ORIGINAL ARTICLE
Journal Section

A data-oriented approach to making new molecules as a student experiment: AI-enabling FAIR publication of NMR data for organic esters

Henry S. Rzepa[1] | **Stefan Kuhn**[2]

[1]Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, UK

[2]School of Computer Science and Informatics, De Montfort University, Leicester, UK

Correspondence
Henry S. Rzepa, Imperial College London, White City Campus, 82 Wood Lane, London W12 OBZ
Email: rzepa@imperial.ac.uk

Funding information
HSR thanks the EPSRC for funding, grant EP/T030534/1

The lack of machine-readable data is a major obstacle in the application of NMR in artificial intelligence. As a way to overcome this, a procedure for capturing primary NMR Spectroscopic instrumental data annotated with rich metadata and publication in a FAIR data repository is described as part of an undergraduate student laboratory experiment in a chemistry department. This couples the techniques of chemical synthesis of a never before made organic ester with illustration of modern data management practices and serves to raise student awareness of how FAIR data might improve research quality and replicability. Searches of the registered metadata are shown which enable actionable Finding and Accessing of such data. The potential for Re-use of the data in AI-applications is discussed.

KEYWORDS
FAIR, NMR Spectroscopy, Data Repository, Metadata Registration, Re-use, Artificial Intelligence, Chemical Education

1 | INTRODUCTION

The basis for artificial intelligence (AI) is data. This is evidently the case for machine learning methods, be it supervised or unsupervised, since those methods rely on inferring rules from large bodies of data. In a similar fashion, the application of other techniques, e.g. signal processing or image analysis, needs enough examples to be successful. In our case, the data in question are Nuclear Magnetic Resonance (NMR) records and their associated metadata. In practice, these

data are mostly reported in a style that they cannot readily by used for AI and other research. Attempts to solve this problem have failed to gain traction so far and are not widely applied. In this article, we explore these themes using a case study based on one emerging modern model for the acquisition and management [1] of what is called primary data from an NMR instrument (a spectrometer) and its subsequent life-cycle. This cycle includes transformation into a processed dataset to enable information to be extracted and derived for the system being studied (the "molecule") and then its management as a so-called FAIR [2] data set to enable its re-use for AI applications.

NMR is one technique of many used in organic and physical chemistry. The data generated by an NMR experiment is transformed into information associated with a chemical substance. This substance could be a single chemical compound with no significant impurities, but it could also be a mixture of different compounds in varying concentrations which would then be regarded as an impure compound. Apart from applying AI to NMR data for inferring information about chemical structure, it can also help with data quality. Quality is an important aspect of any data collection and a combination of data science and AI techniques can help to identify issues such as outliers, unexpected patterns in data, impurities in the chemical sample studied and indeed fraudulent manipulation of spectra. By identifying the data quality, a virtuous circle of data collection and data usage can be completed.

This article is structured as follows. In Section 2, we describe the characteristics of data produced when collected using an NMR instrument and then give an overview of the current situation and identify strengths and shortcomings of some approaches for disseminating the data for scientific re-use, in AI as well as in other areas. Section 3 introduces the management of data generated during an undergraduate teaching module as a case study for NMR data recording and publication according to FAIR (Findable, Accessible, Interoperable and Reusable) Data principles. [2] In Section 4 we give results derived from the case study and discuss recommendations and lessons learnt. In Section 5, we summarize the paper.

2 | BASICS

NMR works by exposing molecules to a strong constant magnetic field and a weak oscillating field delivered as a pulse sequence, causing the nuclei in the atoms to produce a signal at characteristic frequencies. Modern NMR spectroscopy produces a variety of one or two-dimensional data depending on the precise form of the pulse sequence. Those frequencies are measured or acquired in the first instance as a time domain signal (one evolving over time, the FID or Free-Induction-Decay). This so-called primary instrumental data is then processed using convenient Fourier Transform methods, with appropriate weighting of the data, into a frequency domain in which form it is usually referred to as an NMR spectrum. In such a spectrum, the peaks (called chemical shifts in frequency relative to a reference compound) are associated with particular atomic nuclei or collections of such nuclei in an identical environment. This information is then augmented with derived data such as coupling constants, these being perturbations in the magnetic environment of one nucleus induced by the proximity of other nuclei, along with peak intensities and peak shapes.

Once the transformation of the data from the time to the frequency domain is completed, the traditional approach has been to either discard the original primary FID data or at best to archive it locally and probably privately in a manner that rarely has any associated descriptive metadata. Only the remaining processed spectral data is then reported in the scientific literature, often either in condensed tabular or loosely-structured text form or in visual form as a spectrum. [3] This latter presentation is essentially only human-readable and very far from ideal for further machine processing. These traditional processes result in considerable loss of data and hence potentially also loss of information.

Such lossy reporting of data is far from ideal when associated with a new molecule that has never previously been reported. A basic principle of claiming the synthesis of a new molecule is that it be properly characterised as novel

using the information derived from its measured data. This allows others to show that its synthesis can be replicated and/or to infer that the molecule indeed has the constitution claimed for it. Traditionally the spectral data was included in the primary article reporting the new molecule in what became known as the “experimental section” and there it was often presented only in highly condensed formats such as listings of chemical shift values and coupling constants. During the last thirty years or so, this experimental section of the primary article has become increasingly devolved into a separate document known as the “supporting information” and this is now invariably made available in the form of an electronic supplement to the primary reporting article. Hence the term ESI or electronic supporting information. [4, 5] The ESI document can often become very large and can approach 1000 "pages" [3] of a page-broken PDF file, arguably not an entirely useful measurement of the amount of data. It is rarely reliably structured in a fully predictable manner such that it can be used to reconstitute the constituent datasets associated with explicit molecules, for re-use and re-analysis using automation and software. The NMR spectrum often appears only as a visual object with limited resolution in the ESI document and its metadata is only captured in the form of a free-text caption. Referencing the molecule to which the data relates is probably via an index number such as e.g. compound 27, the depiction of which itself may only appear in another image. With practice, humans can acquire the skills to navigate such an ESI document, but it presents considerable hurdles for the automated/unsupervised application of AI techniques.

A solution to these issues is to make NMR data available not as text or images, but in a defined and structured data format. Two of the most common primary NMR data formats are those deriving from the commercial Jeol and Bruker instruments. The former comprises a single .jdf (Jeol data format) binary file and the latter is presented as a fileset of text and binary files organised in a folder and which is commonly distributed as a compressed ZIP (.zip) archive. Both formats have relatively opaque internal structures. One commercially available solution which can access the data in these formats is MestreNova, [6] which can absorb the primary data and automatically transform this into a spectrum to produce a composite .mnova file, a process which ensures no overall loss of data.[1] The format can also hold information about the molecule itself, along with a selection of rich metadata derived from the instrumental parameters and weighting functions which allows automated reprocessing of the primary data. An even closer correspondence between the data and the molecule can be achieved using formats such as NMReDATA annotations, [7, 8] which captures extracted features like peak lists, coupling constants and other inferred correlations between the spectrum and the molecule, as produced either by human assignment or AI techniques. NMReDATA combined with the raw data are referred to as a full NMR Record. An example with full NMReDATA and NMR records has been published. [9] Other less frequently used formats deal either mostly with raw data (nmrML) [10] or extracted spectral data (JCAMP-DX, [11] CMLSpect [12]).

Two approaches can be taken to the long term archival of such loss-free and curated filesets. The first is submission of some or all of such data to dedicated spectral databases which have been designed for the purpose. Examples of such databases are SDBS, [13] BMRB, [14] or nmrshiftdb2. [15] BMRB, which is part of PDB, [16] contains mainly data from macro molecules and focuses on shifts and coupling constants, but not on raw data. Similarly, SDBS does not allow raw data and does not have an open submission interface. nmrshiftdb2 on the other hand, accepts raw data and has a submission interface, but it requires picked peaks in any case. So the type of data and the access mechanisms to it differ according to the databases. Ideally, the availability should include both raw data and extracted data (peaks), a full metadata record and a complete set of spectra for a compound. nmrshiftdb2 has this for some datasets which have full data for a discrete set of publications, [17] but generally such data are rare in NMR spectroscopy. The situation is much better in X-ray crystallography, where data deposition in the CSD [18] is often mandated by journals and a comprehensive and well-curated data repository has been built over a period of more than fifty years. Access to this data is however on a commercial basis. The COD database [19] has no such access restrictions, but it is also only half the size of the CSD. Such a centralized approach has not worked in NMR spectroscopy so far. The second,

more recent modern trend, is to exploit the advantages of data repositories and this latter approach will be addressed in more detail below.

The different types of information can potentially all be used for work in AI. A good deal of this work relies on extracted data, particularly shift positions but also coupling constants. The most prominent application here is for chemical shift prediction, where deep learning-based approaches have recently been shown to give very accurate results [20, 21]. An early paper [22] emphasizes that the method described does not need peak picking, but uses pattern recognition on the spectrum directly. Interestingly, that paper had 188 citations (on 12/2/2021 according to google scholar), of which 93 were from 2000 or earlier, indicating that the approach is still of interest today. Recent work using spectral images [23] performs fragment identification using convolutional neural networks. Other approaches have used signal processing techniques on the original data or the processed data, for example for spectrum de-noising. [24] These are just some examples; a recent review is available. [25] It is clear that NMR data of all levels should be available in a systematic and predictable fashion for exploitation by AI techniques. In the present article, we describe an alternative approach to making data available by publication and which requires an infrastructure including a general metadata registration agency [26] which is complementary to that for used journal publishing [27] and access to a data publication repository which can capture a metadata record rich enough to carry at least information relevant to NMR data. [28]

3 | CASE STUDY

3.1 | Experiment Overview

When the opportunity arose to design a new experiment for an undergraduate chemical synthesis laboratory, we decided to incorporate the modern data-handling procedures which retain primary and loss-free NMR data as an integral part of the experiment, introducing at least partially solutions to the problems mentioned in Section 2. Even more importantly, this experiment would serve the purpose of introducing students to concepts associated with data management above, nicely summarised by the acronym FAIR, [2] where the letters stand for F=Findable, A=Accessible, I=Interoperable and R=Re-usable. The experiment selected was a combinatorial synthesis of an ester, starting from a relatively small library of component organic acids and alcohols (Figure 1).

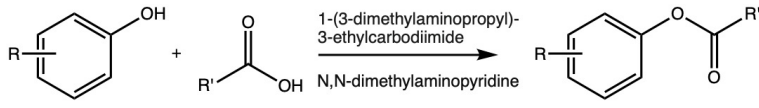


FIGURE 1 The scheme for synthesising an organic ester from an aromatic phenol and aliphatic carboxylic acid

These are selected so that each student is allocated a novel ester which has never previously been reported in the scientific literature. Each year that the course runs, a new combination of the two components would be assigned to each student and over a period of time a library of new esters would be produced, together with spectroscopic data and metadata unique to each molecule. The novelty of a new molecule would not of itself produce a viable publishable unit in terms of the traditional scientific journal model and hitherto the only option for sharing such data with the larger community would be submit it to one of the dedicated spectral databases noted earlier, at which point its provenance as a citable work by the researcher is likely to be lost.

Here we chose to explore a new approach in this experiment, in which the information about each molecule synthesized as part of the course is instead formally published to a data repository in the form of a FAIR fileset. This

is accompanied by a minimal set of structured descriptive information (the metadata) which characterises not only the NMR data and the molecule but also the researcher who synthesized it [29] and the institution they worked at. [30] In a manner similar to using a dedicated NMR database, the metadata is registered with a data registration authority, [26] where it is added to a global aggregated metadata store and there indexed to make it searchable. The researcher in turn receives a citable "receipt" for their deposition in the form of a persistent identifier for the file or dataset and its metadata. The resulting richness of any search of this metadata is then defined by the extent and scope of the submitted metadata. NMR data published in a data repository differs from that of a traditional NMR database in that for the latter the internally defined indexing terms are derived by suitable curation of the specific database, whereas repository-based metadata is obtained from the depositor for global aggregation and indexing. Such registered metadata also differs in that it can be harvested and re-purposed by other agencies, who can add further metadata based on their own more specialised and potentially AI-inferred/informed curation.

3.2 | Student Researcher Learning Objectives

The experiment is designed for new students who may have had limited experience of chemistry laboratories prior to the course. During a two-week module, students are first introduced to experimental techniques associated with molecule synthesis, including safely handling glassware, techniques for following the progress of the reaction to identify reaction completion and simple separation techniques for product isolation and purification. The final sample is then prepared for NMR analysis by dissolution in a suitable solvent and placed in a suitable container for insertion into the NMR instrument. Once the data is recorded by the instrument, the student recovers the raw or primary NMR data files from the spectrometer data server and subjects this data to Fourier Transform processing using a suitable program, in our case using MestreNova. [6] Students then are invited to meet in small supervised groups to interpret the resulting visual spectra in terms of their individual chemical structures, which helps them ascertain if the desired molecule was actually synthesized. Nonetheless, these are inexperienced students who are still learning how to interpret their recorded spectral data and so it is important that a mechanism is in place for the data to be subsequently subjected to some suitable authentication procedure for final quality control of aspects such as specimen purity and reconciliation of the data with the postulated chemical structure.

To enable suitable infrastructures to be in place that can accomplish this last phase, the students are introduced to the basic procedures of data publication in a repository. This includes the student acquiring a persistent identifier unique to them in the form of a registered ORCID (Open Researcher and Collaborator ID) to enable correct attribution of the original creator of the molecule. A workflow is also in place to transform a student-provided atom-connection table of the molecule into algorithmically-generated metadata such as an InChI chemical identifier [31] to help characterise the molecular constitution. The process is completed with the student publishing the filesets resulting from their experiment and recording the resulting registered digital object identifier (DOI), which is now also added to their ORCID publication record along with a time and date stamp for the work. In doing this, the students have learnt about two basic forms of persistent identifier, those that are registered with an agency but carry no other semantic information and those that are semantically derived without the need for registration.

For most traditional laboratory experiments, students are taught how to present the experimental data in the form of visual graphs, images and tables, these being considered the traditional part of the primary process of eventually reporting the experiment and its outcomes and interpretations in a primary and peer-reviewed journal article. Whilst less attention is devoted to teaching how to prepare and structure a more extensive collection of visual and interpreted numerical information (the ESI), there is even less precedent for introducing the management of primary or unprocessed instrumental data to students. In doing this experiment, students are introduced to the concepts of how

processing data into visual information (spectral images) incurs loss of data and the benefits of having access to the original loss-free primary data which can be reprocessed or reanalysed according to need. By publishing a FAIR data archive resulting from their laboratory experiment in synthesizing a new molecule, the students not only start their journey as a scientific researcher by gaining probably their first scientific (data) publication, but enable others to benefit from this unique and open data and the opportunity to re-use it in a re-purposed (the I of FAIR as in inter-operated) form in other experiments and analyses.

This approach has similarities to a Course-Based Undergraduate Research Experience (CURE), where students explore an original research question, typically within the context of the research of their faculty. [32] In our case, we focus on teaching certain practices, similar to students traditionally being taught how to keep lab records. A CURE takes this further by providing a full research project. Ideally, our approach can be integrated into a CURE.

3.3 | Data publication details

This process has distinct phases, a submission stage and editing phase to remove any errors and inspection of the outcome of the process.

1. Firstly, the student has to associate their ORCID with the data repository they are going to use for publication, so that this identifier is automatically propagated to the metadata records for each publication. In our laboratory, we use a local repository which was explicitly designed for this purpose, [33] but other options include registering the ORCID using e.g. the Zenodo repository. [34]
2. Next, the student prepares computer files which will be needed for the data publication process associated with their newly synthesized molecule. A minimal set would include the data obtained from the NMR instrument in the form of the FID. On Bruker spectrometers, this is produced as a folder of files and is conveniently first compressed into a ZIP archive for uploading to the publication repository. The student then produces a transformed version of this file using the MestreNova program, which creates an output that includes not only the original FID data but also adds the processed frequency domain spectrum. The MestreNova program can also be used to produce other optional spectrum files in e.g. a JCAMP-DX format, which is the spectral data in a form suitable for visual representation and expressed in a non-proprietary standard format. A simple PDF visual format, the standard form included in traditional ESI information files is also generated, since this might be useful in some circumstances. The final essential file is a connection specification which defines the 2D atom connectivity of the molecule (i.e. which pairs of atoms in the molecule are considered as sharing a bond). Although useful in itself, this file is actually needed to generate a unique molecular identifier for the molecule known as an InChI in the form of both the cannonical InChI and the shorter InChI key derived from a hash code. Formats which can serve this purpose include the classical MDL Molfile, or the proprietary Chemdraw file.
3. The final publication phase comprises completing a short information record presented on the selected data repository. This would include providing basic metadata fields such as the title of the record and an associated description. For the title, we recommend to students that they use a suitable program such as Chemdraw to automatically generate the systematic name for the molecule, using what could be described as an AI-process. This process also helps to validate the structure, since an incorrectly or incompletely drawn chemical structure will not allow a name to be generated. The description field can be the synthetic procedure used to prepare the molecule, including the quantities of chemicals used and experimental methods. In theory this latter information could also be expressed as metadata, but currently no suitable automated procedure for doing this is available.
4. Each of the files previously prepared is selected for upload, together with a short repository description field

to describe its purpose. This latter field can also be used for versioning. Depending on the repository used, there may be other fields that require selection. For our repository, this includes a license selector, which by default selects the Creative Commons Public Domain Dedication 1.0, or CC0 [35] and a funding grant number if appropriate. Another feature of our repository is that the file or dataset can be associated as a member (a child) of a higher-level collection, which is a method of organising data into a hierarchy to help its Findability. Thus our top level hierarchy is a collection labelled "Undergraduate synthesis laboratories at Imperial College", which contains year groups as in "1st Year undergraduate synthesis 1.2 laboratory (2019-2020)", which in turn then lists the data publications by individual students who have taken the course. The student has to associate their publication with the appropriate collection, so that the publications by the entire year group can be easily viewed and inspected by an instructor.

5. The student will now select the publish option. An internal workflow collects the metadata, expresses it as an XML document conforming to the DataCite schema [36] and registers it with DataCite. The response contains the persistent identifier now associated with this metadata record which is recorded at the repository, along with a link which enables this record to be easily retrieved and a timestamp for the process.
6. The final phase is to allow the depositor to edit some aspects of the metadata record to correct any errors made in the first pass to publication (fields such as the datestamp are not editable, nor can any files be deleted, with the only option being to submit a revised new version). If an experiment had been undertaken as a group, then the depositor can also add the fellow members of the group as co-authors, again identified by their ORCID.
7. Finally the student then records the DOI assigned to their deposition in their laboratory notebook, in the form of a formal citation. They can also cite the overall project collection identifier to illustrate the collaborative nature of this process, again in the form of a collection DOI.

4 | RESULTS AND DISCUSSION

4.1 | Data publication

Figure 2 illustrates the outcome of the activities of class of 2019-2020, showing the overall hierarchy and how some of the metadata captured during the process is presented on the landing page for the item. This landing page is primarily for human visual inspection; the complete metadata record would be used by machines. The DOI assigned to the year collection is shown at the top, together with a link to the overall metadata record should you wish to inspect it. This is followed by the list of contributors, with each link to their ORCID record. The main parts of the collection are the datasets produced by the individual contributors, showing the DOIs and the titles. Most of these titles relate to the systematic naming for the molecule, but some are more generic and reveal less about the identity of the molecule. The title field is free-form; it is not currently controlled in any way to ensure e.g. a sensible molecule name. Finally, to the collection is appended any other associated information which must also have an assigned persistent identifier as a DOI. In this example it is a formal publication describing the FAIR-data processes associated with this laboratory technique [1] and ultimately any DOI resulting from the present article.

Shown in Figure 3 is an individual entry in the collection. There are three generated attributes here that require explanation. The first of these is the presence of additional files created by the submission process and not uploaded by the depositor, taking the form of an .mnpub file. [1] These are a feature unique to the repository we use and which contain in effect a single-use license which allows anyone downloading specifically this .mnpub file to access the dataset associated with it using the Mestrenova program in fully functional form. This is available to any external user and overcomes the aspect that MestreNova is a commercial program, normally requiring a pre-purchased license

to unlock functionality such as providing Fourier Transform capability that converts the primary data (the FID) into a spectrum. Provision of such a licence enables both the accessibility and the inter-operability of the FAIR attributes of the data; effectively highly sophisticated software has to be made available to allow the data (the FID) to be accessed and inter-operated into a spectrum. The other two generated attributes of the data are the InChI string and InChI key, derived from the provided connection table (Chemdraw file in this example), and which allows a broader search in principle of other instances of data in other repositories which relate to molecules having the same molecular identifier. This in turn facilitates the Findability of the data (the F in the FAIR acronym), via the information in the generated metadata record for the entry.

4.2 | Exploiting the metadata record

Having created a collection of NMR datasets each populated with a metadata record, we now address the aspect of machine handling of the metadata records by exploiting two attributes of FAIR data; its Findability and its Accessability. The metadata record is accumulated and indexed in the DataCite metadata store according to the DataCite schema [36] and this store can be searched using the ElasticSearch engine.[37] A selection of searches is illustrated here to illustrate the potential only. More complex searches would be enabled by populating the metadata records with richer, more finely grained information.

1. Seeding can be by a general search illustrated by:
   https://commons.datacite.org/?query=titles.title:1st+AND+Year+AND+undergraduate+AND+synthesis+AND+laboratory
   where the prefix `https://commons.datacite.org/?query=` represents the resolver component of the search query (and is omitted in the examples below). The syntax shown is that of ElasticSearch, [37] although a more human-friendly interface to this syntax has recently been made available. [38] This query returns a single hit, for which the persistent identifier is identified as `10.14469/hpc/6215` and for which the metadata record can be retrieved in several ways using a metadata resolver:
   https://data.datacite.org/application/vnd.datacite.datacite+xml/10.14469/hpc/6215
   or in JSON-LD (JavaScript Object Notation for Linked Data) format via a command line as: `curl -LH "Accept: application/vnd.schemaorg.ld+json" https://doi.org/10.14469/hpc/6215`
   This record shows the item to be a collection (`resourceTypeGeneral="Collection"`) and this collection can now be used for more focused searches of data availability associated with this project.
2. The query:
   relatedIdentifiers.relatedIdentifier:10.14469/hpc/6215+AND+relatedIdentifiers.relationType:IsPartOf
   identifies 47 individual filesets that have been defined as a part of the top level collection for which the PID is `10.14469/hpc/6215`
3. The task now is to identify the type of data that might be available for each of these datasets. This can be achieved by specifying the media type of the desired data:
   relatedIdentifiers.relatedIdentifier:10.14469/hpc/6215+AND+relatedIdentifiers.relationType:IsPartOf
   +AND+(media.media_type:chemical/x-mnpub*+AND+media.media_type:application/zip)
   This specifies the two media types that were uploaded by students. The ZIP archive was created directly from the data retrieved from the NMR instrument and is presumed to be unprocessed in any way. The .mnpub file is automatically created by a repository workflow script to associate a license file allowing the Mestrenova program to process the data. The search returns 35 hits, which implies that 12 of the datasets do not have both these files.
4. An alternative declaration of media types would contain both the original NMR instrumental data and a processed

version as an .mnova type:
relatedIdentifiers.relatedIdentifier:10.14469/hpc/6215+AND+relatedIdentifiers.relationType:IsPartOf
+AND+(media.media_type:chemical/x-mnova*+NOT+media.media_type:application/zip)
returns 10 sets that contain only the .mnova version of the instrumental data. Further changes to the Boolean operator reveal that a total of 45 datasets have either the ZIP or the mnova dataset, implying that two contain neither. In fact, one of these contains the alternative Jeol .jdf format for raw NMR data and the other indeed indicates that no NMR data was uploaded by the student.

5. We next seek how many of the datasets also contain a chemical identifier as an InChI. The search:
relatedIdentifiers.relatedIdentifier:10.14469/hpc/6215+AND+relatedIdentifiers.relationType:IsPartOf
+AND+subjects.subjectScheme:inchikey
indicates that 39 of the datasets also have metadata that identifies the associated molecule by specifying an InChI chemical identifier and that 38 of these also have NMR instrumental data in some form.
6. These searches indicate that even using minimal metadata records, it is possible to identify in a machine-sense a corpus of NMR instrumental data and the asserted molecule each dataset derives from. One can focus on just a single entry by specifying the value of the chemical identifier:
relatedIdentifiers.relatedIdentifier:10.14469/hpc/6215+AND+relatedIdentifiers.relationType:IsPartOf
+AND+subjects.subjectScheme:inchikey+AND+subjects.subject:KKOOECZDSKSVIG-UHFFFAOYSA-N
+AND+(media.media_type:chemical/x-mnova*+OR+media.media_type:application/zip)
7. The metadata record for this dataset:
https://data.datacite.org/application/vnd.datacite.datacite+xml/10.14469/hpc/6468
contains one further information record that would be useful for an unsupervised retrieval by e.g. AI post-processing of the data to potentially allow a verification of the asserted association between the dataset and the chemical structure. The metadata record identifies the availability of an ORE (Object Re-use and Exchange) manifest of additional metadata by the declaration:
`<relatedIdentifier relatedIdentifierType="URL" relationType="HasMetadata">`
https://data.hpc.imperial.ac.uk/resolve/?ore=6468</relatedIdentifier
which in turn allows programmatic retrieval of the desired dataset by specifying its media type and including a declaration of the size of the dataset for checksum verification. An example of this extended metadata record is shown below.
`<atom:link rel="...snip..." href="https://data.hpc.imperial.ac.uk/resolve?doi=6468&file=2"`
`title="Cumyl_ethyl_malonate.mnova" type="chemical/x-mnova" length="326389"/>`
A (non-AI) example of how such a metadata record can be utilised can be viewed at DOI: 10.14469/hpc/6273 where a molecular visualisation and analysis program (JSMol) is requesting data for processing, using the persistent identifier of that data and its media type as defined in the `handle_jmol` script.
`javascript:handle_jmol('10.14469/hpc/7668','...scripting commands...;'.`

5 | CONCLUSIONS

We have here described a newly introduced undergraduate experiment in which each student synthesises a novel chemical compound and records NMR instrumental data to help characterise the molecule. Traditional publishing mechanisms would involve reducing this data to a processed spectrum, expressing the spectrum as a PDF-based vector diagram and then including this diagram in a very loosely structured PDF document which is largely devoid of

descriptive metadata. Here we have introduced students to how a metadata-rich environment for the data can be created, one which involves retaining not only the processed spectrum but also the original unprocessed instrumental data, together with instrumental settings and parameters. This metadata is centered around enabling the data as FAIR by including properties that will allow it to be found by searches of the metadata, as registered with an appropriate authority in exchange for a persistent identifier. The metadata also includes other properties such as media types which identify the document container for the data and a file manifest or resource map which allows the fully-qualified file paths to be defined to allow unsupervised access to the data. Other metadata in this particular collection includes some form of chemical identifier which sets up the capability to correlate features in the spectral data with the nuclear environments in the molecule. This in turn would allow patterns in these features to be detected across a collection of, in this example, closely related molecules. It would also allow anomalies to be detected, for example resulting from either the wrong molecule being formed during the synthesis or the presence of unreacted molecules because the reaction has failed to complete. This capability is important for assessing the quality of the original data, especially so since for this project it has been created by relatively inexperienced researchers. In this regard the FAIR attributes are especially well suited for AI-applications, since unsupervised procedures can be programmed to retrieve and identify the type of data being analysed. Such an AI-oriented perspective is not normally the focus of FAIR data, which often remains oriented towards humans rather than machines.

The metadata described in this data collection should be considered both minimal and very much evolving in richness. One component currently entirely missing is that relating to chemical shift and coupling constant assignments, although recommendations for suitable metadata descriptors are currently being prepared. [39, 40] We also note that most metadata records registered with e.g. DataCite are far more impoverished (sub-minimal) than the ones were are describing here, a failing that we are attempting to correct by educating the students in the Imperial course in the desirability of FAIR-enabling metadata. The records here also adopt the particular approach of exploiting the elements of a particular schema for the data descriptors as defined by DataCite, coupled with the use of the much older media type document property. This is just one of a number of approaches that could be taken and it is to be expected that standards bodies such as e.g. the International Union of Pure and Applied Chemistry (IUPAC) will eventually recommend a more standardised approach to the metadata structures. Currently, the approach taken in this project parallels that in setting up a bespoke database, with unique internal structures and API. The expectation is that eventually these parochial differences will be replaced by a small number of standard metadata schemas for describing data and in particular recommendations for implementations of such schemas, so that an AI-based system can retrieve data from a global environment without any explicit inclusions of local rules. This would allow scaling up to in effect the global output of data acquired as part of the scientific processes in e.g. chemistry.

Conflicts of interest

There are no conflicts of interest in this work.

Supporting Information

All NMR data referred to in this article [41] can be found via the Imperial College data repository at DOI: 10.14469/hpc/6215

Acknowledgements

We would like to acknowledge Drs Ed Smith and Laura Patel, who designed the student course described here and encouraged the addition of a FAIR data component to it.

references

[1] Barba A, Dominguez S, Cobas C, Martinsen DP, Romain C, Rzepa HS, et al. Workflows Allowing Creation of Journal Article Supporting Information and Findable, Accessible, Interoperable, and Reusable (FAIR)-Enabled Publication of Spectroscopic Data. ACS Omega 2019 Feb;4(2):3280–3286. https://doi.org/10.1021/acsomega.8b03005.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016 Mar;3(1):160018. https://doi.org/10.1038/sdata.2016.18.

[3] Lopchuk JM, Fjelbye K, Kawamata Y, Malins LR, Pan CM, Gianatassio R, et al. Strain-Release Heteroatom Functionalization: Development, Scope, and Stereospecificity. Journal of the American Chemical Society 2017 Mar;139(8):3209–3226. https://doi.org/10.1021/jacs.6b13229.

[4] Martinsen D. Primary Research Data and Scholarly Communication. Chemistry International 2017;39(3):35–38. https://doi.org/10.1515/ci-2017-0309.

[5] McEwen L. Research Data Reporting in Chemistry. In: ACS Guide to Scholarly Communication American Chemical Society; 2019. https://doi.org/10.1021/acsguide.30104.

[6] MestreNova; Accessed: 2021-1-20. https://mestrelab.com.

[7] Pupier M, Nuzillard JM, Wist J, Schlörer NE, Kuhn S, Erdelyi M, et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. Magnetic Resonance in Chemistry 2018;56(8):703–715. https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4737.

[8] Kuhn S, Wieske LHE, Trevorrow P, Schober D, Schlörer NE, Nuzillard JM, et al. NMReDATA: Tools and applications. Magnetic Resonance in Chemistry;n/a(n/a). https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mrc.5146.

[9] Hahn M, von Elert E, Bigler L, Díaz Hernández MD, Schloerer NE. 5α-Cyprinol sulfate: Complete NMR assignment and revision of earlier published data, including the submission of a computer-readable assignment in NMReDATA format. Magnetic Resonance in Chemistry 2018;56(12):1201–1207. https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4782.

[10] Schober D, Jacob D, Wilson M, Cruz JA, Marcu A, Grant JR, et al. nmrML: A Community Supported Open Data Standard for the Description, Storage, and Exchange of NMR Data. Analytical Chemistry 2018 Jan;90(1):649–656. https://doi.org/10.1021/acs.analchem.7b02795.

[11] Davies AN, Lampen P. JCAMP-DX for NMR. Applied Spectroscopy 1993 Aug;47(8):1093–1099. https://doi.org/10.1366/0003702934067874.

[12] Kuhn S, Helmus T, Lancashire RJ, Murray-Rust P, Rzepa HS, Steinbeck C, et al. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. Journal of Chemical Information and Modeling 2007 Nov;47(6):2015–2034. https://doi.org/10.1021/ci600531a.

[13] of Advanced Industrial Science NI, Technology, National Institute of Advanced Industrial Science and Technology: SDBSWeb; Accessed: 2021-1-20. https://sdbs.db.aist.go.jp.

[14] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res 2008 Jan;36(Database issue):D402–408.

[15] Kuhn S, Schlörer NE. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories. Magnetic Resonance in Chemistry 2015;53(8):582–589. https://onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4263.

[16] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nature Structural & Molecular Biology 2003 Dec;10(12):980.

[17] Example records at nmrshiftdb2; Accessed: 2021-1-20. http://www.nmrshiftdb.org/collections/ChiuZ%7CClassics%2Bin%2BSpectroscopy.

[18] Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. Acta Crystallographica Section B 2016;72:171–179. https://doi.org/10.1107/S2052520616003954.

[19] Gražulis S, Chateigner D, Downs RT, Yokochi AFT, Quirós M, Lutterotti L, et al. Crystallography Open Database – an open-access collection of crystal structures. Journal of Applied Crystallography 2009 Aug;42(4):726–729. https://doi.org/10.1107/S0021889809016690.

[20] Unzueta PA, Greenwell CS, Beran GJO. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. J Chem Theory Comput 2021 Feb;17(2):826–840.

[21] Jonas E, Kuhn S. Rapid prediction of NMR spectral properties with quantified uncertainty. Journal of Cheminformatics 2019;11:50.

[22] Kowalski BR, Bender CF. K-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Analytical Chemistry 1972;44:1405–1411. https://doi.org/10.1021/ac60316a008.

[23] Kuhn S, Tumer E, Colreavy-Donnelly S, Moreira Borges R. A Pilot Study For Fragment Identification Using 2D NMR and Deep Learning. arXiv e-prints 2021 Mar;p. arXiv:2103.12169.

[24] Lee HH, Kim H. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. Magn Reson Med 2019 07;82(1):33–48.

[25] Cobas C. NMR signal processing, prediction, and structure verification with machine learning techniques. Magnetic Resonance in Chemistry 2020;58(6):512–519. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mrc.4989.

[26] Neumann J, Brase J. DataCite and DOI names for research data. Journal of Computer-Aided Molecular Design 2014 Oct;28(10):1035–1041. https://doi.org/10.1007/s10822-014-9776-5.

[27] The Formation of CrossRef: A Short History; 2009. https://www.crossref.org/pdfs/CrossRef10Years.pdf.

[28] Harvey MJ, McLean A, Rzepa HS. A metadata-driven approach to data repository design. Journal of Cheminformatics 2017 Jan;9(1):4. https://doi.org/10.1186/s13321-017-0190-6.

[29] ORCID; Accessed: 2021-1-20. https://orcid.org/.

[30] ROR (Research Organization Registry); Accessed: 2021-1-20. https://ror.org/about.

[31] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. Journal of Cheminformatics 2015 May;7(1):23. https://doi.org/10.1186/s13321-015-0068-4.

[32] Sun E, Graves ML, Oliver DC. Propelling a Course-Based Undergraduate Research Experience Using an Open-Access Online Undergraduate Research Journal. Frontiers in Microbiology 2020;11:2980. https://www.frontiersin.org/article/10.3389/fmicb.2020.589025.

[33] Imperial College Research Data Repository, entry in FAIRsharing.org;. https://doi.org/110.25504/FAIRsharing.LEtKjT.

[34] Zenodo Data Repository; Accessed: 2021-1-20. https://about.zenodo.org.

[35] Hrynaszkiewicz I, Cockerill MJ. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. BMC Research Notes 2012 Sep;5(1):494. https://doi.org/10.1186/1756-0500-5-494.

[36] working group DM, DataCite Metadata Schema 4.3; 2019. https://doi.org/10.14454/f2wp-s162.

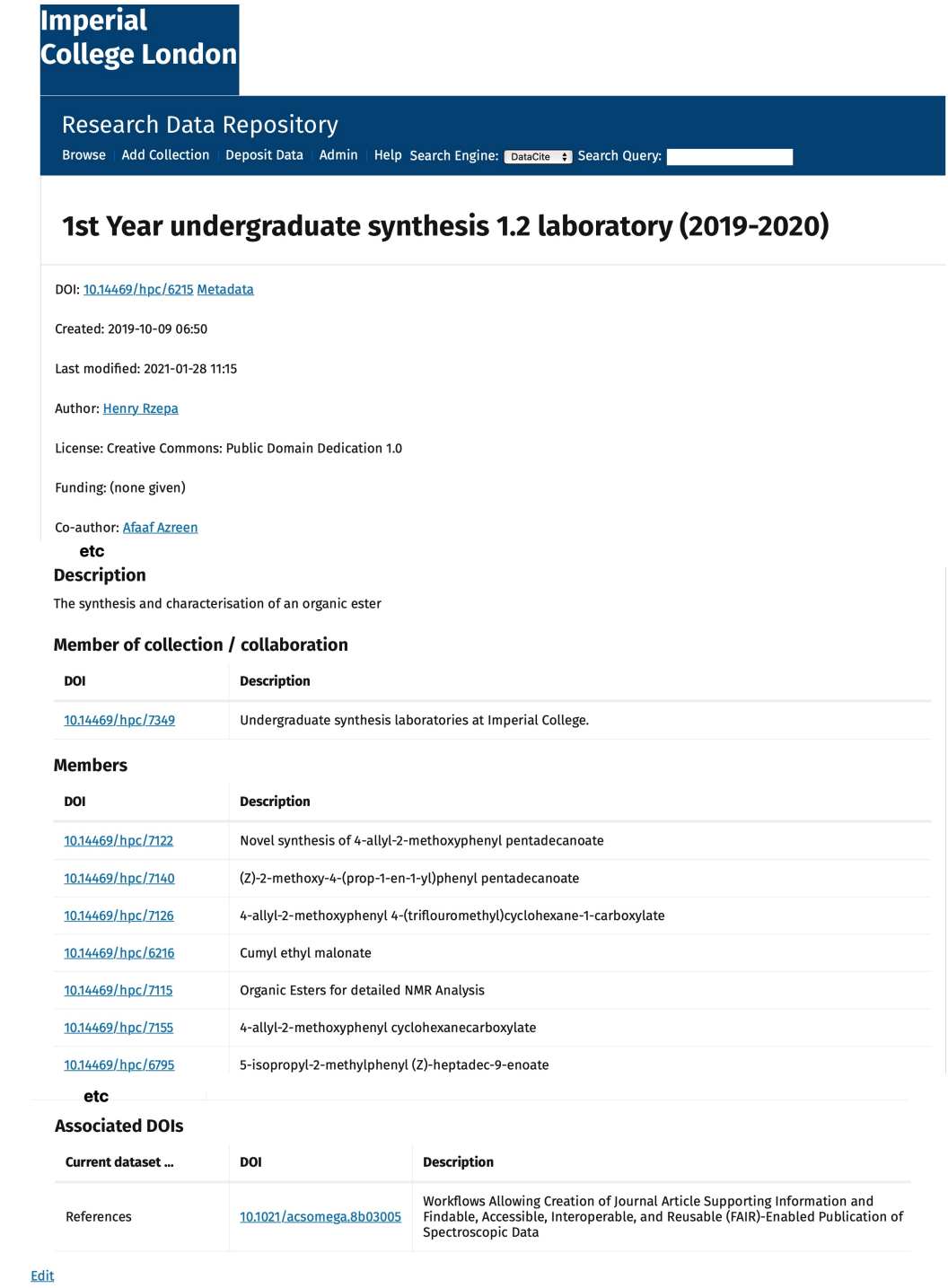
[37] Banon S, ElasticSearch; Accessed: 2021-1-20. https://www.elastic.co/elasticsearch/.

[38] Romain C, Rzepa HS, Query builder for metadata search: NMR data; Accessed: 2020-4-14. https://doi.org/10.14469/hpc/7992.

[39] Hanson R, Jeannerat D, Rzepa HS, Archibald M, Bruno I, Chalk S, et al. Framing FAIR: Scientific Research Data Sharing Policies, Frameworks and Principles. Talk presented at the American Chemical Society spring meeting 2021;.

[40] Davies AN, Hanson RM, Jeannerat D, Bruno I, Chalk S, Lang J, et al. FAIR enough? Spectroscopy Europe 2021;33(2):25–31. https://doi.org/10.1255/sew.2021.a9.

[41] Rzepa HS, et al, 1st Year undergraduate synthesis 1.2 laboratory (2019-2020). Imperial College London data repository; 2020. https://doi.org/10.14469/hpc/6215.


Imperial College London

Research Data Repository

Browse | Add Collection | Deposit Data | Admin | Help Search Engine: DataCite Search Query:

1st Year undergraduate synthesis 1.2 laboratory (2019-2020)

DOI: 10.14469/hpc/6215 Metadata

Created: 2019-10-09 06:50

Last modified: 2021-01-28 11:15

Author: Henry Rzepa

License: Creative Commons: Public Domain Dedication 1.0

Funding: (none given)

Co-author: Afaaf Azreen

etc

Description

The synthesis and characterisation of an organic ester

Member of collection / collaboration

DOI	Description
10.14469/hpc/7349	Undergraduate synthesis laboratories at Imperial College.

Members

DOI	Description
10.14469/hpc/7122	Novel synthesis of 4-allyl-2-methoxyphenyl pentadecanoate
10.14469/hpc/7140	(Z)-2-methoxy-4-(prop-1-en-1-yl)phenyl pentadecanoate
10.14469/hpc/7126	4-allyl-2-methoxyphenyl 4-(triflouromethyl)cyclohexane-1-carboxylate
10.14469/hpc/6216	Cumyl ethyl malonate
10.14469/hpc/7115	Organic Esters for detailed NMR Analysis
10.14469/hpc/7155	4-allyl-2-methoxyphenyl cyclohexanecarboxylate
10.14469/hpc/6795	5-isopropyl-2-methylphenyl (Z)-heptadec-9-enoate

etc

Associated DOIs

Current dataset ...	DOI	Description
References	10.1021/acsomega.8b03005	Workflows Allowing Creation of Journal Article Supporting Information and Findable, Accessible, Interoperable, and Reusable (FAIR)-Enabled Publication of Spectroscopic Data

Edit

FIGURE 2 The landing page for repository collection of NMR data, showing the members

Research Data Repository

Browse | Add Collection | Deposit Data | Admin | Help Search Engine: Google Search Query:

2-isopropyl-5-methylphenyl (E)-octadec-9-enoate

Admin: rzepa

DOI: 10.14469/hpc/6873 Metadata

Created: 2020-02-28 10:53

Last modified: 2020-02-28 15:14

Author: Haotian Fu

License: Creative Commons: Public Domain Dedication 1.0

Funding: (none given)

Description

2-isopropyl-5-methylphenyl (E)-octadec-9-enoate was produced by the reaction of thymol and elaidic acid with heating under reflux, and EDC.HCl acts as a condensing agent, while DMAP acts as a catalyst.

Files

Filename	Size	Type	Description
2-isopropyl-5-methylphenyl (E)-octadec-9-enoate.cdxml	6KB	chemical/x-cdxml	chemdraw of the molecule
Synthesis 1.2 NMR1.mnova	680KB	chemical/x-mnova	NMR spectra of the molecule
Synthesis 1.2 NMR1.mnpub	0	chemical/x-mnpub	Mestrenova signature file for Synthesis 1.2 NMR1.mnova
Synthesis 1.2 NMR1.zip	290KB	application/zip	Zip NMR spectra of the molecule
Synthesis 1.2 NMR1.mnpub	0	chemical/x-mnpub	Mestrenova signature file for Synthesis 1.2 NMR1.zip

Member of collection / collaboration

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

Subject Keywords

Keyword	Value
inchi	InChI=1S/C28H46O2/c1-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-28(29)30-27-23-25(4)21-22-26(27)24(2)3/h12-13,21-24H,5-11,14-20H2,1-4H3/b13-12+
inchikey	LTKBJOQAMVFHEE-OUKQBFOZSA-N

Edit

FIGURE 3 The landing page for a collection item, showing files and selected metadata present