance computing (HPC) resources as user-friendly as possible. Rather than expecting users of this resource to master the complex command-line instructions which HPC services then required, we chose to develop a bespoke web-based solution HPC portal to interface both with the batch control system of the HPC service and to the SPECTRa data repository. The portal would also serve to act as an electronic laboratory notebook (ELN) by preserving both the submitted input files and the generated output files for the user, allowing searches of the submissions made over time and providing an identifier for each computational experiment generated by submission to the data repository.[7] The portal would also use defined workflows to generate suitable computational metadata for the repository. This coupling of three concepts, that of a portal to HPC resources, an ELN for managing day to day activity and a publication data repository for use in citing the resulting data was quite unusual at the time in the chemical sciences (Figure 1).

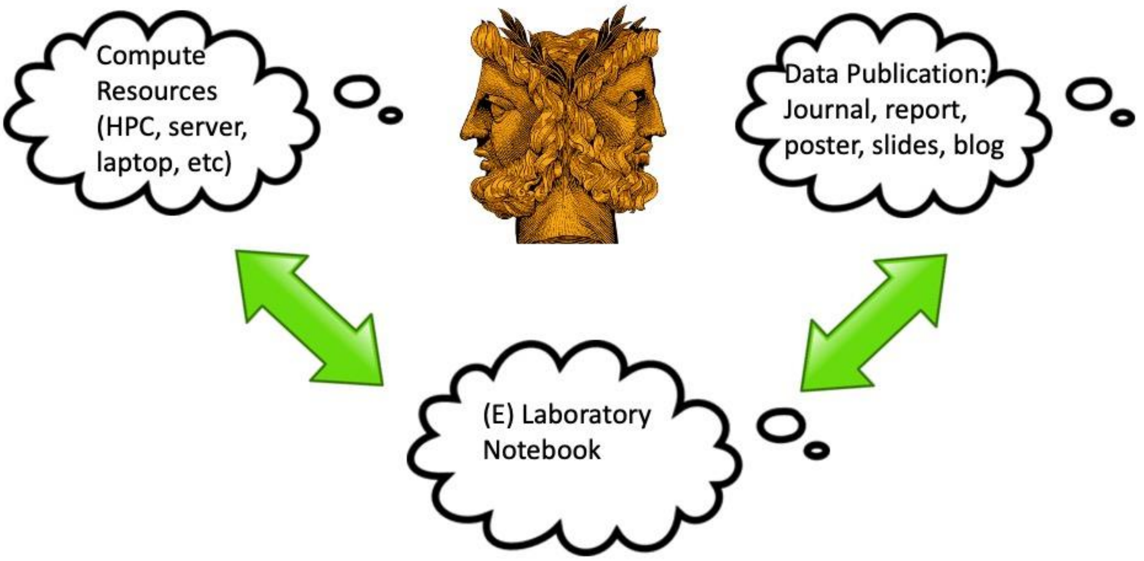
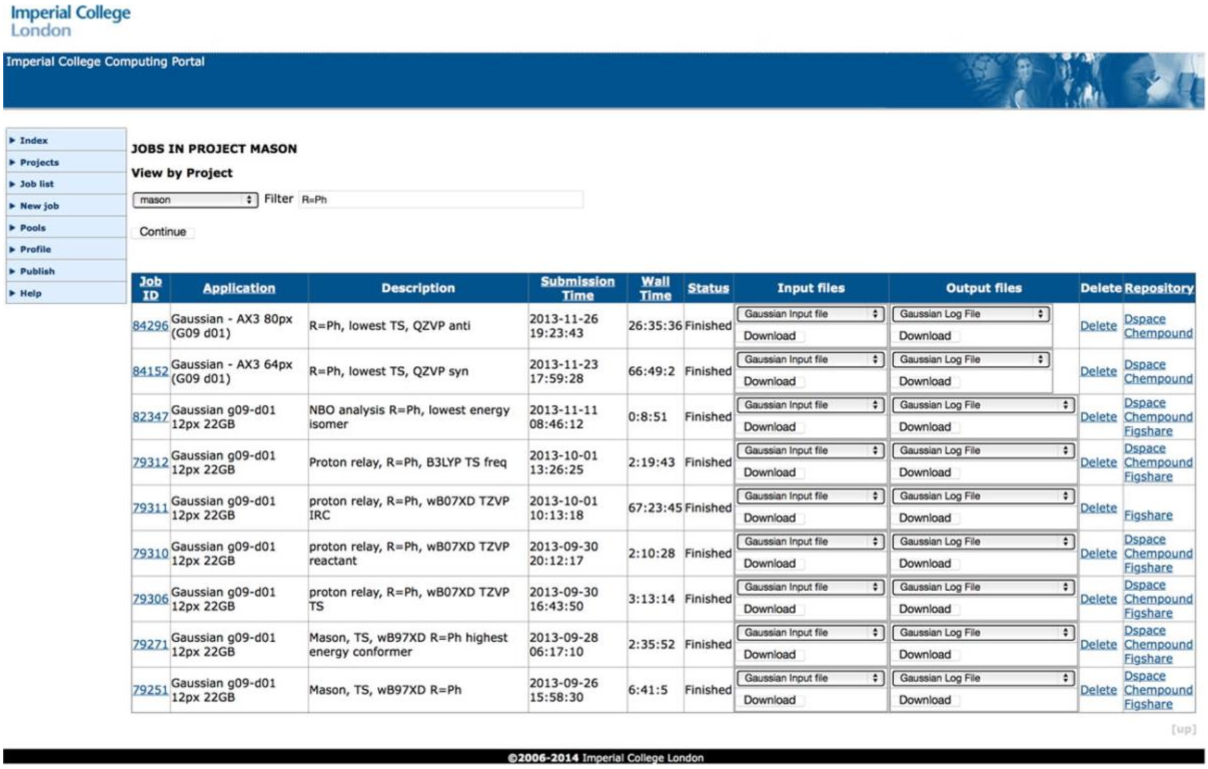

Compute Resources (HPC, server, laptop, etc)
Data Publication: Journal, report, poster, slides, blog
(E) Laboratory Notebook

Figure 1. The Janus-like relationship of the ELN in connecting computing resources with the data publication repository.

As the project developed, we realised that simply generating a locally unique identifier which would be used to identify experiments in both the portal ELN, and repository was very limiting. We decided to ensure that the identifier was what is now called "persistent" (a PID) by using the Handle registration authority[8] to issue a globally unique PID for each deposition. The Handle authority would also register some of the metadata captured using our automated workflows, thus enabling the entire system to move from a local to potentially a globally **F**indable resource. This particular design decision involving generating a PID has turned out to perhaps be the most far-sighted we made at the time.

By 2007, our infrastructure was in place and was ready to be used by undergraduate students as well as researchers. A typical presentation of the ELN to the student or researcher is shown in Figure 2 illustrating from left to right what might be called the flow of the data pipeline and was starting to be used for student research projects. One student's experiences were included in the article describing the operation of the ELN.[7] The final project outputs in the form of PIDs for the computational research data were incorporated into the article[9] describing the actual science (Figure 3). By the time this particular student project was undertaken, we had added the capability of publishing from the ELN not only to our SPECTRa repository but also to the new external service started in 2012 by Figshare[10] (Figure 4).The system was also used for laboratory experiments in which bench chemistry was merged with computational modelling, the latter including *e.g.* predictive calculations of ^{13}C spectra of molecules synthesized by the students and prediction of expected optical rotations for molecules made by enantioselective reactions such as Jacobsen and Shi epoxidations.[11] For some of the more computationally intensive models which would take more time than allocated for the laboratory, we stored pre-computed results in a repository for students to **R**euse and derive conclusions from (Figure 4).



JOBS IN PROJECT MASON

View by Project

mason ⬍ Filter R=Ph

Job ID	Application	Description	Submission Time	Wall Time	Status	Input files	Output files	Delete	Repository
84296	Gaussian - AX3 80px (G09 d01)	R=Ph, lowest TS, QZVP anti	2013-11-26 19:23:43	26:35:36	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound
84152	Gaussian - AX3 64px (G09 d01)	R=Ph, lowest TS, QZVP syn	2013-11-23 17:59:28	66:49:2	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound
82347	Gaussian g09-d01 12px 22GB	NBO analysis R=Ph, lowest energy isomer	2013-11-11 08:46:12	0:8:51	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare
79312	Gaussian g09-d01 12px 22GB	Proton relay, R=Ph, B3LYP TS freq	2013-10-01 13:26:25	2:19:43	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare
79311	Gaussian g09-d01 12px 22GB	proton relay, R=Ph, wB07XD TZVP IRC	2013-10-01 10:13:18	67:23:45	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Figshare
79310	Gaussian g09-d01 12px 22GB	proton relay, R=Ph, wB07XD TZVP reactant	2013-09-30 20:12:17	2:10:28	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare
79306	Gaussian g09-d01 12px 22GB	proton relay, R=Ph, wB07XD TZVP TS	2013-09-30 16:43:50	3:13:14	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare
79271	Gaussian g09-d01 12px 22GB	Mason, TS, wB97XD R=Ph highest energy conformer	2013-09-28 06:17:10	2:35:52	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare
79251	Gaussian g09-d01 12px 22GB	Mason, TS, wB97XD R=Ph	2013-09-26 15:58:30	6:41:5	Finished	Gaussian Input file ⬍ Download	Gaussian Log File ⬍ Download	Delete	Dspace Chempound Figshare

©2006-2014 Imperial College London

Figure 2. The presentation of the ELN as a computing portal, showing each experiment as a pipeline ending in the data repository entry for that experimental computation.[7]

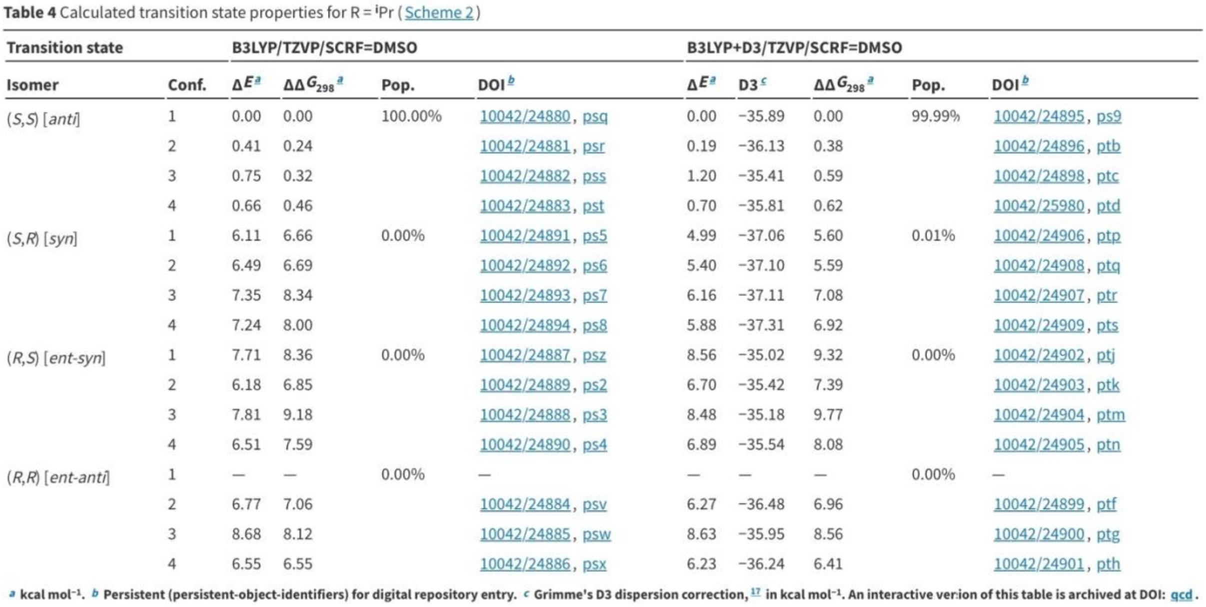
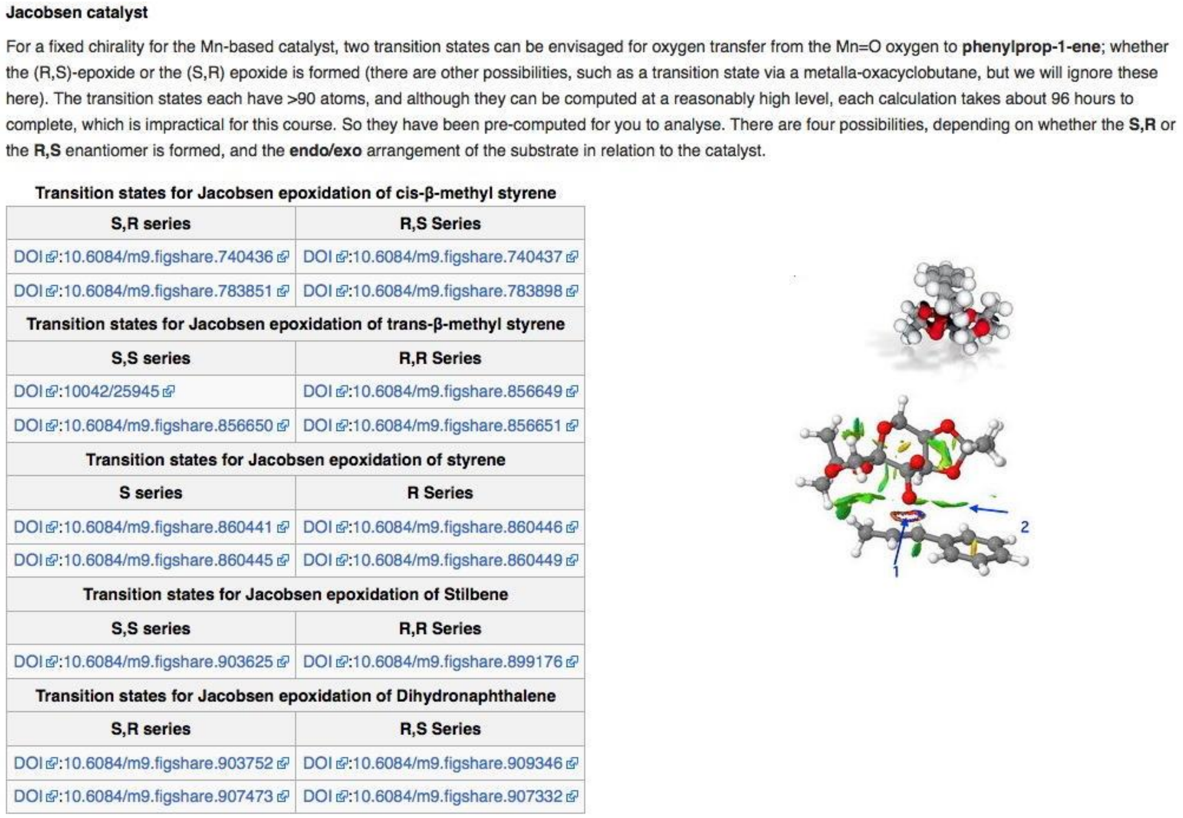


Table 4 Calculated transition state properties for R = ^{i}Pr (Scheme 2)

Transition state		B3LYP/TZVP/SCRF=DMSO				B3LYP+D3/TZVP/SCRF=DMSO				
Isomer	Conf.	ΔE [a]	$\Delta\Delta G_{298}$ [a]	Pop.	DOI [b]	ΔE [a]	D3 [c]	$\Delta\Delta G_{298}$ [a]	Pop.	DOI [b]
(*S*,*S*) [*anti*]	1	0.00	0.00	100.00%	10042/24880, psq	0.00	−35.89	0.00	99.99%	10042/24895, ps9
	2	0.41	0.24		10042/24881, psr	0.19	−36.13	0.38		10042/24896, ptb
	3	0.75	0.32		10042/24882, pss	1.20	−35.41	0.59		10042/24898, ptc
	4	0.66	0.46		10042/24883, pst	0.70	−35.81	0.62		10042/25980, ptd
(*S*,*R*) [*syn*]	1	6.11	6.66	0.00%	10042/24891, ps5	4.99	−37.06	5.60	0.01%	10042/24906, ptp
	2	6.49	6.69		10042/24892, ps6	5.40	−37.10	5.59		10042/24908, ptq
	3	7.35	8.34		10042/24893, ps7	6.16	−37.11	7.08		10042/24907, ptr
	4	7.24	8.00		10042/24894, ps8	5.88	−37.31	6.92		10042/24909, pts
(*R*,*S*) [*ent-syn*]	1	7.71	8.36	0.00%	10042/24887, psz	8.56	−35.02	9.32	0.00%	10042/24902, ptj
	2	6.18	6.85		10042/24889, ps2	6.70	−35.42	7.39		10042/24903, ptk
	3	7.81	9.18		10042/24888, ps3	8.48	−35.18	9.77		10042/24904, ptm
	4	6.51	7.59		10042/24890, ps4	6.89	−35.54	8.08		10042/24905, ptn
(*R*,*R*) [*ent-anti*]	1	—	—	0.00%	—	—	—	—	0.00%	—
	2	6.77	7.06		10042/24884, psv	6.27	−36.48	6.96		10042/24899, ptf
	3	8.68	8.12		10042/24885, psw	8.63	−35.95	8.56		10042/24900, ptg
	4	6.55	6.55		10042/24886, psx	6.23	−36.24	6.41		10042/24901, pth

[a] kcal mol^{-1}. [b] Persistent (persistent-object-identifiers) for digital repository entry. [c] Grimme's D3 dispersion correction,[17] in kcal mol^{-1}. An interactive ver:ion of this table is archived at DOI: qcd.

Figure 3. Illustration of how the PIDs generated using the SPECTRa data repository were incorporated into published articles.[9]



Jacobsen catalyst

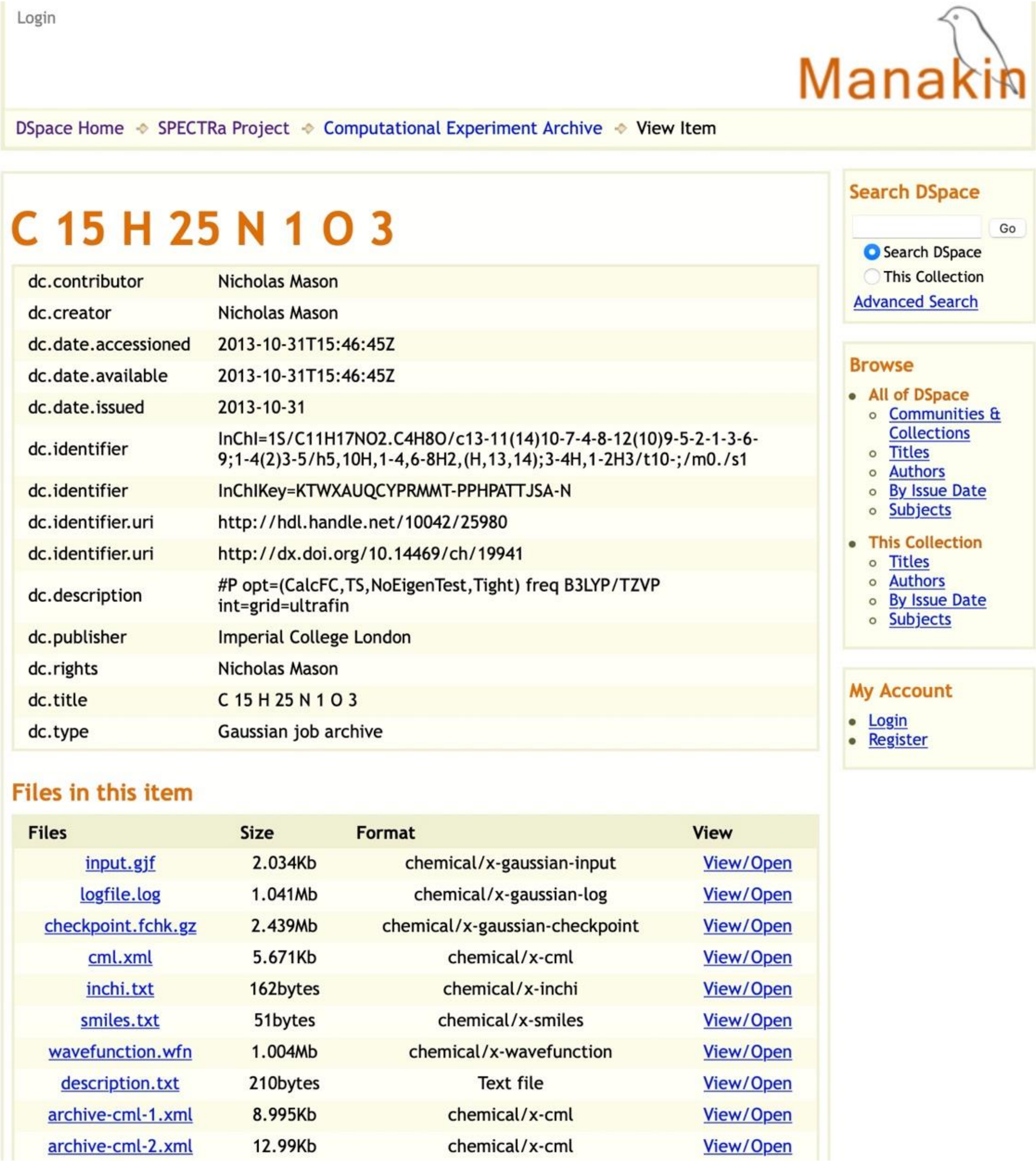
For a fixed chirality for the Mn-based catalyst, two transition states can be envisaged for oxygen transfer from the Mn=O oxygen to **phenylprop-1-ene**; whether the (R,S)-epoxide or the (S,R) epoxide is formed (there are other possibilities, such as a transition state via a metalla-oxacyclobutane, but we will ignore these here). The transition states each have >90 atoms, and although they can be computed at a reasonably high level, each calculation takes about 96 hours to complete, which is impractical for this course. So they have been pre-computed for you to analyse. There are four possibilities, depending on whether the **S,R** or the **R,S** enantiomer is formed, and the **endo/exo** arrangement of the substrate in relation to the catalyst.

Transition states for Jacobsen epoxidation of cis-β-methyl styrene	
S,R series	**R,S Series**
DOI:10.6084/m9.figshare.740436	DOI:10.6084/m9.figshare.740437
DOI:10.6084/m9.figshare.783851	DOI:10.6084/m9.figshare.783898
Transition states for Jacobsen epoxidation of trans-β-methyl styrene	
S,S series	**R,R Series**
DOI:10042/25945	DOI:10.6084/m9.figshare.856649
DOI:10.6084/m9.figshare.856650	DOI:10.6084/m9.figshare.856651
Transition states for Jacobsen epoxidation of styrene	
S series	**R Series**
DOI:10.6084/m9.figshare.860441	DOI:10.6084/m9.figshare.860446
DOI:10.6084/m9.figshare.860445	DOI:10.6084/m9.figshare.860449
Transition states for Jacobsen epoxidation of Stilbene	
S,S series	**R,R Series**
DOI:10.6084/m9.figshare.903625	DOI:10.6084/m9.figshare.899176
Transition states for Jacobsen epoxidation of Dihydronaphthalene	
S,R series	**R,S Series**
DOI:10.6084/m9.figshare.903752	DOI:10.6084/m9.figshare.909346
DOI:10.6084/m9.figshare.907473	DOI:10.6084/m9.figshare.907332

Figure 4. A collection of precomputed transition state structures associated with a student laboratory experiment in the form of DOIs (digital object identifiers), with a derived property shown on the right in the form of a non-covalent-interaction isosurface.

In 2012, we extended the use of Handles as data repository identifiers to the now more familiar DOI (digital object identifier), in this instance issued by the DataCite registration authority. Such DOIs until then had been associated exclusively with journal articles and the CrossRef registry, but their use to identify data was now being extended to join the more general identifier or PID ecosystem. This allowed a larger and more flexible metadata record designed specifically to describe data to be associated with each item, as defined by the DataCite schema.[12] This meant that the metadata could **F**ound or searched using the globally aggregated and indexed DataCite metadata store. These DOIs were included by students in their final laboratory report as a means of facilitating **A**ccess to their data[11] to the assessor of their report. It also enabled the data to be cited in any resulting journal publication (Figure 3). Another unique feature for that time was the inclusion of calculation checkpoint files as part of the fileset associated with a quantum calculation (Figure 5), which in turn allowed facile (re)generation of properties associated with the calculation such as *e.g.* non-covalent interaction surfaces[13] (as seen in Figure 3). The strategy here was to allow the data retrieved from a repository to be repurposed or **I**nteroperated. Here and in preceding paragraphs, we have already referred to attributes such as **F**inding, **A**ccessing and **R**eusing and these became collectively known as **FAIR** data in 2015.[14] FAIR data in turn is increasingly required by *e.g.* funders and recently journals, making it nowadays an essential component of a taught laboratory in an experimental science such as chemistry.

Another advantage of using DOI identifiers, as demonstrated by a student project was the ability to include them in interactive versions of journal tables and figures (Figure 3).[9] We had been experimenting with such objects for some years,[15] but the introduction of Handles and later DOIs allowed these tables to be constructed from data stored in a repository, to create for example rotatable models of molecules deriving from the calculations.[16] During the period 2005-present, some 73 research articles were enhanced with such tables and figures.[17]



C 15 H 25 N 1 O 3

dc.contributor	Nicholas Mason
dc.creator	Nicholas Mason
dc.date.accessioned	2013-10-31T15:46:45Z
dc.date.available	2013-10-31T15:46:45Z
dc.date.issued	2013-10-31
dc.identifier	InChI=1S/C11H17NO2.C4H8O/c13-11(14)10-7-4-8-12(10)9-5-2-1-3-6-9;1-4(2)3-5/h5,10H,1-4,6-8H2,(H,13,14);3-4H,1-2H3/t10-;/m0./s1
dc.identifier	InChIKey=KTWXAUQCYPRMMT-PPHPATTJSA-N
dc.identifier.uri	http://hdl.handle.net/10042/25980
dc.identifier.uri	http://dx.doi.org/10.14469/ch/19941
dc.description	#P opt=(CalcFC,TS,NoEigenTest,Tight) freq B3LYP/TZVP int=grid=ultrafin
dc.publisher	Imperial College London
dc.rights	Nicholas Mason
dc.title	C 15 H 25 N 1 O 3
dc.type	Gaussian job archive

Files in this item

Files	Size	Format	View
input.gjf	2.034Kb	chemical/x-gaussian-input	View/Open
logfile.log	1.041Mb	chemical/x-gaussian-log	View/Open
checkpoint.fchk.gz	2.439Mb	chemical/x-gaussian-checkpoint	View/Open
cml.xml	5.671Kb	chemical/x-cml	View/Open
inchi.txt	162bytes	chemical/x-inchi	View/Open
smiles.txt	51bytes	chemical/x-smiles	View/Open
wavefunction.wfn	1.004Mb	chemical/x-wavefunction	View/Open
description.txt	210bytes	Text file	View/Open
archive-cml-1.xml	8.995Kb	chemical/x-cml	View/Open
archive-cml-2.xml	12.99Kb	chemical/x-cml	View/Open

Figure 5. The fileset associated with a quantum calculation as deposited into a data repository, with associated Media types and metadata.

2. The second generations of portal and data repository.

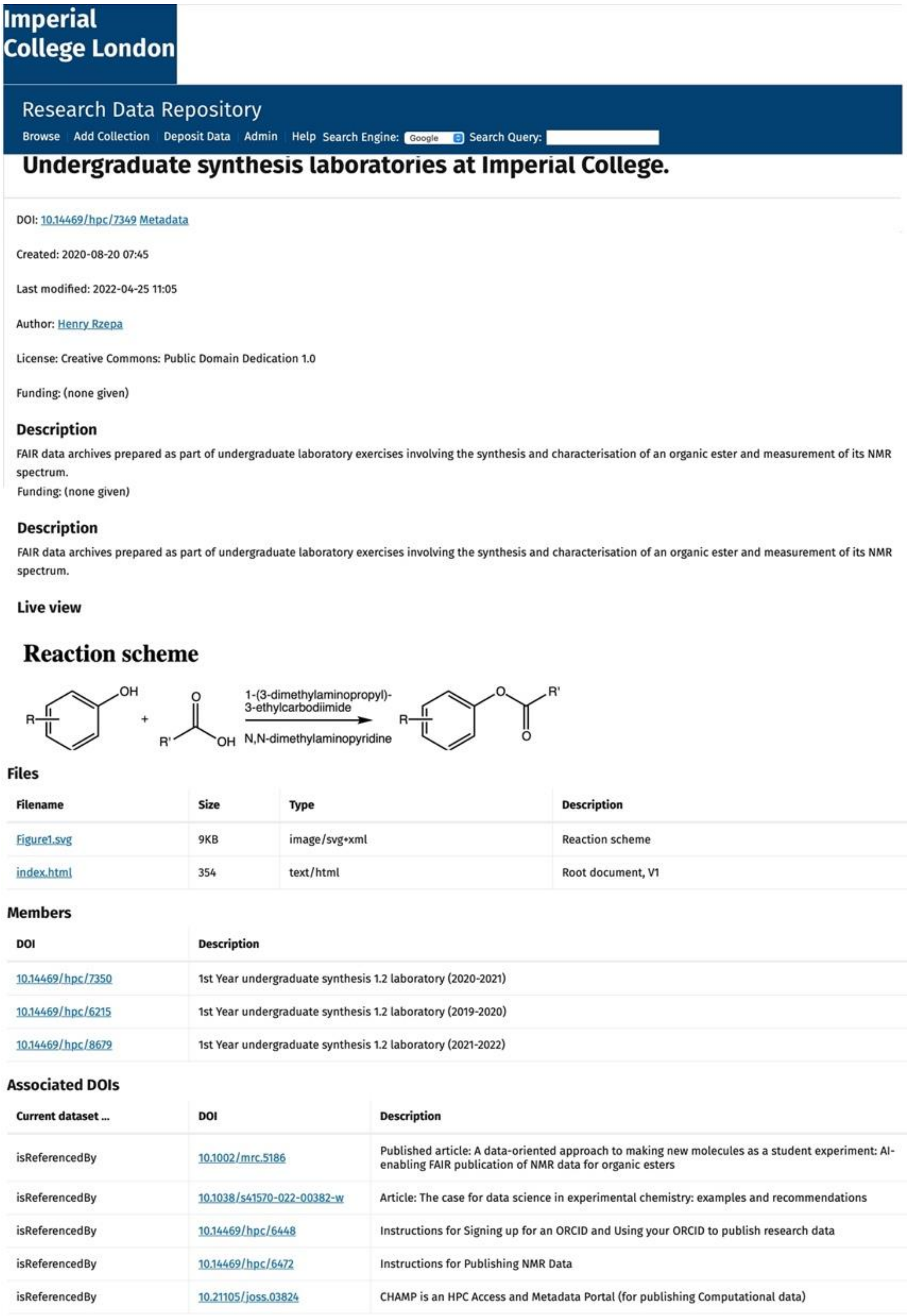
By 2016, we realised a new generation of data repository was needed to accompany the ELN. The DataCite metadata schema[12] was constantly being extended and we needed to match the new capability this offered with the functionality of the repository – the existing Dspace technology used for SPECTRa was not then considered as easy to adapt for this purpose. Further persistent identifiers such as ORCID[18] had now been introduced and we wanted to make their use mandatory with the repository. Finally, the Dspace technology itself was considered non-trivial to upgrade and we needed a more lightweight tool.

The resulting new repository[19] offered further features. These included:

1. Further metadata workflows to *e.g.* capture the Gibbs Energy from a computation that included vibrational frequencies - we had realised that this energy (and associated total energies) made for excellent identifiers of a calculation, in that the numerical values could be searched using the DataCite search feature.[20]
2. Media types[21] were another searchable file attribute that served a useful purpose for identifying different types of chemical data.[20] Further examples of such FAIR searches are collected here.[22]
3. Since including an ORCID (Open Researcher and Collaborator Identifier) was now part of the repository initiation procedure, students – as researchers - could now get full and open credit for their activity by automated linking to their ORCID record.[23]
4. We recognized that research data increasingly requires sophisticated software to allow interpretation, but that such software is often only available commercially. In collaboration with the software vendor MestreLabs, we introduced a novel single-use licensing system called MNPUB.[24] Each dataset identified as NMR data is assigned a digitally signed authority for unrestricted use of the Mnova software, albeit for that dataset only, without requiring the program itself to be commercially licensed. The MNPUB file itself is acquired from the repository and when loaded into the Mnova software, it assigns a license and then fetches the relevant dataset for re-use. This feature also introduces the concept of having access to lossless original instrumental data, often also referred to as raw or primary data, rather than the considerably reduced or lossy spectral version. Unless the original data is simply too large to be so captured, there is no compelling reason nowadays not to offer this complete version of the data.[25]
5. Continuing the previous point, the new repository now regularly carries original x-ray diffraction images as well as the processed crystal structure data files. The former carries sufficient information to objectively reproduce the original structure determination. An example of student work with such data can be seen here.[26]
6. The new repository extended the concept of data collections. This allowed students to contribute to a class experiment, such as that shown in Figure 6.[27] Each year's class collection itself has a DOI assigned[28] with its members being the individual student contributions and the overall experiment containing the year collections.[29] The student submissions of the NMR spectra of their newly synthesized and original compounds also exploited the point made in item 4 above, containing not only a spectrum but also the FID (Free Induction Decay) as captured by the NMR spectrometer. The FAIR attributes could also allow *e.g.* an AI/ML (artificial intelligence and machine learning) algorithm in the

future to retrieve the original data and in an automated manner reanalyse it for *e.g.* structural correctness or purity.[27]

7. The repository itself could now host the interactive data tables[30] as used in an article publication.[31] Importantly, the registered metadata can formally link such data to the published article, and in the future the link is expected to become bidirectional with the article linking **via** a citation to the data.



Imperial College London

Research Data Repository

Undergraduate synthesis laboratories at Imperial College.

DOI: 10.14469/hpc/7349 Metadata

Created: 2020-08-20 07:45

Last modified: 2022-04-25 11:05

Author: Henry Rzepa

License: Creative Commons: Public Domain Dedication 1.0

Funding: (none given)

Description

FAIR data archives prepared as part of undergraduate laboratory exercises involving the synthesis and characterisation of an organic ester and measurement of its NMR spectrum.

Funding: (none given)

Description

FAIR data archives prepared as part of undergraduate laboratory exercises involving the synthesis and characterisation of an organic ester and measurement of its NMR spectrum.

Live view

Reaction scheme

Files

Filename	Size	Type	Description
Figure1.svg	9KB	image/svg+xml	Reaction scheme
index.html	354	text/html	Root document, V1

Members

DOI	Description
10.14469/hpc/7350	1st Year undergraduate synthesis 1.2 laboratory (2020-2021)
10.14469/hpc/6215	1st Year undergraduate synthesis 1.2 laboratory (2019-2020)
10.14469/hpc/8679	1st Year undergraduate synthesis 1.2 laboratory (2021-2022)

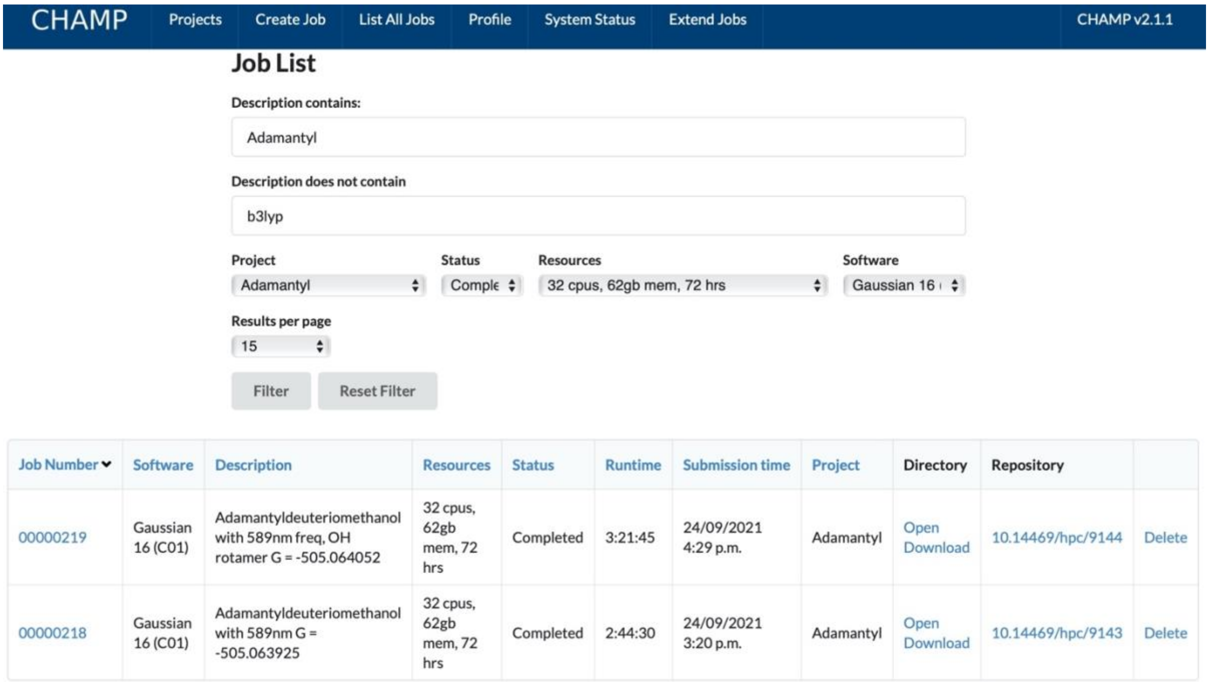
Associated DOIs

Current dataset ...	DOI	Description
isReferencedBy	10.1002/mrc.5186	Published article: A data-oriented approach to making new molecules as a student experiment: AI-enabling FAIR publication of NMR data for organic esters
isReferencedBy	10.1038/s41570-022-00382-w	Article: The case for data science in experimental chemistry: examples and recommendations
isReferencedBy	10.14469/hpc/6448	Instructions for Signing up for an ORCID and Using your ORCID to publish research data
isReferencedBy	10.14469/hpc/6472	Instructions for Publishing NMR Data
isReferencedBy	10.21105/joss.03824	CHAMP is an HPC Access and Metadata Portal (for publishing Computational data)

Figure 6. The top-level repository data collection for a laboratory experiment, showing the annual year versions, to which student depositions are linked.

After six years in use, a third generation of repository is now being built, this time to flexibly accommodate new FAIR data standards being considered by an IUPAC working party[32] allowing richer and more complete metadata to be specified in conformance with *e.g.* the DataCite schema,[12] as well as being based on a general repository toolkit for future ease of maintenance.[33] It is expected that further working parties will be constituted to consider metadata standards specifically for various areas of computational chemistry.

The maintenance aspects also caused us to refactor our ELN in 2021.[34] Version 1 had been written in 2006 and based on locally written code using the php language, but this has now been replaced with Version 2 called CHAMP (Figure 7) as based on a toolkit written in the more future-proof python language that had been published a year or so earlier,[35] and where the design philosophy was similar to our original ELN.[7]



Job List

Description contains: Adamantyl

Description does not contain: b3lyp

Project: Adamantyl

Status: Comple

Resources: 32 cpus, 62gb mem, 72 hrs

Software: Gaussian 16 (

Results per page: 15

Job Number	Software	Description	Resources	Status	Runtime	Submission time	Project	Directory	Repository	
00000219	Gaussian 16 (C01)	Adamantyldeuteriomethanol with 589nm freq, OH rotamer G = -505.064052	32 cpus, 62gb mem, 72 hrs	Completed	3:21:45	24/09/2021 4:29 p.m.	Adamantyl	Open Download	10.14469/hpc/9144	Delete
00000218	Gaussian 16 (C01)	Adamantyldeuteriomethanol with 589nm G = -505.063925	32 cpus, 62gb mem, 72 hrs	Completed	2:44:30	24/09/2021 3:20 p.m.	Adamantyl	Open Download	10.14469/hpc/9143	Delete

Figure 7. The CHAMP portal/ELN for computational chemistry, illustrating the data pipeline.

The improvements include better search features for filtering experiments, inspection of job outputs in real time, the ability to construct a user profile which allows *e.g.* resources provided by funded projects to be used and includes the ability to publish using repositories such as Zenodo. CHAMP has already been used for student research projects, proving particularly convenient for both researchers and students during the COVID19 lockdowns.[31]

3. Conclusions.

For almost two decades now, students have had an opportunity at Imperial College to practise modern data management techniques using two tools, the Janus-like computational portal or ELN, where one face links to high performance facilities and the other face to data publication repositories. Students and researchers alike have learnt about the importance of persistent identifiers, increasingly used in future-facing knowledge graphs[36] and which in turn will be a central feature of developments in AI/ML. More hidden from view is perhaps the increasingly sophisticated metadata associated with data publications, but which nevertheless allows a researcher's work to be readily found and re-used by others. As the repositories and their embedded data and metadata workflows become further enriched, one might anticipate a much wider diversity of data will be added, including more forms of spectroscopy and other instrumental methods. As both students and researchers develop their portfolio of what are referred to as "works" by ORCID, we might imagine their future career could be influenced not only by their published journal works, but by their registered activities in data publication.

Acknowledgements. I thank the EPSRC for funding with grant EP/T030534/1. Dedicated to the memory of Nick Mason, who as a student contributed greatly to these projects.

Data availability statement. All the PIDs for the primary research data referenced in this article are collected at DOI: 10.14469/hpc/11240

[1] Based on an invited presentation at the WATOC congress, Vancouver, July 2022, and available for viewing at DOI: https://doi.org/hsgk
[2] Dspace. See https://dspace.lyrasis.org for an overview.
[3] Spiral. See https://spiral.imperial.ac.uk/about.jsp for an overview.
[4] J. Downing, P. Murray-Rust, A. P. Tonge, P. Morgan, H. S. Rzepa, F. Cotterill, N. Day, and M. J. Harvey, *J. Chem. Inf. Mod***.**, **2008**, *48*, 1571-1581. DOI: 10.1021/ci7004737
[5] See https://www.inchi-trust.org for an overview and S. R. Heller, A. McNaught, I. Pletnev, S. Stein and D. Tchekhovskoi, *J Cheminform*, **2015**, *7*, 23. DOI: 10.1186/s13321-015-0068-4
[6] N. M. O'Boyle, M. Banck, C.A. James, et al. *J Cheminform*, **2011**, *3*, 33. DOI: 10.1186/1758-2946-3-33
[7] M. J. Harvey, N. J. Mason and H. S. Rzepa, *J. Chem. Inf. Model*., **2014**, *54*, 2627-2635. DOI: 10.1021/ci500302p
[8] See https://en.wikipedia.org/wiki/Handle_System for an excellent overview and https://www.dona.net for the current administrators.
[9] A. Armstrong, R. A. Boto, P. Dingwall, J. Contreras-García, M. J. Harvey, N. Mason and H. S. Rzepa, *Chem. Sci*., **2014**, *5*, 2057-2071. dataDOI: DOI: 10.1039/C3SC53416B
[10] Figshare data repository. See https://knowledge.figshare.com/about for an overview.
[11] E. H. Smith, H. S. Rzepa and M. Hii, *J. Chem. Ed*, **2015**, *92, 1385-1389.* DOI: 10.1021/ed500398e For a representative student report for the experiment illustrating the use of PIDs, see DOI: https://doi.org/hz84
[12] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V., **2021**, DOI: 10.14454/3w3z-sa82
[13] E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen, and W. Yang, *J. Am. Chem. Soc*., **2010**, *132*, 6498-6506 DOI: 10.1021/ja100936w and H. S. Rzepa, Script for creating an NCI surface as a JVXL compressed file from a (Gaussian) cube of total electron density, DOI: 10.14469/hpc/3660
[14] M. Wilkinson, M. Dumontier, I. Aalbersberg *et al*,, *Sci Data*, **2016**, *3*, 160018. DOI: 10.1038/sdata.2016.18
[15] H. S. Rzepa "Chemistry with a twist blog", **2022**, DOI: 10.14469/hpc/10850 and appended comments.
[16] See A. Armstrong, R. A. Boto, P. Dingwall, J. Contreras-García, M. J. Harvey, N. Mason and H. S. Rzepa, for an interactive version of Table 1 taken from ref 9, **2014**, DOI: 10.14469/hpc/11014
[17] H. S. Rzepa, Imperial College Research Data Repository, **2022**, DOI: 10.14469/hpc/11238
[18] ORCID, see https://orcid.org/ for an overview.
[19] M. J. Harvey, A. McLean, and H. S. Rzepa, **2017**, *9*, 4. DOI: 10.1186/s13321-017-0190-6.
[20] A. N. Davies and H. S. Rzepa, *Spectros. Eu.*, **2022**, DOI: 10.1255/sew.2022.a10
[21] H. S. Rzepa, P. Murray-Rust and B. J. Whitaker, J. *Chem. Inf. Comp. Sci*., **1998**, *38*, 976-982.
[22] H. S. Rzepa, Imperial College Research Data Repository, **2022**, DOI: 10.14469/hpc/11109
[23] An example of an ORCID record, https://commons.datacite.org/orcid.org/0000-0001-9787-8853
[24] A. Barba, S. Dominguez, C. Cobas, D.P. Martinsen, C. Romain, H. S. Rzepa and F. Seoane, *ACS Omega*, **2019**, *4*, 3280-3286. DOI: 10.1021/acsomega.8b03005, dataDOI: 10.14469/hpc/4751

[25] For an early example of student-derived NMR data, see J. Clarke, K. J. Bonney, M. Yaqoob, S. Solanki, H. S. Rzepa, A. J. P. White, D. S. Millan, and D. C. Braddock, *J. Org. Chem*. **2016**, **81**, 20, 9539–9552 10.1021/acs.joc.6b02008 with the NMR data available at DOI: 10.14469/hpc/1267

[26] D. C. Braddock, N. Limpaitoon, K. Oliwa, D. O'Reilly, H. S. Rzepa and A. J. P. White, Imperial College Research Data Repository, **2022**, 10.14469/hpc/9241

[27] H. S. Rzepa and S. Kuhn, *Mag. Res. Chem*., **2021**, DOI: 10.1002/mrc.5186

[28] See a year collection for 2019-2020 at DOI: 10.14469/hpc/6215 and a student contribution at DOI: 10.14469/hpc/7122

[29] See the overall laboratory experiment repository collection at DOI: 10.14469/hpc/7349

[30] D. C. Braddock, N. Limpaitoon, K. Oliwa, D. O'Reilly, H. S. Rzepa and A. J. P. White, Imperial College Research Data Repository, **2022**, DOI: 10.14469/hpc/9869

[31] D. C. Braddock, N. Limpaitoon, K. Oliwa, D. O'Reilly, H. S. Rzepa and A. J. P. White , *Chem. Comm*., **2022**, *58*, 4981-4984.DOI: 10.1039/D2CC01136K

[32] R. M. Hanson, D. Jeannerat, M. Archibald, I. Bruno, S. J. Chalk, A. N. Davies, R. J. Lancashire J. Lang and H. S. Rzepa, *Pure App. Chem*., **2022**, *94*, 623-636. DOI: 10.1515/pac-2021-2009.

[33] Invenio. See https://inveniosoftware.org/about/ for an overview.

[34] C. Cave-Ayland, M. Bearpark, C. Romain and H. S. Rzepa, *J. Open Source Software*, **2022**, *7*, 3824. DOI: 10.21105/joss.03824

[35] D. Hudak *et al*, *J. Open Source Software*, **2018**, *3*, 622. DOI: 10.21105/joss.00622

[36] H. Cousijn, R. Braukmann, M. Fenner, C. Ferguson, R. van Horik, R. Lammey, A. Meadows and S. Lambert, *Patterns*, **2020**, *2*, 100180. DOI: 10.1016/j.patter.2020.100180