

COMMENTARY

Open Access

Chemical datuments as scientific enablers

Henry S Rzepa

Abstract

This article is an attempt to construct a chemical datument as a means of presenting insights into chemical phenomena in a scientific journal. An exploration of the interactions present in a small fragment of duplex Z-DNA and the nature of the catalytic centre of a carbon-dioxide/alkene epoxide alternating co-polymerisation is presented in this datument, with examples of the use of three software tools, one based on Java, the other two using Javascript and HTML5 technologies. The implications for the evolution of scientific journals are discussed.

Background

Chemical sciences are often considered to stand at the crossroads of paths to many disciplines, including molecular and life sciences, materials and polymer sciences, physics, mathematical and computer sciences. As a research discipline, chemistry has itself evolved over the last few decades to focus its metaphorical microscope on both far larger and more complex molecular systems than previously attempted, as well as uncovering a far more subtle understanding of the quantum mechanical underpinnings of even the smallest of molecules. Both these extremes, and everything in between, rely heavily on data. Data in turn is often presented in the form of visual or temporal models that are constructed to illustrate molecular behaviour and the scientific semantics. In the present article, I argue that the mechanisms for sharing both the underlying data, and the (semantic) models between scientists need to evolve in parallel with the increasing complexity of these models. Put simply, the main exchange mechanism, the scientific journal, is accepted [1] as seriously lagging behind in its fitness for purpose. It is in urgent need of reinvention; one experiment in such was presented as a data-rich chemical exploratorium [2]. My case here in this article will be based on my recent research experiences in two specific areas. The first involves a detailed analysis of the inner kernel of the Z-DNA duplex using modern techniques for interpreting the electronic properties of a molecule. The second recounts the experiences learnt from modelling the catalysed alternating co-polymerisation of an alkene epoxide and carbon dioxide.

An attempt will here be made to present both stories in the form of a chemical datument. This portmanteau word refers to a data-rich document, and is used here to mean a document that describes a story of chemical research in a manner which allows the data underpinning the discourse to be provided as an integral part of that story. Although the term *datument* was originally explicitly coined in a scientific context in 2004 [3], arguably the first true datument on the topic of molecular science was published in a mainstream peer-reviewed chemistry journal had appeared as early as 2001 [4]. This latter article has several unusual attributes. It attracted an editorial comment [5] that describes the article as an "*interesting experiment*", but which also concludes that "*it wasn't easy to deal with by any means*", referring to the production process. In this sense, this article was also arguably ahead of its time, since it required an early beta version of a Web browser to expose the available data to the reader (Internet Explorer 6.0 or 6.5) using a combination of XML as the carrier of the data/content and XSLT stylesheets to transform this for browser presentation. Modern browsers support newer versions of the standards used for these operations and some 11 years on, the original article now needs "maintenance" to recover these aspects. But nevertheless, the data contained with it, expressed in XML and CML [6] as the principle carrier of chemical information retains all of its original semantic meanings, and it is specifically the presentational layer that requires the maintenance. This of itself raises some interesting issues which will need to be addressed in the future. In turn, it may also mean that the presentational mechanisms used in the current article may equally need curation in the future. In the last 11 years nevertheless, made major advances in this

Correspondence: rzepa@imperial.ac.uk
Department of Chemistry, Imperial College London, South Kensington
Campus, London, UK

area of semantic scientific publishing have been made, and the reader is referred to several excellent reviews of this area for further information [7,8].

Case 1. The inner secrets of the structure of Z-DNA

In a previous article on the topic [9], I recounted how early papers describing the molecular structure of the DNA double-helix were quite data-impooverished. The issue related to why this molecule adopted a left or right handed helical wind and what the factors influencing this balance might have been. To analyse these features requires evaluating the wavefunction of a (small fragment) of the system. This is then inspected for not only the electronic interactions between covalent bonds themselves but also the nature of any close (non-covalent) contacts between pairs of atoms which do not classify as bonds or appear in a bond connection table (and are therefore un-indexed and hence neglected) [2].

Two analytical tools were utilised and the results are presented here.

1. The first was a so-called Natural-bond-orbital analysis [10,11], the basis of which is to transform the computed wavefunction of the molecule into localised functions called NBOs, which take two basic forms. The first has a two-electron occupancy, and is deemed to be a potential *donor* of these two electrons (BD). The second is an NBO with zero-electron occupancy, and which is deemed to be an acceptor of electrons (BD*). The extent to which the latter influences the former is quantified by a perturbation energy $E(2)$. The magnitude of this term in turn depends on both the difference in energy between the two interacting NBOs and the degree of overlap between them. Whilst the former can be expressed simply by an energy, the latter lends itself to visual presentation as a set of overlapping isosurfaces.
2. The NBO procedure, by definition, explores how bonds BD interact with anti-bonds, BD*, within a molecule. But almost as important are the regions which conventionally are not defined as bonds, but are instead referred to as non-covalent interactions within a molecule. A hydrogen bond is one example of this type, but they can also refer to weaker interactions. It is important to appreciate that although any single such interaction may be quite weak, repeated occurrence in a large molecule will tend to accumulate the effect. This NCI procedure [12-14] involves computing the reduced (electron) density gradient isosurfaces for the molecule in question, and filtering the range of this value to that which focuses only on the weakly interacting regions. A further property (the density Laplacian) can be

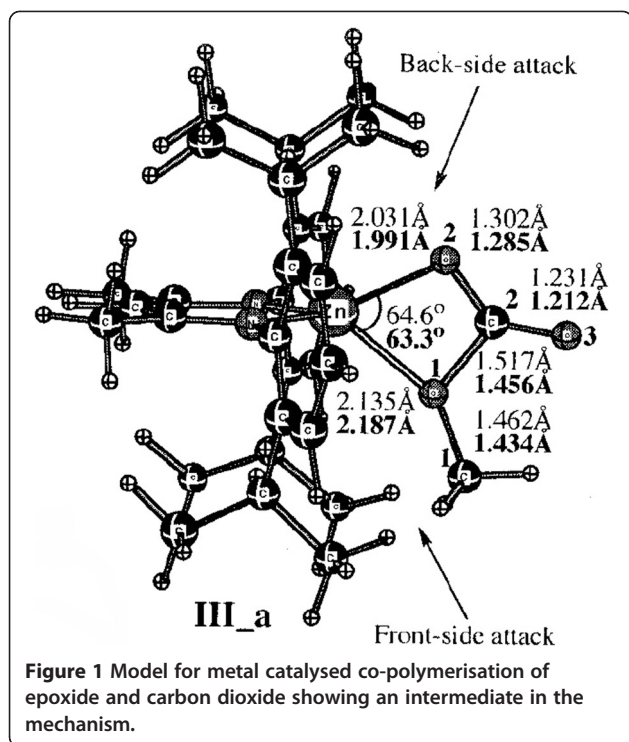
used in conjunction with colour coding of the isosurface to indicate whether the interaction is attractive or repulsive. Again, this is a complex visual surface generated from a computed molecular wavefunction.

You might appreciate that communicating these concepts to the reader using merely descriptive text and static diagrams (even the use of colour by authors in diagrams can incur very substantial/additional costs charged by the publisher) may be very limiting. Of course, I have also selected this example for precisely such difficulty, and to introduce how a datument might go a long way towards addressing this problem.

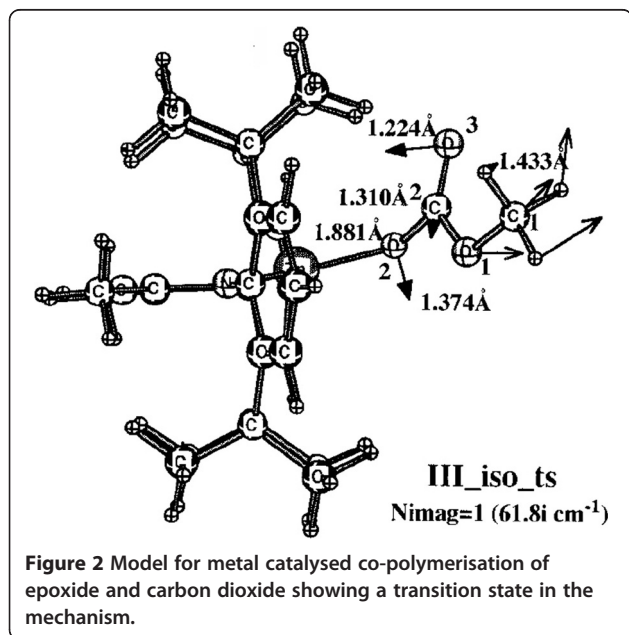
Case 2. Unravelling the mechanism of co-polymerisations

Optimising the effective use of carbon dioxide as a C₁ feedstock for manufacturing polymers is a pressing scientific challenge [15]. The answer lies in understanding the complex catalytic chemistries of quite large molecular systems. These chemistries can increasingly be successfully modelled using modern quantum chemical theories. The capability to address these complex catalytic systems first emerged around 2002 when, in something of a tour-de-force for that period, Morokuma and co-workers reported their exploration of the mechanism of Zinc(II)-catalysed alternating co-polymerisation of carbon dioxide with cyclohexene epoxide [16]. The objective then was to develop a rational explanation for why the polymer alternated, *i.e.* by addition of one molecule of carbon dioxide monomer was invariably followed by one molecule of cyclohexene epoxide, and then again by CO₂. By 2012 the complexity and subtle nature of the challenge had increased, the challenge now being an understanding [17] of how asymmetric induction in the resulting polymer can be achieved. Answering such questions involves a detailed and intricate knowledge of the reacting (covalent or ionic) bonds themselves and (as with the DNA project discussed above) the nature of the non-covalent interactions [12-14]. The model one builds to explore such aspects may contain up to 200 atoms (in a polymer of course, it is potentially much larger, and one has to truncate the model to manageable proportions). Two examples of this complexity are shown below (Figures 1 and 2).

Firstly, I should state that this article is typical of the period (a mere ten years ago), whereby limitations on the page length precluded inclusion of any tables of coordinates (=data) associated with this figure. Seeking to explore the attributes of the bonds and appropriate non-covalent interactions requires data. It was however reasonably common in that period, but certainly not mandatory, for such data to be included in the *supporting information* (note it is styled as information, not



data). In this particular instance, no such information is actually available. Even in 2011 when another article on this topic was published [18], and for which supporting information was available, one finds a paucity of the type of data required to reconstitute any of the models on which all the assertions in the body of the article are based. Unlike the area of crystallography, where data deposition is mandatory [19], no such requirements exist



for many other areas of data-rich chemistry, including e.g. for computational chemistry.

The figures themselves (Figure 1) contain information that only a human could attribute meaning to (the figures could not for example be automatically mined for information in the manner that eg the OSCAR project [20] has demonstrated). Even this author (as a human), struggled to reconstitute a usable model from these figures. There is indeed some numerical information associated (note that text mining software cannot access this information) with the figure, styled in *lightface* and *boldface* (each being the result of using a different theoretical method, as explained in the original figure caption). but for a system with perhaps 100 atoms (counting the actual number from the figure is essentially impossible), this represents only 7 of the 294 ($=3N-6$) variables required to precisely define the three dimensional model. Other annotations are present in the figure; the designation *back-side* or *front-side* attack have a semantic meaning that is not immediately obvious to someone not very familiar with the original research. One of the diagrams (Figure 2) has lines carrying arrowheads, the other (Figure 1) does not. Although not explained in the figure caption, an experienced (human) computational chemist can nevertheless infer the semantics that these are normal vibrational mode displacement vectors. Furthermore it is likely that this particular mode is selected because it represents the vibration for (harmonic) motion of the atoms at the transition state for the reaction. These vectors may or may not be mass-weighted, so their length may carry no significance. One might associate this feature with the (non-minable) text below it indicating that the (imaginary) wave number of this mode is $61.8i \text{ cm}^{-1}$. One might infer that the diagram in which such vectors are absent (Figure 1) is not a transition state, but an equilibrium structure in which all the normal modes are real and not imaginary. That is a lot of implicit semantics, which a (trained) human can cope with, but which again a machine is unlikely to. I have gone into this one figure in some detail, not at all to criticise the authors for providing a deficient figure, but to illustrate the extent to which the data from which the model can be built is lacking, together with the semantics needed to put that data to good use.

The preceding analysis of these articles [16,17,20] originated as an outcome of the exploration of the mechanism of this reaction [15]. In order to compare new models with the earlier ones, it was essential to analyse the information carried in Figures 1 and 2, in particular whether the system shown in these figures represented a mono or a bi-metallic system. Because of the particular projection onto two dimensions used in the published figures, it was not possible to establish with absolute certainty whether a second Zn atom might be present, but be obscured in the

figure. After about an hour of such analysis, a key (semantic) connection was made, the realisation that bimetallic models had only been explicitly discussed in the literature from the year 2003 onwards. From this chronology, it was possible to conclude that it was probable that the 2002 model did not represent a bimetallic system. Unfortunately, there were no data presented in the form of atom lists which would have instantly clarified this aspect.

Discussion

When it came to communicating our own researches on these topics [15], it was imperative that we should explore how to not propagate the difficulties we ourselves had experienced with the earlier literature onto future readers of our own article. How could the appropriate data (and semantics) be incorporated into a journal article in 2012? One publisher is already making a virtue of this aspect in advertising *the-article-of-the-future* [21]. These articles feature data-based components such as compound information, experimental flowcharts and embedded video (although potentially data rich, this data can often be inaccessible in the same sense that was discussed for Figures 1 and 2. An animation for example can only be viewed from the author's predetermined viewpoint, and not from the reader's). At the time of writing, there are no examples of articles-of-the-future [21] suitable for describing the type of research discussed above. In fact, we had started an exploration in 2006 [22] of data-rich articles which contained so-called web-enhanced objects (tables and figures) in conjunction with other publishers [23]. These constitute datuments in the sense that not only can a human easily re-use the data carried in such an article, but in theory so could a (much more pedantic) software agent tasked to mine the data. Such mining is facilitated by using XHTML to express the datument (PDF versions of the articles were also made available by the publisher, but the semantic data-enrichment is not present in these versions). These articles however were not optimised for their semantic attributes. Our 2001 datument [4] was expressed entirely in XML, and presented to the reader by an on-the-fly transformation of that XML using appropriate XSLT stylesheets. Our 2006+ web-enhanced figures and tables accepted the practical reality that publishers were not yet ready to accept XML/XSLT submissions, as well as the observation that very few authors had the time and skills to author datuments in this format.

The digital data repository and data semantics

It is important to distinguish between data and the wrapper by which it is presented to the (human) reader. There are several considerations.

1. The raw unprocessed data may be too large to reasonably include in a datument.

2. Or it may take a lot of processing power, or require complex computer code, to transform the data into a meaningful visual appearance.
3. The raw data may have no meta-data associated with it, and hence may not be semantically processable or searchable.

The expedient adopted here is to include at least sufficient well-structured (i.e. XML-based) data to allow regeneration of the original (large) dataset with almost no effort required to achieve this. If the transformation of the data for visual presentation is itself too complex to be handled by a browser in real-time, then the result of that transform can itself be included in the datument (again ideally as an XML dataset). Finally, to complete the utility of the datument, the (possibly large) inputs and outputs from which the dataset derives can be linked to a digital repository where the semantic enrichment can be added in a largely automatic manner. This in turn would allow either humans or software agents to process them if desired.

Examples of digital repositories containing molecular data

1. The DSpace-based SPECTRa repository [24]. Each entry here is created from raw data files, and the metadata is added by post-processed recognition of regular patterns in the data, along with meta-data captured from the user or system at time of deposition. The resulting data-collection is identified with a unique handle, which can be resolved by the same resource as the digital object identifier (DOI) now ubiquitously used in journal articles such as the one you are current reading. A typical set of meta-data and raw data for the type of calculation reported in this article can be seen in Figure 3. The raw files themselves are associated with appropriate MIME types [25] to enable automated processing when downloaded. The entire collection is created automatically from the job submission portal used to create the data in the first place, ensuring it is as free of human error as possible. The unique molecule identifiers (the InChIKey) are captured as assigned to Dublin-core fields, and can also serve well as nodes in an RDF description [26], although Dspace itself cannot be used to invoke a semantic query based on such RDF declarations.
2. Figshare [27] is a new, more general data repository, carrying much the same meta-data as the DSpace example (Figure 4). It too associates a DOI with the data set, and can be used in the same manner.
3. ChemPound [28] was designed specifically to archive chemical information, and generates meta-data for RDF declaration at a much more finely grained level.

C 1 H 5 N 1 Si 2

dc.contributor	Henry S Rzepa
dc.creator	Henry S Rzepa
dc.date.accessioned	2012-07-04T20:47:08Z
dc.date.available	2012-07-04T20:47:08Z
dc.date.issued	2012-07-04
dc.identifier	InChI=1S/CH5NSi2/c2-1(3)4/h2H,3-4H2
dc.identifier	InChIKey=C1JBPJWLXUDLY-UHFFFAOYSA-N
dc.identifier.uri	http://hdl.handle.net/10042/20199
dc.description	# rwb97xd/6-311g(d,p) opt(calcall,ts,noeigentest)
dc.publisher	Imperial College London
dc.rights	Henry S Rzepa
dc.title	C 1 H 5 N 1 Si 2
dc.type	Gaussian job archive

Files in this item

Files	Size	Format	View
input.gif	691bytes	chemical/x-gaussian-input	View/Open
logfile.log	93.84Kb	chemical/x-gaussian-log	View/Open
checkpoint.fchk.gz	138.8Kb	chemical/x-gaussian-checkpoint	View/Open
cml.xml	1.301Kb	chemical/x-cml	View/Open
inchi.txt	73bytes	chemical/x-inchi	View/Open
smiles.txt	22bytes	chemical/x-smiles	View/Open
wavefunction.wfn	0bytes	Unknown	View/Open
description.txt	51bytes	Text file	View/Open
archive-cml-1.xml	4.325Kb	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

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](#)

Figure 3 A data repository entry in DSpace, showing associated chemical metadata. The original can be retrieved at handle: 10042/20199.

For example, the final total energy in a quantum mechanical calculation is identified and associated with an RDF triple. The Chempound repository is also the only one specifically designed for RDF-SPARQL like semantic queries of the triple store. Unlike the first two repositories however, Chempound does not (yet) generate a unique handle for identification of each entry.

Examples of digital repositories containing other types of data

Examples of other projects for depositing and curating data include DataOne [29], Dryad [30] (another DSpace-based repository), DataCite [31] (which also provides DOI identifiers for each collection) and DataShare (an online digital repository of multi-disciplinary research datasets produced at the University of Edinburgh) [32]. There are also separate initiatives for developing standards for the deposition and searching of data [33]. It is becoming clear that such repositories are bifurcating into two types; those for general data that carry only general meta-data descriptors for the content, and

subject-specific repositories which serve to harvest much more finely tuned meta-data, in turn allowing much more specific searches of the repository to be made. If the fragmentation into increasingly subject-specific content continues, then the challenge will refocus on searching across different repositories for related data sets between which there may be valuable synergies.

Two examples of how such a strategy may be deployed are discussed next.

The Java-based datument

Transclusion of (chemical) data-objects into HTML pages for humans to read has evolved in three phases. One of the earliest was introduced around 1996 and benefited from the close physical proximity in San Francisco of two commercial organisations, Netscape and MDL Ltd and the earlier publication of an article on the topic [34]. The reader had to download the software (Chime) and install it each computer they wished to use. This was replaced a few years later by the use of Java, whereby the necessary software archive (.jar) is downloaded automatically when the data-object is loaded. This mode is used to

The image shows a screenshot of a Figshare data repository entry. At the top, the Figshare logo is on the left, and a search bar with 'Browse' and 'Upload' buttons is in the center. On the right, there are 'Sign up' and 'Login' buttons. The main content area is titled 'Gaussian Job Archive for B2(2-)'. Below the title, there is a list of files: 'checkpoint.fchk.gz', 'cml.xml', and 'logfile.log'. Each file has a 'download' button, and 'cml.xml' and 'logfile.log' also have a 'preview' button. To the right of the file list, there are statistics: '146 views' in a blue circle, '0 shares' in a green circle, and a grey circle with 'cites coming soon'. Below the statistics, it says 'Published on 15 Jun 2012 - 14:13 (GMT)' and 'Filesize in total is 303.21 KB'. There is a 'Download all' button. Below the file list, there is a 'Share this:' section with buttons for Facebook (Share 0), Twitter (Tweet 0), and a '+1' button (0). Below that, it says 'Cite this: Gaussian Job Archive for B2(2-). M J Harvey, Henry Rzepa. [figshare](#). Retrieved 13:08, Jul 06, 2012 (GMT) <http://dx.doi.org/10.6084/m9.figshare.92393>'. Below the citation, there is a 'Description' section with the text 'Gaussian Job Archive'. Below that, there is a 'Comments (0)' section with the text 'You must be [logged in](#) to post comments.'. On the right side of the page, there are sections for 'Categories' (Organic chemistry), 'Authors' (M.J Harvey, Henry Rzepa), 'Tags' (CCSD(T)/aug-cc-pvqz, B2(2-), B#B, InChIKey=QSJRRLLWJRLPVID-UHFFFA..., InChI=1/B2H4/c1-2/h1-2H2), and 'Export' (Export to Ref. Manager, Export to Endnote, Export to Mendeley). At the bottom right, there is a QR code.

Figure 4 A data repository entry in Figshare, showing associated chemical metadata. The original can be retrieved at doi: 10.6084/m9.figshare.95816.

the present day on most conventional operating systems and is illustrated below. It makes use of a digitally signed .jar file, which allows data to be extracted from the display by the user (hence the prompt the user receives to accept the datument source when it is loaded). An example of this is illustrated in Additional file 1: Interactivity box 1.

The HTML5 based datument

This mode of presentation takes advantage of the new generation of mobile devices such as touch-screen

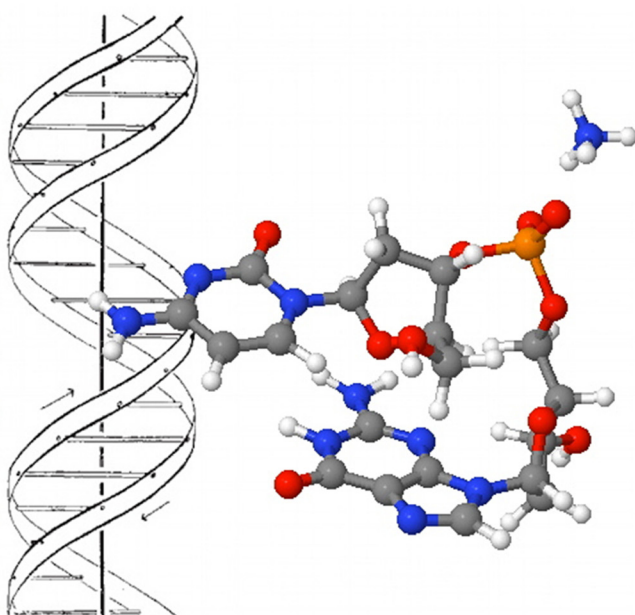
tablets. It also presents a strategy for browser and device independence, since there does seem to be a trend towards increased adoption of standards centered around HTML5 for both browsers and the devices they run on. In this regard, the design of mobile devices appears to be evolving away from dependency on power-consuming software environments such as Java, and towards data-handling environments such as JSON [35] and JavaScript, utilising lightweight graphics renderings based on WebGL in combination with HTML5 that can take full advantage of such an environment. An advantage over

Interactivity box 1:^a Data-based object illustrating various aspects of the interactions at the heart of Z-DNA.

We will pose the question: what are the origins of the relative stability of the left handed CG-rich duplex ([View/download model](#))^{b,c} compared with the isomeric [right](#) handed helix? The following are some of the chemically interesting interactions that can be identified at the heart of the system:

1. A furanose oxygen approaches to within **2.85Å** of a guanine ring
2. A furanose oxygen approaches to within **at least of 2.85Å** of a guanine ring
3. A furanose OC-H hydrogen is identified within **2.48Å** of a second furanose oxygen
4. This hydrogen bond is facilitated by a C-H_σ (BD)/C-O_σ* (BD*) [anti-periplanar](#) alignment, which results in acidification of the C-H bond (shown as magenta bonds). The NBO interaction energy is **E(2) 5.8 kcal mol⁻¹** which may be considered of a chemically significant magnitude.
5. There is an [anomeric interaction](#) between guanine and the ribose. The [overlap and interaction between](#) the N_{Lp} and C-O_σ* orbitals results in a NBO perturbation energy of E(2) of **11.6 kcal mol⁻¹**. The colour coding of the NBO orbital isosurface (contoured at 0.02au) indicates a positive orbital overlap between the blue and purple surfaces (of the acceptor and donor orbitals respectively), and likewise between the red and orange surfaces. This stereoelectronic interaction helps to rigidify the orientation of the plane of the nitrogen base with respect to the sugar ring, and might be expected to influence the higher order structures of the polymer such as supercoiling and unwinding.
6. There is a further rather weaker anomeric interaction in a [cytosine-furanose pair](#) (E2 6.8).
7. A *Gauche*-like conformation of the [ethane-1,2-diol fragment](#) (gold bond) is also aided by a stereoelectronic interaction between the O_{Lp} and an adjacent C-O_σ* bond.
8. A non-covalent-interaction (NCI) analysis reveals [further weaker interactions](#) in this system, which might otherwise be overlooked. The isosurface colour coding maps to blue = attractive, green = weakly attractive, yellow = weakly repulsive and red = strongly repulsive interactions
9. Data can be retrieved from this object by invoking the user menu or [an appropriate script](#).

Mindful that all these interactions are repeated throughout the DNA polymer, their accumulation might be expected to influence the higher order structures of this important biopolymer.



Jmol_S

NCI(small molecules) Search

^a The model includes two bases, two furanose rings linked by a phosphate, of which the charge is balanced by an ammonium cation. Data must be loaded by invoking the hyperlinks present in the preceding text. To invoke the **user menu**, right (or ctrl) click in the viewing area, a feature which can also be used to save the model data. The original complete data set is also available at <http://dx.doi.org/10042to-13541> via a digital repository. ^bThe interactive components are created by invoking the Java-based Jmol applet in the browser window. If your browser does not support Java, the display will instead default to using ChemDoodle. This currently does not support the scripted interactive functionality and only the model coordinates will be loaded. ^c This file is downloaded as a [ZIP-compressed archive](#) containing the components required to reconstitute the scene.

a Java-based solution is that the necessary display code is much smaller and runs natively within the browser rather than as a Java virtual environment. Two such implementations for HTML5 are ChemDoodle [36] and GLMol [37], for which examples of different types of transcluded data Additional files 2-4 (Interactivity Boxes) are shown below [38]. Data can also be flexibly retrieved from such objects [39]. A comparison between the static Figures 1 and 2 and the data-rich interactivity boxes serves to illustrate how an enhanced perception by reader can be achieved when they are allowed to interact with the datument.

The authoring perspective

I should also describe the experience of creating such figures from an author's perspective. The data-carrying components are embedded in the form of scripts, which themselves can be regarded by the author as (publisher-provided?) templates, and the only real task is to provide appropriate variable names.

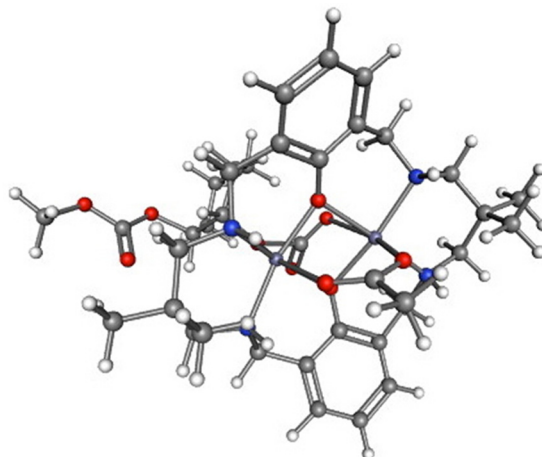
1. Additional file 1: Interactivity box 1 is created using a script for a device-sensitive display, which supports either a Java-based Jmol applet, or a Javascript-based ChemDoodle canvas:

```
<script type="text/javascript" title="Script for creating a Canvas with device-sensitive display"> Figure 3 = Jmol.getApplet ("Figure 3", Info1) Jmol.script (Figure 3,"load dna_mo148.cub.xyz;background image 'helix-back.jpg';spin 5;#alt:LOAD dna_mo148.cub.xyz") </script>
```
2. The links in the interactivity box of Figure 3 are created as:

```
<a href="javascript:Jmol.loadFile (Figure 3,'1ZNA-H.mol',';background%20image%20"helix-back.jpg";measure%2083%20114;measure%20124%20155; measure%2045%2076;measure%204%2035;write%20jmol%20Figure3-1.jmol;)">(View/download model)</a>
```
3. Additional file 2: Interactivity box 2 is created using just a Javascript-based ChemDoodle canvas;

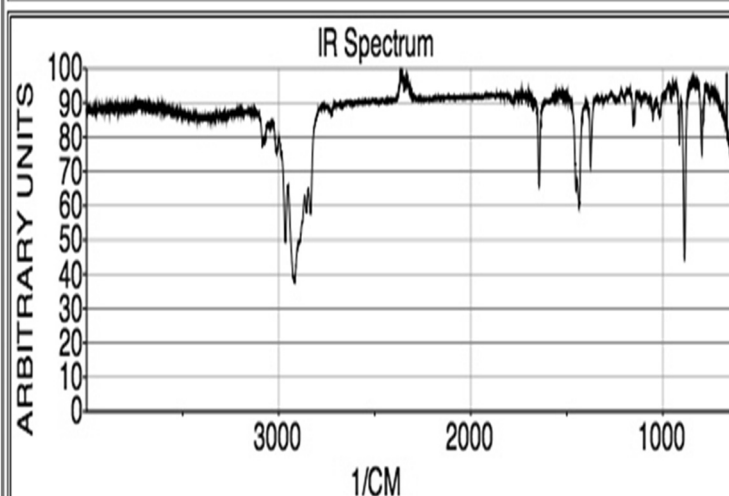
```
<script type="text/javascript" id="a1"> if(ChemDoodle.
```

Interactivity box 2.^a Data-rich molecular model rendered using ChemDoodle, illustrating one structure involved in the co-polymerisation of carbon dioxide and cyclohexene epoxide.



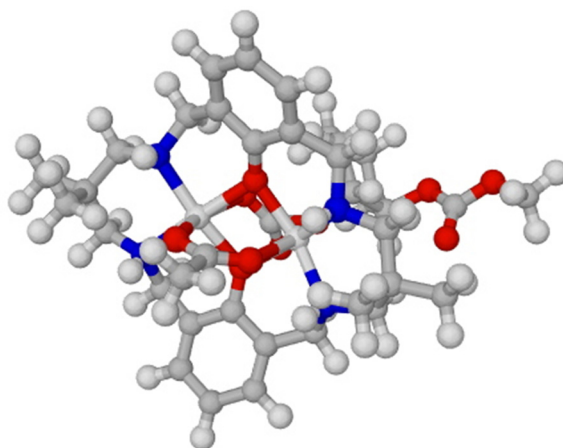
^aThe full data for the object above can be inspected at <http://dx.doi.org/10042-to-8229>

Interactivity box 3.^a Data-rich IR spectrum rendered in the datument using ChemDoodle.



^aThe spectrum can be expanded by suitable "gestures"

Interactivity box 4.^aData-rich molecular coordinate model created using GLMol,³⁶ illustrating one structure involved in the co-polymerisation of carbon dioxide and cyclohexene epoxide.



^aThe full data for the object above can be inspected at <http://dx.doi.org/10042/te-8229>

```
featureDetection.supports_webgl(){ var
transformBallAndStick1 = new ChemDoodle.
TransformCanvas3D('transformBallAndStick1', 550,
450); transformBallAndStick1.specs.
projectionWidthHeightRatio_3D = 550 / 450;
transformBallAndStick1.specs.set3DRepresentation
('Ball and Stick');
4. TransformBallAndStick1.specs.backgroundColor =
'white'; var molFile = httpGet('datument.mol'); var
molecule = ChemDoodle.readMOL(molFile,2);
transformBallAndStick1.loadMolecule(molecule); }
else{document.write('');}
</script>
```

The important variables in the above are simply the names of the data file (*e.g. datument.mol*). Other important attributes such as the size of the canvas etc. can be defined using information arrays (*e.g. Info1*). Such scripts are easily wrapped into *e.g.* HTML5 components such as *widgets* (interactive, and potentially 3D objects), which in turn can be absorbed into authoring environments (such as iBooks author) [40] as transcluded objects, a category that also includes tables, charts and image media. Whilst the average compositor of a scientific article is currently well acquainted with the latter type, familiarity with the concept of including *e.g.* a data-handling widget may well become a skill essential to authoring the science article of the future.

Perhaps the most realistic starting point might be to encourage (require?) Ph.D. theses to be prepared and examined in such enhanced formats. Certainly it is increasingly a requirement imposed by examiners to have available the data underlying the theses in digital, easily viewed form. It is also becoming accepted that theses can contain DOI resolvers to pertinent data-sets supporting the research being examined. The conversion of such material into journal articles might not then appear a challenge.

Conclusions

The ever increasing molecular complexity of modern chemistry presents interesting new challenges for how the underlying models may best be shared between scientists. A researcher should not have to use what may amount to inspired guesswork to reconstitute such a model from a journal article. Here I have taken two examples of complex molecular structures and by embedding descriptive data within this article, have created a working tool, a datument for the researcher. I noted earlier that one of the issues that needs addressing is whether the necessary tools for doing so would be accessible for the average scientific author. This particular datument was in fact written and assembled over a period of two days, although several of its components were already available (having been prepared as part of teaching notes on conformational analysis for lectures

delivered by the author). Higher order tools (such as Apple iBooks author [40]) show how some of the functionality needed could be absorbed into a simple to use tool. Another source of publishable datuments might come from the new generations of electronic laboratory notebooks in chemistry, and these are also increasingly interfacing to digital repositories.

There are also signs that after a long induction period, some publishers are starting to adopt such technologies for journal publication. But there are also dangers. For example, will a datument simply come to be treated as a rights-managed document, with both the full text and the data ardently protected by the publisher's commercial model? Will such enriched publications result in significantly more expensive journals? Will publishers allow datuments to be mined for their data by software agents [20] such as OSCAR? And can a datument be appropriately curated to ensure accessibility long into the future? These are important issues, but we must ensure that resolution includes active participation from both the authors of scientific datuments and their consumers.

Competing interests

The author declares that they have no competing interests.

Acknowledgements

I would like to especially thank Bob Hanson, Kevin Theisen and Takanori Nakane for helpful assistance with Jmol (and the associated scripts for device-sensitivity) ChemDoodle and GLmol respectively and the referees for helpful comments.

Received: 29 March 2012 Accepted: 26 September 2012

Published: 23 January 2013

References

- Bourne PE, Clark T, Dale R, de Waard A, Herman I, Hovy E, Shotton D: "Improving Future Research Communication and e-Scholarship". In *Force 11 Manifesto*. Edited by Allen BP, Birukou A, Blake JA, Bourne PE, Buckingham Shum S, Burns GAPC, Chan L, Olga C, Ciccarese P, Clark T, Czerniewicz L, Dale R, De Liddo A, De Roure D, De Waard A, Decker S, Garcia Castro A, Goble C, Gray E, Groth P, Hahn U, Herman I, Hovy EH, Kurtz MJ, Murphy F, Neylon C, Pettifer S, Rogers MW, Rosenthal DSH, Shotton D, Siren J, van de Sompel H, van den Besselaar P, Vison T: http://www.force11.org/white_paper. Accessed: 2012-07-10. (Archived by WebCite® at <http://www.webcitation.org/6933lu04w>).
- Rzepa HS: **The importance of being bonded**. *Nature Chem* 2009, **1**:510–512. doi:10.1038/nchem.373.
- Murray-Rust P, Rzepa HS: **The Next Big Thing: From Hypermedia to Datuments**. *J Digital Inf* 2004, **5**:Article 248. 2004-03-18. URL: <http://journals.tdl.org/jodi/article/view/130>.
- Murray-Rust P, Rzepa HS, Wright M: **Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content**. *New J Chem* 2001, **25**:618–634. doi:10.1039/b008780g.
- Kidd R, Harlow K: *Royal Society of Chemistry*, 2001. <http://www.rsc.org/suppdata/nj/b0/b008780g/comments.htm>. Accessed: 2012-03-27. (Archived by WebCite® at <http://www.webcitation.org/66TJyFYS>).
- Murray-Rust P, Rzepa HS: **Chemical markup Language and XML Part I. Basic principles**. *J Chem Inf Comp Sci* 1999, **39**:928–942. doi:10.1021/ci990052b.
- Shotton D, Portwin K, Klyne G, Miles A: **A "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article"**. *PLoS Comput Biol* 2009, **5**:e1000361. doi:10.1371/journal.pcbi.1000361.
- Shotton D: **Semantic publishing: the coming revolution in scientific journal publishing**. *Learned Publishing*, **22**:85–94. doi:10.1087/2009202.
- Rzepa HS: **The past, present and future of Scientific discourse**. *J Cheminformatics* 2011, **3**:46. doi:10.1186/1758-2946-3-46.
- Weinhold F, Landis CR: *Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective*. New York: Cambridge University Press; 2005. For an example of the application of the NBO technique to analysing unexpected bonding attributes, see ref 11.
- Rzepa HS: **The Nature of the Carbon-Sulfur bond in the species H-CS-OH**. *J Chem Theory Comput* 2010, **49**:97–102. doi:10.1021/ct100470g.
- Johnson ER, Keinan S, Mori-Sánchez P, Contreras-García J, Cohen AJ, Yang W: **Revealing Noncovalent Interactions**. *J Am Chem Soc* 2010, **132**:6498–6506. doi:10.1021/ja100936w.
- Contreras-García J, Yang W, Johnson ER: **Analysis of Hydrogen-Bond Interaction Potentials from the Electron Density: Integration of Noncovalent Interaction Regions**. *J Phys Chem A* 2011, **115**:12983–12990. doi:10.1021/jp204278k.
- Arbore JL, Rzepa HS, Contreras-García J, Adrio LA, Barreiro EM, Hii KKM: **Silver-catalysed enantioselective additions of O-H and N-H to allenes: a new model for stereoselectivity based on non-covalent interactions**. *Chem Euro J* 2012, **45**:6781–6795. doi:10.1021/ma300803b.
- Buchard A, Jutz F, Kember MR, Rzepa HS, Williams CK: **Experimental and Computational Investigation of the Mechanism of Carbon Dioxide/Cyclohexene Oxide Copolymerization Using A Di-zinc Catalyst**. *Macromolecules* 2012, **45**:6781–6795. doi:10.1021/ma300803b.
- Liu Z, Torrent M, Morokuma K: **Molecular Orbital Study of Zinc(II)-Catalyzed Alternating Copolymerization of Carbon Dioxide with Epoxide**. *Organometallics* 2002, **21**:1056–1071. doi:10.1021/om01110843.
- Peng Wu G, Ren WM, Luo Y, Li B, Zhang WZ, Lu XB: **Enhanced Asymmetric Induction for the Copolymerization of CO₂ and Cyclohexene Oxide with Unsymmetric Enantiopure SalenCo(III) Complexes: Synthesis of Crystalline CO₂-Based Polycarbonate**. *J Am Chem Soc* 2012, doi:10.1021/ja300667y.
- Lehenmeier MW, Bruckmeier C, Klaus S, Dengler JE, Deglmann P, Ott AK, Rieger B: **Differences in Reactivity of Epoxides in the Copolymerisation with Carbon Dioxide by Zinc-Based Catalysts: Propylene Oxide versus Cyclohexene**. *Chem Euro J* 2011, **17**:8858–8869. doi:10.1002/chem.201100578.
- See for example the author instructions for IUCr (International union of crystallography), where the preparation of a CIF file containing the relevant data is required: <http://journals.iucr.org>.
- Jessop DM, Adams SE, Willighagen EL, Hawzy L, Murray-Rust P: **OSCAR4: a flexible architecture for chemical text-mining**. *J Cheminformatics* 2011, **3**:41. doi:10.1186/1758-2946-3-41.
- Elsevier Journals: **Elsevier Journals**: <http://www.articleofthefuture.com>.
- Dove AP, Gibson VC, Marshall EL, Rzepa HS, White AJP, Williams DJ: **Synthetic, Structural, Mechanistic and Computational Studies on Single-Site β -Diketiminato Tin(II) Initiators for the Polymerization of rac-Lactide**. *J Am Chem Soc* 2006, **128**:9834–9843. doi:10.1021/ja061400a.
- Rzepa HS: **A full list of articles published with characteristics of datuments is given**: <http://www.ch.imperial.ac.uk/rzepa/blog/?p=701> Accessed: 2012-03-27. (Archived by WebCite® at <http://www.webcitation.org/66THiGP8E>).
- Downing J, Murray-Rust P, Tonge AP, Morgan P, Rzepa HS, Cotterill F, Day N, Harvey MJ: **SPECTRA: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories**. *J Chem Inf Mod* 2008, **48**:1571–1581. doi:10.1021/ci7004737. We are also exploring the use of Figshare as an open-access repository in this context; <http://figshare.com>.
- Rzepa HS, Murray-Rust P, Whitaker BJ: **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.
- Casher O, Rzepa HS: **SemanticEye: A Semantic Web Application to Rationalise and Enhance Chemical Electronic Publishing**. *J Chem Inf Mod* 2006, **46**:2396–2411. doi:10.1021/ci060139e.
- See <http://figshare.com>.
- Adams S, Murray-Rust P: **Chempond - a Web 2.0-inspired repository for physical science data**. *J Digital Information* 2012, **13**:5873. <http://journals.tdl.org/jodi/article/viewArticle/5873>.
- Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M: **Data Sharing by Scientists: Practices and Perceptions**. *PLoS One*, **6**:e21101. doi:10.1371/journal.pone.0021101. for a review. <http://www.dataone.org/>.
- Dryad. <http://datadryad.org/about>.
- Rosemann U, Sens I: *Abstracts of Papers, 242nd ACS National Meeting & Exposition*. United States: Denver, CO; CINF-24; 2011. See <http://datacite.org/> For a commentary.
- See <http://datashare.is.ed.ac.uk>.

33. See <http://swordapp.org/category/sword2/> for details of the protocols and other information.
34. Casher O, Chandramohan G, Hargreaves M, Leach C, Murray-Rust P, Sayle R, Rzepa HS, Whitaker BJ: **Hyperactive Molecules and the World-Wide-Web Information System.** *J Chem Soc, Perkin Trans* 1995, **2**:7–11. doi:10.1039/P2995000007.
35. JSON data handling, <http://www.json.org>.
36. Theisen KJ: *ChemDoodle Mobile: Leveraging mobile apps in chemistry*, *Abstracts of Papers, 243rd ACS National Meeting & Exposition*. San Diego, CA, United States: CINF-69; 2012. March 25-March 29.
37. Nakane T: *GLmol - Molecular Viewer on WebGL/JavaScript*. webglmol.sourceforge.jp.
38. Williams AJ, Shevelev S, Lang AS, Bradley JC, Theisen K: *Chemistry in the hand: The delivery of structure databases and spectroscopy gaming on mobile devices*, *Abstracts of Papers, 242nd ACS National Meeting & Exposition*. Denver, CO: United States; CINF-12; <http://onswebservice.wikispaces.com/NMR>.
39. Theisen KJ (Ed): *The process of extracting data from a Jmol object was described previously (Ref 9)*. <http://web.chemdoodle.com/tutorial/retrieving-data> and <http://web.chemdoodle.com/demos/chemical-markup-language-cml>.
40. Apple Inc: *iBooks autho*. <http://www.apple.com/ibooks-author>.

doi:10.1186/1758-2946-5-6

Cite this article as: Rzepa: Chemical datuments as scientific enablers. *Journal of Cheminformatics* 2012 **5**:6.

Publish with **ChemistryCentral** and every scientist can read your work free of charge

“Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge.”

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
<http://www.chemistrycentral.com/manuscript/>



ChemistryCentral