Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Digital Data Repositories in Chemistry and Their Integration with Journals and Electronic Notebooks

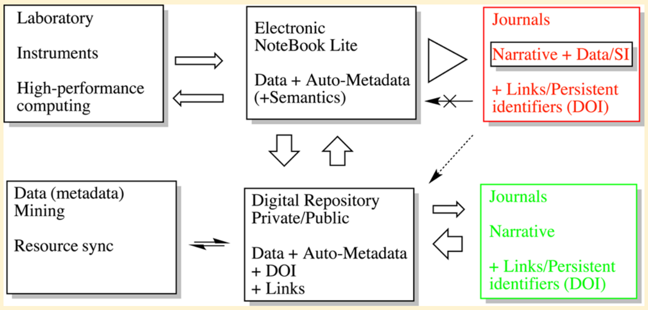
Matthew J. Harvey,[†] Nicholas J. Mason,[‡] and Henry S. Rzepa*,[‡]

[†]High Performance Computing Unit, ICT Division, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

[‡]Department of Chemistry, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

Web-Enhanced Feature

ABSTRACT: We discuss the concept of recasting the data-rich scientific journal article into two components, a narrative and separate data components, each of which is assigned a persistent digital object identifier. Doing so allows each of these components to exist in an environment optimized for purpose. We make use of a poorly-known feature of the handle system for assigning persistent identifiers that allows an individual data file from a larger file set to be retrieved according to its file name or its MIME type. The data objects allow facile visualization and retrieval for reuse of the data and facilitates other operations such as data mining. Examples from five recently published articles illustrate these concepts.


Laboratory
Instruments
High-performance computing
Electronic NoteBook Lite
Data + Auto-Metadata (+Semantics)
Journals
Narrative + Data/SI
+ Links/Persistent identifiers (DOI)
Data (metadata) Mining
Resource sync
Digital Repository Private/Public
Data + Auto-Metadata
+ DOI
+ Links
Journals
Narrative
+ Links/Persistent identifiers (DOI)

INTRODUCTION

Reporting scientific investigations in the form of a periodic journal is a concept dating back some 350 years to the 17th century.[1] For much of that time, the only mechanism for dissemination involved bound paper (the "volume" or "issue"). This of course has changed in the last 20 years, at least in terms of delivery, but the basic structure and format of the scientific article has undergone less change. The article continues to interleave a narrative supported by reporting experimental data. The data is presented in the form of figures, tables, schemes, and plain text in the highly visual form that humans can easily absorb. Often, only a small subset of the data actually available can be presented for reasons of space. The online era, dating back perhaps 30 years, has allowed Supporting Information (data) to be separated from the physical constraints of the printed journal article and deposited with the publisher or a national library as a separate archive.

The logical connection between data present in the main article and its supporting counterpart is in fact tenuous; both sets of data continue to depend on the human reader to extract value from them. Text and data content mining (TDM) however is making enormous strides[2] in allowing machines to harvest data, being capable of far higher and less error prone throughputs than humans. This then facilitates human verification of assertions made in the narrative component of an article or indeed the discovery of connections and patterns between a corpus of articles. The paperbound article, including electronic emulations of paper such as the PDF format, is not well set up for a clear separation of the narrative and the data on which the former is so dependent. The two types of content are also caught up in complex issues of copyright; are the narrative rights and ownership held by the publisher or retained by the author? Is the ability to perform TDM an unrestricted one or prescribed by the publisher? Unfortunately, the data component, which we may presume is not covered by copyright (one cannot copyright the boiling point of water) is often entrained in these complexities. Here, we suggest a new model of how the scientific journal can take advantage of some of the many technical advances in publishing by emancipating the data from its interleaved coexistence with narratives.[3,4] We reflect on how this might allow the journal to evolve in a manner more appropriate for a fully online environment and present several examples.

THE ELECTRONIC NOTEBOOK

A slowly growing innovation is the electronic notebook as the primary holding stage for the capture of data from, for example, instruments, databases, computational resources, and other data-rich sources.[5,6] This model can be represented as in Figure 1. The flow of data is primarily from data source to notebook, and much less information is likely to flow in the other

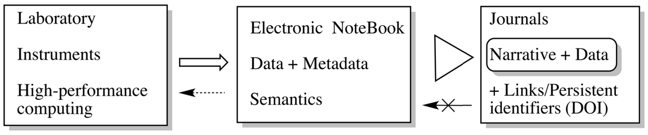


Figure 1. Standard model for the flow of data from the laboratory to the scientific journal.

Received: May 20, 2014
Published: August 29, 2014

direction. When a project is complete, the data held in the notebook is then assembled into a narrative + data visuals in a word processor and submitted in this latter form to a journal. The data, and its expression and semantics, are rarely well preserved in this latter process. There is no communication at all in the reverse direction from the journal article to the notebook; if nothing else, the notebook security model would not permit this.

Consider however an alternative model (Figure 2). It now incorporates perhaps the single most important game-changing technology introduced a little less than 10 years ago, the digital repository in chemistry[7] and other subject domains.[8−10] This combines the concept of using rich (reliable) metadata to describe a dataset with an infrastructure that allows automated retrievals of the datasets, potentially on a vast scale. The other basic component of a digital repository is the idea of a persistent identifier for the data, one that can be abstracted away from any explicit hardware installation. There are other differences from the first model:

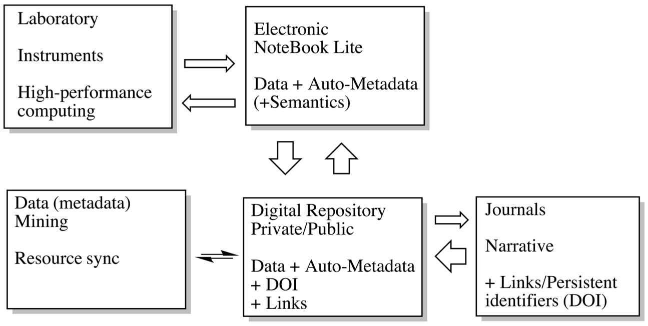


Figure 2. A proposed model for the bidirectional data flows between the laboratory and the scientific journal.

(1) The electronic notebook is replaced by what might be termed the "lite" model, more customized for the type of data being collected and more bidirectional in its data flows.

(2) The lite notebook continues with a privacy model, but the interface between private and shared or public data is now handled instead by the next link in the chain, the repository. Again the data flow model is bidirectional.

(3) It is the repository that can have further (bidirectional) data flows associated. The first, a data-mining model, is not described further in this article. The second becomes the journal.

(4) The links between the repository and the journal are based on persistent identifiers. Generically, they are called handles, of which a specific example much adopted throughout the publishing industry is the Digital Object Identifier, or DOI.[11,12] Again the data flow can be bidirectional; a journal article can reference an entry in a repository and *vice versa*.

(5) In this model, one can achieve a desirable separation of narrative from data. Each of these components can be held in an environment optimized for its primary purpose; neither need be compromised.

In the remainder of the current narrative, we will describe how a working implementation of this model was constructed.

■ UPORTAL: A LITE ELECTRONIC NOTEBOOK

Our starting point was computational chemistry, although solutions for molecular synthesis and spectroscopy have also been trialled[7] and will be described elsewhere. The basic resource used for this is known as a high-performance computing center. Because the latter operates in a fully digital manner, it provides a good test-bed to construct an electronic notebook customized for the purpose. It is referred to here as **uportal** (Figure 3). The design of such a system has to factor in requirements specific to computations:

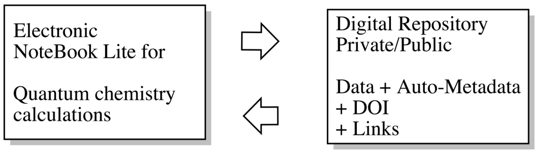

Electronic NoteBook Lite for
Quantum chemistry calculations
Digital Repository Private/Public
Data + Auto-Metadata
+ DOI
+ Links

Figure 3. Data flows between the uportal notebook and a digital repository.

(1) It requires domain knowledge. The need for specific software; the type of calculation that the software is asked to undertake; the resources needed to complete it.

(2) It requires a technical knowledge. What type of computer and software is needed? How much CPU/memory is required? What is the interface to the batch system?

(3) An imposed structured workflow captures all input and output data to provide a full context record that ensures future reproducibility for the calculation. The workflow includes actions such as listing all previous jobs submitted to the system, an interface for creating a new job within a declared project area, and configuring software for computing properties and generating new data.

(4) Naturally arising from the workflow are metadata such as user name, date stamps, and free text descriptions. We have come to believe that such metadata is best generated as part of that automatic workflow; it should not be added in an *ad hoc* and relatively uncontrolled manner at a later stage.

(5) Silently created as part of the computation are CML (XML) representations[13] of the objects being modeled (molecular structures in our case).

(6) Other chemistry-specific autogenerated (meta)data includes portable application-specific data formats (e.g., Gaussian formatted checkpoint file) and tags abstracted from input/output files (InChI, SMILES, calculation type, etc.).

The uportal notebook interfaces via a developer API[14] to a digital repository to enable deposition there of the following:

(1) Raw input and output data files for the computation (ORCA input, Gaussian input, formatted checkpoints).

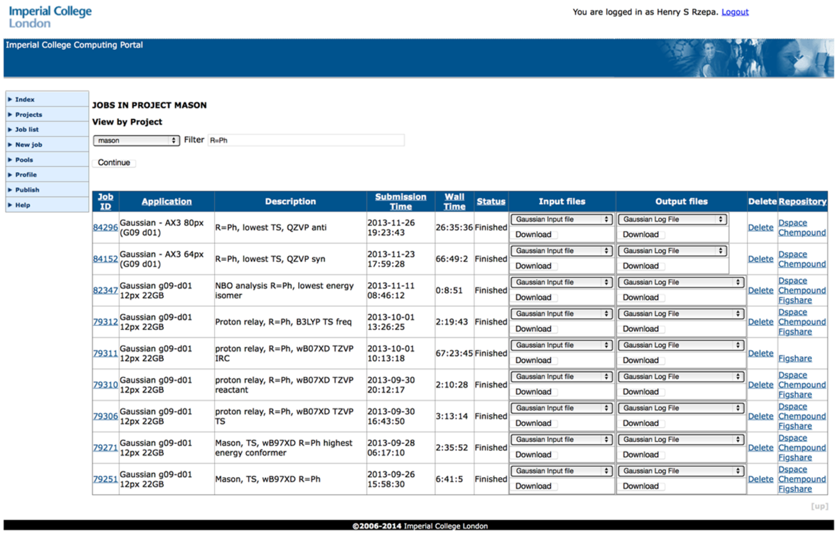
(2) Derived and extracted metadata (program used, calculation types, XML representations, InChI, SMILES identifiers).

(3) Validated metadata from an appropriate validation procedure such as an institutional login or ORCID credentials[15,16] and all date stamps.

(4) Captured metadata (free-text description supplied by depositor, and a project code assigned to each item).

(5) One or more unique and persistent identifiers for each deposition in the form of a DOI/Handle, as assigned by a Handle Server.

(6) Each deposition derives from a specific computation on a specified molecule, and these entries are further grouped by project codes. This latter attribute can be inherited by a digital repository (for example, Figshare) where depositions can be viewed by their project identifier and where project collections can be shared (privately or publically) among a specified group of collaborators.



You are logged in as Henry S Rzepa. Logout

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 4. Uportal: A job submission and notebook system interfacing to a high-performance computing resource and digital repositories.

It is challenging and expensive to build/acquire/configure a general purpose electronic laboratory notebook system that can accomplish all these specialized tasks for a local environment. We believe it is more straightforward to instead construct a lightweight portal using standard scripting environments such as python or php. An overview of such a system is shown in Figure 4, where some of these attributes are listed for each entry. As observed from the sequential ID, the system can easily scale to ~100,000 entries accumulated over a period of around seven years, corresponding to about 14,000 entries each year. This was achieved by around 600 users distributed among staff, postgraduates, and undergraduates. The last column shows the interface to the next component, the digital repository. The bidirectional nature is reflected in the capture of the assigned repository DOI back into the lite notebook if it has been published. Other actions include deleting the entry or simply leaving it unpublished.

■ THE DIGITAL REPOSITORY

We have described elsewhere[7] the principles behind our DSpace-based repository (SPECTRa), introduced in 2006. Two others have been added since then, Chempound[17] and Figshare.[18] There is no limitation to the number of repositories that can be associated with any given electronic notebook.

(1) DSpace contains a handle server that is used to assign unique persistent identifiers and to resolve incoming requests. Our system currently assigns two identifiers, the first being a generic Handle and the second issued by DataCite[19] as a DOI (digital object identifier).

(2) Chempound was added as an example of a semantically enabled system, including the ability to create semantic queries. This system however does not assign its own persistent identifiers via a dedicated handle server.

(3) Figshare is an open repository with a published application programming interface (API). Because it is an external organization, it contains a two-stage authentication system that establishes trust between the uportal and itself via encrypted tokens. Figshare also relies exclusively on DataCite[19] to issue persistent identifiers (DOIs).

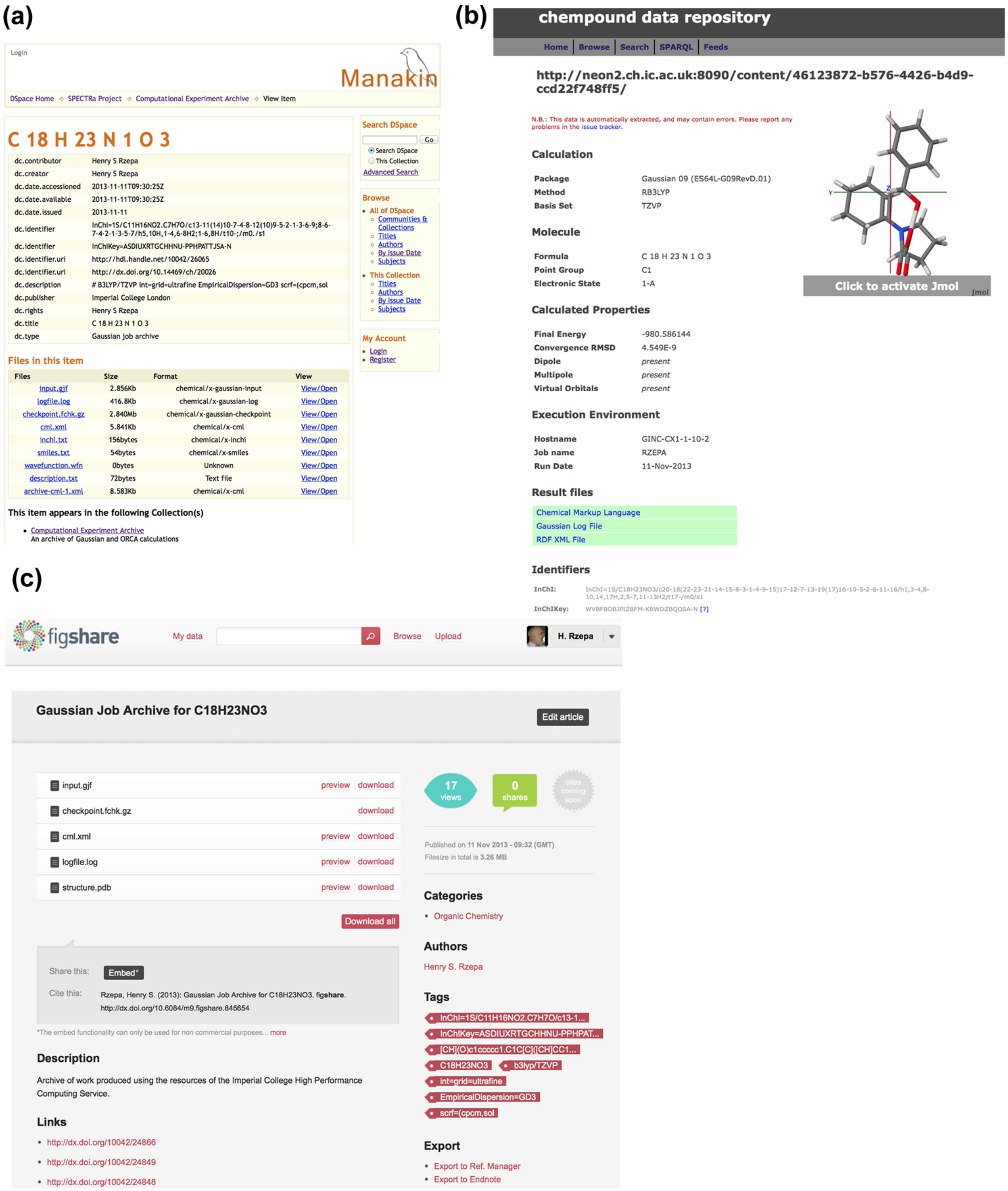
In general, therefore, any dataset collected at the uportal as a result of a job submitted to the high-performance computing resource can be simultaneously published into any combination of these three repositories. An example of how one particular deposited data or file set is presented in these three repositories is shown in Figure 5, highlighting the autodetermined metadata and other attributes. The metaphor is that each dataset relates to a specific molecule with specific metadata for that molecule, and that this collection of data is then assigned its own repository identifier. Sets of such depositions can then be grouped into collections in the form of, for example, datuments (see below).

The Figshare repository differs in one regard from the other two. The initial deposition process reserves a persistent identifier for the object, inherits any project associated with the original entry from the uportal, and creates a private entry within that project. At this stage, Figshare allows collaborators to be assigned exclusive permission to access the items in any given project, but the item is otherwise not open. Only when the project is deemed complete and submitted for publication is each entry converted to public mode. One aspect of this process is not yet supported: a private but nevertheless anonymous mode to enable referees only to view the depositions as part of any review process. Currently, we make the data fully public even at the review stage, with priority afforded by date stamp and other metadata associated with the deposition. Of the three repositories noted above, only one (Chempound) is enabled with semantic Web technology.[17] Link-outs from, for example, Figshare, to the corresponding entries in SPECTRa and Chempound are automatically added to enable semantic searches if they are needed, and a link to a data descriptor is provided for more information about the data structures, syntax, and context.[20]

■ THE HANDLE SYSTEM

Handles are analogous to Web URIs (uniform resource identifiers) in being a hierarchic descriptor containing an authority and a path to a resource. A technology for assigning and resolving persistent identifiers for digital objects (IETF RFCs 3650-2) was developed and is maintained by the Corporation for National Research Initiatives (CNRI). This has the following features:

(1) The identifiers have the form XXXX/YYYY, for example, 10042/28000 or 10.6084/m9.figshare.1234.



(a)

Login

Manakin

DSpace Home › SPECTRa Project › Computational Experiment Archive › View Item

C 18 H 23 N 1 O 3

dc.contributor	Henry S Rzepa
dc.creator	Henry S Rzepa
dc.date.accessioned	2013-11-11T09:30:25Z
dc.date.available	2013-11-11T09:30:25Z
dc.date.issued	2013-11-11
dc.identifier	InChI=1S/C11H16NO2.C7H7O/c13-11(14)10-7-4-8-12(10)9-5-2-1-3-6-9;8-6-7-4-2-1-3-5-7/h5,10H,1-4,6-8H2;1-6,8H/t10-;/m0./s1
dc.identifier	InChIKey=ASDIUXRTGCHHNU-PPHPATTJSA-N
dc.identifier.uri	http://hdl.handle.net/10042/26065
dc.identifier.uri	http://dx.doi.org/10.14469/ch/20026
dc.description	# B3LYP/TZVP int=grid=ultrafine EmpiricalDispersion=GD3 scrf=(cpcm,sol
dc.publisher	Imperial College London
dc.rights	Henry S Rzepa
dc.title	C 18 H 23 N 1 O 3
dc.type	Gaussian job archive

Files in this item

Files	Size	Format	View
input.gjf	2.856Kb	chemical/x-gaussian-input	View/Open
logfile.log	416.8Kb	chemical/x-gaussian-log	View/Open
checkpoint.fchk.gz	2.840Mb	chemical/x-gaussian-checkpoint	View/Open
cml.xml	5.841Kb	chemical/x-cml	View/Open
inchi.txt	156bytes	chemical/x-inchi	View/Open
smiles.txt	54bytes	chemical/x-smiles	View/Open
wavefunction.wfn	0bytes	Unknown	View/Open
description.txt	72bytes	Text file	View/Open
archive-cml-1.xml	8.583Kb	chemical/x-cml	View/Open

This item appears in the following Collection(s)

- Computational Experiment Archive
  An archive of Gaussian and ORCA calculations

Search DSpace

Go

● Search DSpace
○ This Collection

Advanced Search

Browse

- All of DSpace
  - Communities & Collections
  - Titles
  - Authors
  - By Issue Date
  - Subjects
- This Collection
  - Titles
  - Authors
  - By Issue Date
  - Subjects

My Account

- Login
- Register

(b)

chempound data repository

Home | Browse | Search | SPARQL | Feeds

http://neon2.ch.ic.ac.uk:8090/content/46123872-b576-4426-b4d9-ccd22f748ff5/

N.B.: This data is automatically extracted, and may contain errors. Please report any problems in the issue tracker.

Calculation

Package	Gaussian 09 (ES64L-G09RevD.01)
Method	RB3LYP
Basis Set	TZVP

Molecule

Formula	C 18 H 23 N 1 O 3
Point Group	C1
Electronic State	1-A

Calculated Properties

Final Energy	-980.586144
Convergence RMSD	4.549E-9
Dipole	*present*
Multipole	*present*
Virtual Orbitals	*present*

Execution Environment

Hostname	GINC-CX1-1-10-2
Job name	RZEPA
Run Date	11-Nov-2013

Result files

Chemical Markup Language
Gaussian Log File
RDF XML File

Identifiers

InChI:	InChI=1S/C18H23NO3/c20-18(22-23-21-14-15-8-3-1-4-9-15)17-12-7-13-19(17)16-10-5-2-6-11-16/h1,3-4,8-10,14,17H,2,5-7,11-13H2/t17-/m0/s1
InChIKey:	WVBFBOBJPIZBFM-KRWDZBQOSA-N [?]

(c)

figshare My data Browse Upload H. Rzepa

Gaussian Job Archive for C18H23NO3

Edit article

input.gjf	preview download
checkpoint.fchk.gz	download
cml.xml	preview download
logfile.log	preview download
structure.pdb	preview download

Download all

Share this: Embed*

Cite this: Rzepa, Henry S. (2013): Gaussian Job Archive for C18H23NO3. figshare. http://dx.doi.org/10.6084/m9.figshare.845654

*The embed functionality can only be used for non commercial purposes... more

Description

Archive of work produced using the resources of the Imperial College High Performance Computing Service.

Links

- http://dx.doi.org/10042/24866
- http://dx.doi.org/10042/24849
- http://dx.doi.org/10042/24848

17 views 0 shares cites coming soon

Published on 11 Nov 2013 - 09:32 (GMT)
Filesize in total is 3.26 MB

Categories

- Organic Chemistry

Authors

Henry S. Rzepa

Tags

InChI=1S/C11H16NO2.C7H7O/c13-1...
InChIKey=ASDIUXRTGCHHNU-PPHPAT...
[CH](O)c1ccccc1.C1C[C]([CH]CC1...
C18H23NO3
b3lyp/TZVP
int=grid=ultrafine
EmpiricalDispersion=GD3
scrf=(cpcm,sol

Export

- Export to Ref. Manager
- Export to Endnote

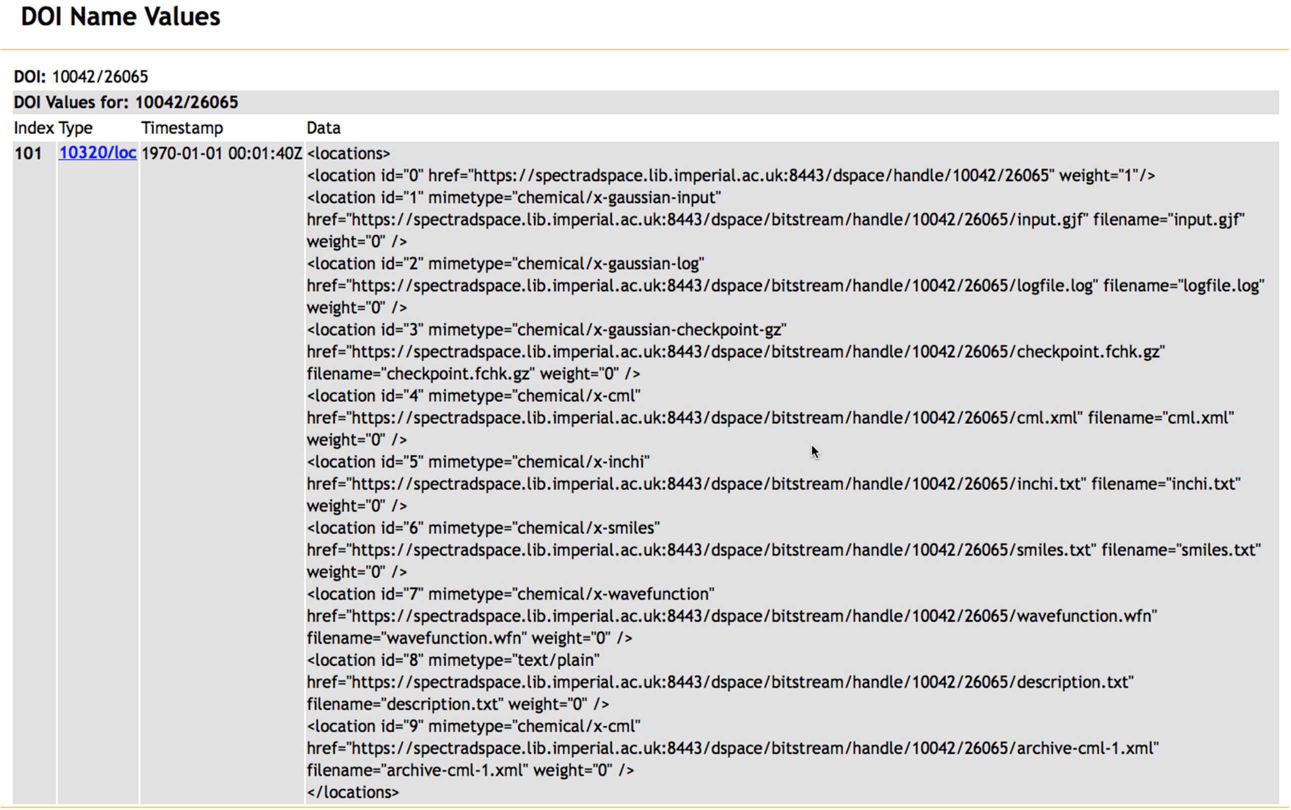
Figure 5. Repository metadata for the same dataset as sequentially published in (a) Dspace/SPECTRa, (b) Chempound, and (c) Figshare.

(2) XXXX is the prefix and is a unique identifier assigned by CNRI to an organization creating the persistent identifiers. For example, the unique prefix for our DSpace repository is 10042.

(3) YYYY is an arbitrary identifier assigned by the prefix owner that uniquely identifies a digital object; it can be any length (within limits).

The most common implementation of a handle by journal publishers is the **Digital Object Identifier** (DOI) System.[10,11] A short-form of the standard DOI has recently been introduced that limits the identifier length to seven characters and can be as few as three; the purpose being to facilitate their use by humans. Handles are typically resolved using http://hdl.handle.net or http://doi.org. These resolvers can display the records returned from the prefix server via the syntax: doi.org/10042/26065?noredirect or doi.org/10.6084/m9.figshare.845654?noredirect.

It is common to use the Handle resolver to immediately redirect the client to the destination page, often also referred to as the "landing page", using a "URL" record type. Although it would be possible to also assign such a URL record to individual data files, this rapidly becomes unwieldy, and the associations between related files are lost. Such URL records also have the limitation that there is no easy way of specifying what action is required for the file, the default being simply to attempt to display the contents in the browser DOM (document object model). A standard more flexible way is therefore needed to directly specify the individual files within a deposited record and one that may be off the landing page. This can in fact be achieved by an extension to the handle system, the poorly-known Locatt feature of the "**10320/loc**" record type that was developed to improve the selection of specific resource URLs and to add features to the handle-to-URL



DOI Name Values

DOI: 10042/26065

DOI Values for: 10042/26065

Index	Type	Timestamp	Data
101	10320/loc	1970-01-01 00:01:40Z	(see below)

```
<locations>
<location id="0" href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/handle/10042/26065" weight="1"/>
<location id="1" mimetype="chemical/x-gaussian-input"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/input.gjf" filename="input.gjf"
weight="0" />
<location id="2" mimetype="chemical/x-gaussian-log"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/logfile.log" filename="logfile.log"
weight="0" />
<location id="3" mimetype="chemical/x-gaussian-checkpoint-gz"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/checkpoint.fchk.gz"
filename="checkpoint.fchk.gz" weight="0" />
<location id="4" mimetype="chemical/x-cml"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/cml.xml" filename="cml.xml"
weight="0" />
<location id="5" mimetype="chemical/x-inchi"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/inchi.txt" filename="inchi.txt"
weight="0" />
<location id="6" mimetype="chemical/x-smiles"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/smiles.txt" filename="smiles.txt"
weight="0" />
<location id="7" mimetype="chemical/x-wavefunction"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/wavefunction.wfn"
filename="wavefunction.wfn" weight="0" />
<location id="8" mimetype="text/plain"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/description.txt"
filename="description.txt" weight="0" />
<location id="9" mimetype="chemical/x-cml"
href="https://spectradspace.lib.imperial.ac.uk:8443/dspace/bitstream/handle/10042/26065/archive-cml-1.xml"
filename="archive-cml-1.xml" weight="0" />
</locations>
```

Figure 6. Response returned for a query http://doi.org/10042/26065?noredirect showing the **filename** and **mimetype** records.

resolution process.[21] This type includes an XML-encoded list of entries, each containing a resource URL, a numeric index, an optional weight attribute for random selection, and arbitrary key:value pairs.

The resolvers hdl.handle.net or doi.org accept the URL-encoded ?locatt=key:value and return the URL of the first entry matching **key=value**. We define the keys **filename** and **mimetype**[22] in our custom DSpace handle records (Figure 6). This now works as follows:

(1) hdl.handle.net/10042/26065 or doi.org/10042/26065 resolve to the DSpace landing page for that entry.

(2) hdl.handle.net/10042/26065?locatt=filename:input.gjf or doi.org/10042/26065?locatt=filename:input.gjf resolve to a Gaussian input file (filename matching) on that repository.

(3) hdl.handle.net/10042/26065?locatt=mimetype:chemical/x-Gaussian-input or doi.org/10042/26065?locatt=mimetype:chemical/x-Gaussian-input resolve to the Gaussian input file using MIME-type matching.

(4) hdl.handle.net/10042/26065?locatt=id:1 or doi.org/10042/26065?locatt=id:1 resolve to the Gaussian input file (ID matching)

The most valuable feature of this extended experimental system is that the resolvers hdl.handle.net/api/10042/26065 or doi.org/api/10042/26065 return the JSON-encoded (JavaScript Object Notation) full handle record, which we use for processing in Javascript. There do remain issues that will need eventual resolution:

(1) Support for the 10320/loc record type is currently limited to our DSpace repository, where we control the handle resolution.

(2) Keys and values are assigned on a per-repository basis. We have used **filename** suffixes and **mimetype** that are generic keypairs. We propose that these be used as standards for any repository supporting this feature.

(3) Nevertheless, the issue arises of whether the 10320/loc record should continue to be extended in this manner or whether a new record type should instead be defined for the purpose.

(4) Everything in the handle system retrieved in this manner is invisible to search engines.

(5) Searching is still limited to indices of the messy human-readable landing pages.

FILE SETS AND MORE COMPLEX REPOSITORY OBJECTS SUCH AS DATUMENTS

With a system established that can directly and automatically address individual files (objects) held in a repository store, we can now consider how more complex objects such as datuments[10] might be constructed. A table or a figure can be authored from basic HTML5/CSS3/SVG/Javascript components, as is typical for a complex marked-up Web page. Such a table or figure can itself reference if need be hundreds of data files (a file set). In chemistry, such datument collections have been in use since 2006,[23] with descriptions such as Web-enhanced objects (WEO, by the American Chemical Society) or interactivity boxes. As these imply, they are a combination of data together with a scripted environment that renders the data into an interactive visual presentation to the reader (a datument).[10] Most of the existing examples are interwoven with the narrative of a journal article[23] and occur in the HTML version of the article, whereas a static equivalent is presented in the printable PDF version. The infrastructure described above now allows us to formally separate such datuments from the narrative by depositing the data file set into a repository and assigning it a persistent identifier of its own. The datument can then be reabsorbed back from the repository using, for example, an <iframe> declaration.

We have now created a number of such datuments on the Figshare repository, which have the following features:

(1) An HTML5 document is created that will be defined as the root document of the object.

(2) This document can invoke, for example, Javascript-based utilities that serve the purpose of transforming a data file into a visual representation. We will use the JSmol molecular viewer[24] to achieve this, which is also based on Javascript files included in the file set.

(3) Another Javascript module **resolve-api.js** has been written to implement the handle processing described above. It retrieves a JSON-array from, for example, hdl.handle.net/api/10042/26065 and converts the information into a string suitable for passing through to the JSmol system. The author provides the persistent identifier for a given datument, and the system uses the above to retrieve the appropriate individual file (dataset) from the digital repository for visual display. A specific example helps to illustrate the process:

<a href="javascript:handle_jmol('10042/26065',';frame 21;connect (atomno=1) (atomno=11) partial;')">log</a>,

with components elaborated below:

(1) The function handle_jmol (defined in resolve-api.js) does the work of producing a Jmol or JSmol figure from data obtained from the specified handle. The first argument in the script above, the handle '10042/26065', is resolved via http://doi.org into a structured JSON object. From this, the repository URL for a Gaussian logfile is obtained by MIME-type matching and passed as input to a JSMol instance. The second argument to handle_jmol() is a script passed directly to the JSMol instance to configure the rendering. In this case, the script is used to select the 21st frame from the log file (which in fact corresponds to a normal vibrational analysis showing the imaginary mode for a transition state calculation) to render the bond between atoms 1 and 11 as a partial one and to display the result. Many other script options are available, both for specifying the visual display and indeed computing new properties from the data provided. These include requesting a copy of the data to be saved to the user's local storage system.

(2) This is all presented as an HTML hyperlink, and the action results if the hyperlinked object (the text **log** in this instance, but it can be any valid object) is activated by the reader.

(3) Further unscripted actions can also be interactively initiated by the reader using the menu-driven interface of JSmol itself.

(4) At the Figshare server, the landing page for resolution of the DOI describing this datument is declared as the root document, and so when the DOI (10.6084/m9.figshare.840483 in the example below) is requested, the reader immediately receives a visual presentation as provided by JSmol.

(5) The object can also be invoked from Figshare using the following code:

<iframe src="http://wl.figshare.com/articles/840483/embed?show_title=1" width="850" height="300" frameborder="1"></iframe>

the effect being shown in WEO 1. In principle, the *iframe* declaration could itself be derived purely from the datument DOI using the locatt selection method described above; in this specific instance, it was obtained manually from the Figshare DOI landing page. It is also possible that this HTML element will be superseded by the use of a link element, regarded as having superior document properties:

<link rel="import" href="/path/to/imports/stuff.html">

(6) We have investigated two other ways in which data can be emancipated from narrative. The most simple and the one that most closely corresponds to traditional Supporting Information is a file set collection, DOI: 10.6084/m9.figshare.777773 (shortdoi: rnf). In this example (WEO 2), it corresponds to instrumental spectra captured in PDF and available for download from Figshare for inspection.

(7) A utility might be similarly packaged. If you invoke the DOI: 10.6084/m9.figshare.811862 (shortdoi: n5b) this will take you through the process of uploading data in the form of a cube of electron density values from a local file store and converting it into a so-called *noncovalent interaction* (NCI) surface and storing the isosurface in a new data file. Such a utility could, for example, be included within an article[31] describing the generation and use of such surfaces.

(8) Persistent data identifiers can also be usefully deployed in other contexts such as, for example, blogs. In conjunction with a citation extension such as Kcite,[25] a dataset can be referenced using the simple syntax [cite]10.6084/m9.figshare.840483[/cite]. This uses the handle API to retrieve the metadata associated with the item and appends this into the data citation enumerated list.[26]

■ REDUNDANCY

The functionality implemented in the *resolve-api.js* script is linear. One or more persistent identifiers for datasets specified in a datument are each resolved using a handle server into **10320/loc** handle record types pointing to a data repository server. This returns the specified files to the calling datument, which itself can be requested by a journal server via its own persistent identifier. A total of up to four services in possibly four locations can be involved in this sequence, each being a potential point of failure. Here, we briefly discuss what redundancies could be built into the system.

The general repository structure would be as follows:

Repository 1 Handle records

URL − URL of landing page (repository 1)

URL − URLs/persistent identifiers of landing page (e.g., repository 2)

URL− URLs/persistent identifiers of landing page (e.g., repository 3)

10320/loc - locations of files at repository 1

Repository 2 Handle records

URL − URL of landing page (repository 2)

URL − persistent identifier for additional deposition (e.g., repository 1)

URL − persistent identifier for additional deposition (e.g., repository 3)

10320/loc − locations of files at repository 2 etc.

(1) An attempt to retrieve a handle record from repository 1 (e.g., 10042) is made using a call of the type, for example, doi.org/api/10042/26065. This returns the JSON-encoded full handle record for repository 10042.

(2) If this record contains a **10320/loc** entry for the file of interest (e.g., a Gaussian log file), an attempt is made to retrieve the file.

(3) If this fails for whatever reason, switch to repository 2 (e.g., 10.6084) as specified in the URL records of repository 1.

(4) If 10.6084 has a matching handle record, try the calls in step 1. If not try repository 3.

(5) etc.

This scheme relies on the alternative resources having the same or a similar handle record structure, including the **10320/loc** type. Currently, only our DSpace/SPECTRa server has this specified, so the scheme above is not yet capable of practical resolution. It is nevertheless useful to include all instances of alternative depositions in the handle record if possible, in anticipation of other repositories implementing this scheme.

In the **10320/loc** scheme, locatt is a selection method that selects a location based upon a specified key-pair attribute. This scheme also allows two other selection methods, **country** and **weighted**, specified by a **chooseby** attribute.[21] If this attribute is not defined, it defaults to **locatt,country,weighted**. Our implementation (which allows the value of **chooseby** to default) uses **locatt** followed by **weighted**. We suggest it is good practice to include an explicit **chooseby** attribute in the handle records to anticipate any changes or enhancements in the repository structures. We also note that increasing consideration is being given to country records because it can be desirable to select these based on the legal frameworks in place for cloud-based data.

The redundancy model described above is suitable for a tightly coupled set of repositories into which deposition is managed by, for example, the **uportal** front end. Specifications for a complementary solution known as **SWORD1** and **SWORD2** have recently been published[27] to enable remote systems to synchronize metadata resources.

■ JOURNAL EXAMPLES

Data emancipation along the lines of the model set out in Figure 2 has been used in five articles to date.[28–32]

(1) *The Vinylcarbene–cyclopropene equilibrium of silicon: An isolable disilenyl silylene*

Narrative:[28] DOI: 10.1038/NCHEM.1751

Emancipated Data:[33] DOI: rng

(2)*Mechanistic and chiroptical studies on the desulfurization of epidithiodioxopiperazines reveal universal retention of configuration at the bridgehead carbon atoms*

Narrative: DOI:[29] 10.1021/jo401316a

Emancipated Data:[34] DOI: rns, rnf

(3) *Epoxidation of bromoallenes connects red algae metabolites by an intersecting bromoallene oxide–Favorskii manifold*

Narrative:[30] DOI: 10.1039/C3CC46720A

Emancipated Data:[35] DOI: n6q, n6r

(4) *The Houk–List transition states for organocatalytic mechanisms revisited*

Narrative: DOI:[31] 10.1039/C3SC53416B

Emancipated Data:[36] DOI: qd8, p9d, qd7, qcc, qcd, qcs, qc3, qc4

(5) *N-Heterocyclic carbene or phosphine-containing copper(I) complexes for the synthesis of 5-iodo-1,2,3-triazoles: Catalytic and mechanistic studies*[24]

Narrative:[32] DOI: 10.1021/CS500326E

Emancipated Data:[37] DOI: rnt, rfk, rfm

In all five cases, much of the data originated from the quantum modeling of the systems controlled using the uportal. The individual calculations were published into both Dspace and Figshare simultaneously as public objects. The assigned DOIs were then incorporated into tables and figures as hyperlinks using HTML. For articles 1–3 and 5, explicit data files were also included in the file collection, and the complete set was then converted to a datument, uploaded to the Figshare repository, and then itself assigned a DOI (the one quoted above). The reader can either retrieve this local copy of the data and view it in JSmol or use the original DOI for that item to download a more complete set of set, which includes the input specification that defines how the calculation was performed, and a checkpoint file containing the complete set of calculated properties. For the fourth article above, no local copies of the data files are present in the complex data object, and the calculation log files are retrieved on-demand from the original repository (in this instance DSpace/SPECTRa). There is one exception to this type of retrieval. Some of the original datasets were converted into electron density cubes using the calculation checkpoint file and a noncovalent interaction (NCI) surface was then generated. This was as two files: a.xyz coordinate file and a.jvxl surface file. Such operations can take 10–15 min or longer per molecule and are too long to be implemented as an on-demand interactive process. These specific surface files were therefore included into the datument as local files.

The model above was clearly developed to handle and illustrate the type of data we are interested in. It is not necessarily a generic solution for chemistry. It does serve to demonstrate that the entire workflow can be successfully implemented and offers an example of how solutions for many other kinds of chemical data could be developed.

■ DATA SEARCH AND DERIVED ANALYSIS

Search engines are starting to appear that focus on citable data. For example, all the metadata associated with persistent data identifiers issued by DataCite[19] is available for querying. Thus, http://search.datacite.org/ui?q=InChIKey%3DLQPOSWKBQVCBKS-PGMHMLKASA-N will return all deposited data objects associated with the InChIKey chemical structure identifier[38] LQPOSWKBQVCBKS-PGMHMLKASA-N. As the "SEO" (search engine optimization) of the metadata included in the depositions becomes more effective, so too will, for example, searches for molecular information held in digital repositories. Similar features are also offered by Google scholar[40] and ORCID (open-researcher-and-collaborator-id).[16] A search using either of these sites for one of the present authors reveals multiple data citation entries from both the DSpace/SPECTRa and Figshare repositories. Although data citations cannot be directly compared with article citations in terms of impact, the infrastructure is appearing to construct useful altmetrics to do so. Thus, one example of how added value can accrue is illustrated by a resource[40] harvesting metadata from the data repository Figshare,[18] the ORCID database of researchers[16] and collaborators, and Google Scholar.[40] This information includes metrics that allow usage of the data to be estimated and hence rather indirectly some measure of its scientific impact.

■ CONCLUSIONS AND THE WAY FORWARD

A two-component narrative–data model for the journal article (Figure 2) has the potential for solving one major current problem associated with scientific journals: the serious and permanent loss and emasculation of data from its point of creation in the laboratory to its final permanent presentation to the community in the published article. A very recent publication serves to illustrate the serious extent of the data loss.[41] A significant computational resource was used to create 123,000 sets of optimized molecular coordinates in an impressive exploration of the conformational space of four pyranosides. Only 907 of these coordinate sets are available via the article Supporting Information, and they are presented in the form of a double-column unstructured monolithic PDF document containing page breaks and numbering and with very little associated metadata for each entry. A fair amount of effort would be required by the reader of this article to (re)create a usable database from this collection for further analysis. Absence of what could be regarded as key data is unfortunately often the norm rather than the exception. Some reported data can be recast into a structured reusable form by data-mining techniques, but where it occurs has traditionally been conducted by commercial abstracting agencies, the substantial cost of which is also passed back to the scientist. In the example described here, if the data is absent in the first place, it cannot be recovered by any form of data or content mining. It is worth at this stage noting a recent and concise declaration of principles known as the *Amsterdam Manifesto*[42] regarding data sharing, which are reproduced here in full:

(1) Data should be considered citable products of research.

(2) Such data should be held in persistent public repositories.

(3) If a publication is based on data not included with the article, those data should be cited in the publication.

(4) A data citation in a publication should resemble a bibliographic citation and be located in the publication's reference list.

(5) Such a data citation should include a unique persistent identifier (a DataCite DOI recommended or other persistent identifiers already in use within the community).

(6) The identifier should resolve to a page that either provides direct access to the data or information concerning its accessibility. Ideally, that landing page should be machine-actionable to promote interoperability of the data.

(7) If the data are available in different versions, the identifier should provide a method to access the previous or related versions.

(8) Data citation should facilitate attribution of credit to all contributors.

Of particular note are articles 6 and 7 above, which we address above using the 10320/loc records. Article 8 proposes that credit for data citation should be facilitated, which resources such as ImpactStory[39] are starting to do.

The clear separation of narrative and data also addresses the vexed issues of copyright; data need no longer be constrained by limitations and costs imposed upon the narrative. There are costs of course associated with the data; the repositories must have a sound business model to ensure their long-term permanence. Whether these responsibilities are borne by the agencies where the research is initially conducted or by new agencies set up for the purpose, they should be determined by the communities involved.

Other than the infrastructures implicit in, for example, Figure 2, there is also the issue of how to persuade the authors of a scientific article to create the two components we propose. The narrative is straightforward, but can authors be persuaded to use and create the data objects? Very few are currently well versed or confident in using the HTML5/CSS3/SVG/Javascript toolkit to write scientific articles, although it is worth noting that this combination of tools is specified in an open distribution and interchange format standard for digital publications and documents known as epub3.[43] As adoption of such standards increases, so will familiarity with the concepts. An interim solution for promoting adoption may lie in the creation of standard templates; most of the complex detail is actually carried in Javascript and stylesheet declarations that need not be edited. They can be easily transcluded into a document template via header declarations. Indeed, almost the only actual code that they need be aware of is that shown earlier:

<a href="javascript:handle_jmol('persistent identifier', 'presentation script')">linked hypertext</a>

A more challenging problem is that most authors would see little reward in the current system for undertaking such tasks. Such rewards accrue from the narrative they present (be it scientific article or Ph.D. dissertation) and not currently from the data they associate with that narrative (except of course that the narrative would not stand on its own if no data had been presented somehow). Here, the scientific community must agree that preserving data and curating it for the future is a worthwhile activity and bestow the appropriate rewards for doing so or indeed apply sanctions if it is not done. Technically at least, there is nothing preventing the scientific journal from evolving in this manner. Indeed, there are already other emerging examples of such evolution toward data-rich publishing. The Royal Society of Chemistry has launched Chemspider Synthetic pages,[44] an example of what they call micro-publishing, and Nanopublications[45] is an example of a Semantic Web technology similar in concept to Chempound.[17] We are encouraged that such resources will become mainstream in the very near future.

■ ASSOCIATED CONTENT

Ⓦ Web-Enhanced Features

WEOs 1 and 2 are available in the HTML version of the paper.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rzepa@ic.ac.uk, h.rzepa@imperial.ac.uk.

Notes

All additional information is available via the digital repository links cited in the text and references.

The authors declare no competing financial interest.

■ REFERENCES

(1) Oldenburg, H. Epistle Dedicatory. *Philos. Trans.* **1665**, 1.

(2) Jessop, D. M.; Adams, S. E.; Willighagen, S. E.; Hawizy, E. L.; P. Murray-Rust, P. OSCAR4: A flexible architecture for chemical text-mining. *J. Cheminf.* **2011**, 3, 41.

(3) James, D.; Whitaker, B. J.; Hildyard, C.; Rzepa, H. S.; Casher, O.; Goodman, J. M.; Riddick, D.; Murray-Rust, P. The case for content integrity in electronic chemistry journals: The CLIC Project. *New Rev. Inf. Networking* **1995**, *1*, 61−69.

(4) Rzepa, H. S. Emancipate your data. *Chemistry World*, **2013**, *09*, DOI: 10042/a3uxk.

(5) Rubacha, M.; Rattan, A. K.; Hosselet, S. C. A review of electronic laboratory notebooks available in the market today. *J. Lab. Autom.* **2011**, *16*, 90−98.

(6) Coles, S. J.; Frey, J. G.; Bird, C. L.; Whitby, R. J.; Day, A. E. First steps towards semantic descriptions of electronic laboratory notebook records. *J. Cheminf.* **2013**, *5*, 52.

(7) Downing, J.; Murray-Rust, P.; Tonge, A. P.; Morgan, P.; Rzepa, H. S.; Cotterill, F.; Day, N.; Harvey, M. J. SPECTRa: The deposition and validation of primary chemistry research data in digital repositories. *J. Chem. Inf. Model.* **2008**, *48*, 1571−1581.

(8) Shotton, D.; Portwin, K.; Klyne, G.; Miles, A. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *PLoS Comput. Biol.* **2009**, *5*, e1000361.

(9) Shotton, D. Semantic publishing: The coming revolution in scientific journal publishing. *Learned Publishing* **2009**, *22*, 85−94.

(10) For a summary, see Rzepa, H. S. Chemical datuments as scientific enablers. *J. Chemoinf.* **2013**, *5*, 6.

(11) Paskin, N. Digital object identifiers for scientific data. *Data Sci. J.* **2005**, *4*, 12−20.

(12) Paskin, N. Digital Object Identifier (DOI) System. *Encyclopedia of Library and Information Sciences*, Third ed. **2011**.

(13) Murray-Rust, P.; Rzepa, H. S. CML: Evolution and design. *J. Cheminf.* **2011**, *3*, 44.

(14) Figshare application programming interface. http://api.figshare.com/docs/intro.html (accessed July 17, 2014).

(15) Butler, D. Scientists: Your number is up. *Nature* **2012**, *485*, 564.

(16) For an example relating to one of the authors, see http://orcid.org/0000-0002-8635-8390 (accessed July 17, 2014).

(17) Adams, S.; Murray-Rust, P. Chempound − A Web 2.0-inspired repository for physical science data. *J. Digital Inf.* **2012**, *13*, 5873.

(18) Figshare. http://api.figshare.com/docs/intro.html (accessed July 17, 2014).

(19) DateCite. http://www.datacite.org/ (accessed July 17, 2014).

(20) See http://www.nature.com/sdata/, where a data descriptor is defined as "a primary article-type, the Data Descriptor...designed to make your data more discoverable, interpretable and reusable.".

(21) Kahn, R., Wilensky, R. A framework for distributed digital object services. *Int. J. Digital Libraries*, **2006**, *6*, 115−123. http://www.handle.net/documentation.html, specifically http://www.handle.net/overviews/handle_type_10320_loc.html and http://www.handle.net/overviews/handle_type_10320_loc.html#conneg (accessed July 17, **2014**).

(22) Rzepa, H. S.; Murray-Rust, P.; Whitaker, B. J. The application of chemical multipurpose internet mail extensions (Chemical MIME) internet standards to electronic mail and World-Wide Web information exchange. *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 976−982.

(23) Rzepa, H. S. Activated data in chemistry publications, 2006−2013. *Figshare*. 2014, DOI: rnp (accessed July 17, 2014).

(24) Hanson, R. M.; Prilusky, J.; Zhou, R.; Nakane, T.; Sussman, J. JSmol and the next-generation Web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.* **2013**, *53*, 207−256.

(25) Cockell, S.; Lord, P. KCite Plugin. *Knowledge Blog*, **2011**. http://knowledgeblog.org/kcite-plugin (accessed July 17, 2014).

(26) Rzepa, H. S. Intersecting Paths in Molecular Energy Surfaces. *Chemistry with a Twist (Blog)*, **2014**. DOI: 10042/a3uzb (accessed July 17, 2014).

(27) SWORD Specifications. http://swordapp.org/sword-v1/the-specification/ and http://swordapp.org/sword-v2/sword-v2-specifications (accessed September 8, 2014).

(28) Scheschkewitz, D.; MCowley, M. J.; Huch, V.; Rzepa, H. S. The vinylcarbene−cyclopropene equilibrium of silicon: An isolable disilenyl silylene. *Nature Chem.* **2013**, *5*, 876−879.

(29) Cherblanc, F. L.; Lo, Y.-P.; Herrebout, W. A.; Bultinck, P.; Rzepa, H. S.; Fuchter, M. J. Mechanistic and chiroptical studies on the desulfurization of epidithiodioxopiperazines reveal universal retention of configuration at the bridgehead carbon atoms. *J. Org. Chem.* **2013**, *78*, 11646−11655.

(30) Braddock, D. C.; Clarke, J.; Rzepa, H. S. Epoxidation of bromoallenes connects red algae metabolites by an intersecting bromoallene oxide−Favorskii manifold. *Chem. Commun.* **2013**, *49*, 11176−11178.

(31) Armstrong, A.; Boto, R. A.; Dingwall, P.; Contreras-García, J.; Harvey, M. J.; Mason, N. J.; Rzepa, H. S. The Houk−List transition states for organocatalytic mechanisms revisited. *Chem. Sci.* **2014**, *5*, 2057−2071.

(32) La, S.; Rzepa, H. S.; Díez-González, S. N-Heterocyclic carbene or phosphine-containing copper(i) complexes for the synthesis of 5-iodo-1,2,3-triazoles: Catalytic and mechanistic studies. *ACS Catal.* **2014**, *4*, 2274−2287.

(33) Scheschkewitz, D.; Cowley, M. J.; Huch, V.; Rzepa, H. S. Table 1: The vinylcarbene−cyclopropene equilibrium of silicon: An isolable disilenyl silylene. *Figshare*. 2014, rng (accessed September 10, 2014).

(34) Cherblanc, F. L.; Lo, Y.-P.; Herrebout, W. A.; Bultinck, P.; Rzepa, H. S.; Fuchter, M. J. Mechanistic and chiroptical studies on the desulfurization of epidithiodioxopiperazines reveal universal retention of configuration at the bridgehead carbon atoms. *Figshare*. 2014, rns (accessed 2014).

(35) Braddock, D. C.; Clarke, J. H. S. Interactivity box. Epoxidation of bromoallenes connects red algae metabolites by an intersecting allene oxide−Favorskii manifold. *Figshare*. 2013, DOI: n6q and n6r (accessed September 10, 2014).

(36) Armstrong, A.; Boto, R. A.; Dingwall, P.; Contreras-García, J.; Harvey, M. J.; Mason, N. J.; Rzepa, H. S. Table 3: Calculated transition state properties for R=Ph (scheme 2). *Figshare*. 2014, DOI: qcc (accessed September 10, 2014).

(37) La, S.; Rzepa, H. S.; Diez-González, S. N-Heterocyclic carbene or phosphine-containing copper(I) complexes for the synthesis of 5-iodo-1,2,3-triazoles: Catalytic and mechanistic studies. *Figshare* . 2014, rnt, rfk, rfm (accessed September 10, 2014).

(38) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI − The worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7.

(39) ImpactStory. www.impactstory.org/rzepa/ (accessed July 17, 2014).

(40) Henry S. Rzepa. http://scholar.google.co.uk/citations?user=ljZtPwkAAAAJ&hl=en&oi=ao&cstart=304 (accessed July 17, 2014).

(41) Mayes, H. B.; Broadbelt, L. J.; Beckham, G. T. How sugars pucker: Electronic structure calculations map the kinetic landscape of five biologically paramount monosaccharides and their implications for enzymatic catalysis. *J. Am. Chem. Soc.* **2013**, *136*, 1008−1022.

(42) Crosas, M., Carpenter, T. Shotton, D., and Borgman, C. Joint Declaration of Data Citation Principles. http://www.force11.org/AmsterdamManifesto (accessed March 2013).

(43) International Digital Publishing Forum. A distribution and interchange format standard for digital publications and documents. http://idpf.org/epub/30 (accessed September 8, 2014).

(44) Tkachenko, V.; Batchelor, C.; Karapetyan, K.; Sharpe, D.; Williams, A. J. ChemSpider reactions: Delivering a free community resource of chemical syntheses, Abstracts of Papers, 245th ACS National Meeting and Exposition, New Orleans, LA, United States, April 7−11, 2013, CINF-95. http://cssp.chemspider.com (accessed September 8, 2014).

(45) Patrinos, G. P.; Cooper, D. N.; van Mulligen, E.; Gkantouna, V.; Tzimas, G.; Tatum, Z.; Schultes, E.; Roos, M.; Mons, B. Micro-attribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum. Mutat.* **2012**, *33*, 1503−12 and http://nanopub.org/wordpress (accessed September 10, 2014).