2009 Fifth IEEE International Conference on e-Science

CHIC – Converting Hamburgers Into Cows

Joseph A. Townsend, Jim Downing, Peter Murray-Rust
Unilever Centre for Molecular Science Informatics, University of Cambridge
Cambridge, UK
jat45@cam.ac.uk

***Abstract*— We have developed a methodology and workflow (CHIC) for the automatic semantification and structuring of legacy textual scientific documents. CHIC imports common document formats (PDF, DOCX and (X)HTML) and uses a number of toolkits to extract components and convert them into SciXML. This is sectioned into text-rich and data-rich streams and stand-off annotation (SAF) is created for each. Embedded domain specific objects can be converted into XML (Chemical Markup Language). The different workflow streams can then be recombined and typically converted into RDF (Resource Description Format).**

workflow; semantics;conversion;SAF;XML

I. Introduction

The "data deluge" has not yet occurred in many physical sciences because although they are data rich there is no semantic vehicle for interchange [1]. Many scientific reports end up as "full text papers" or theses which record the data in a number of conventional but non-semantic forms. Universal examples over physical science include: tables, graphs, inline numeric data with the addition of domain specific information objects such as mathematical equations, chemical formulae, spectra and other line diagrams.

We show here that modern tools for information processing and extraction can recover much of this information in reusable form. Examples include: indexing on non-textual data (*e.g.* numeric quantities), resubmission as input to computational chemistry and physics and further semantic enhancement through domain specific tools and ontologies (*e.g.* extraction of chemical reactions as described in [2]).

Most scientific research is communicated in a formal manner through scholarly publications and theses. Although the primary data is often carefully managed for the research group itself, it rarely gets exported to the rest of the community [3]. This problem is now recognised as a critical failure in the re-usability of scientific output, exemplified in the recent call from the JISC in the UK for infrastructure for research data management [4]. In this they show that without mandates data are rarely deposited and even with enforced mandates compliance is patchy. By contrast much of the researcher effort goes into producing the "full-text" (FT) since this is highly valued in the research metrics system.

There is a welcome realization by some publishers that the "raw data" should be made available and the common method is for publishers to allow or require deposition of "supplemental information" (SI). Even though this is often only a small part of the experiment, the data contained in the FT and SI are highly valuable, as they are normally the most important values and objects on which the proof of scientific validity rests. As such, the current literature can provide, in principle, millions of data sets a year over many disciplines.

Unlike the situation found in more data-savvy disciplines (bioscience, space, environment) where the re-use is appreciated by the community, in most sciences there is no tradition of providing data in semantic form (*e.g.* as input to computer programs, spreadsheets or relational databases). Despite the publications being "electronic" the publication metaphor is universally "e-paper" – they are crafted for human eyes, not machine input. This paper shows the possibility of converting significant amounts of current e-publication into semantic form. By "semantic form" we mean that the following requirements have been met:

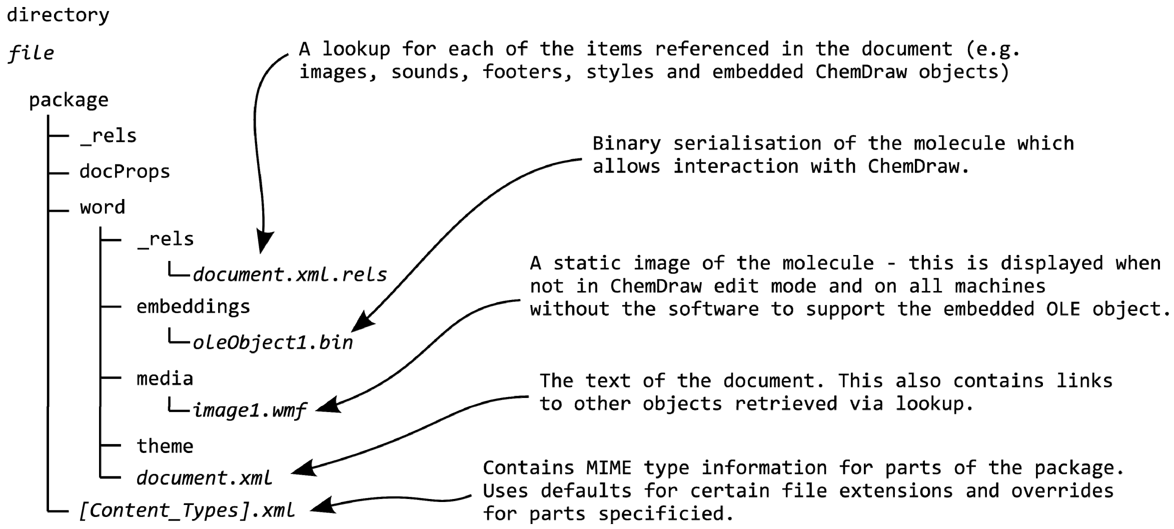
- the character encoding is clearly stated (many papers are unreadable because the various glyphs are unresolved)
- the document uses a markup technology to identify components (most commonly eXtensible Markup Language (XML), but Resource Description Format (RDF) is also common in the Semantic Web)
- the components have meaning and possibly behavior associated with them. The commonest approach is to provide ontologies in Web Ontology Language (OWL) or similar languages.

It is not essential that the complete document be converted to semantic form. It is common to identify only the domain-specific entities such as genes, proteins, organisms, diseases, stellar objects, geospatial coordinates and chemical compounds. Although the complete meaning may be unclear these are extremely useful for classification and discovery. Similarly it may be useful to process a subsection of the document, such as the `Results` or the `Methodology`.

In this paper we show that this is practicable in physical sciences and highlight the software infrastructure that is needed to promote widespread use and generation of the deluge. We first describe the common types of document used to communicate science: Portable Document Format (PDF), Microsoft® Office Word's DOC(X) and (eXtensible) HyperText Markup Language ((X)HTML). These do not generally provide structural information (sections) and we show how heuristics can be used to create these.

We note the high potential information content of domain specific (chemical) text and graphical material. We then describe the CHIC workflow in which different types of information are extracted, processed and recombined.

978-0-7695-3877-8/09 $26.00 © 2009 IEEE
DOI 10.1109/e-Science.2009.54



A lookup for each of the items referenced in the document (e.g. images, sounds, footers, styles and embedded ChemDraw objects)

Binary serialisation of the molecule which allows interaction with ChemDraw.

A static image of the molecule - this is displayed when not in ChemDraw edit mode and on all machines without the software to support the embedded OLE object.

The text of the document. This also contains links to other objects retrieved via lookup.

Contains MIME type information for parts of the package. Uses defaults for certain file extensions and overrides for parts specificied.

Figure 1. The structure of a DOCX package (simplified).

II. Turning Hamburgers into Cows

The initial and generally most intractable problem of creating semantic data is turning information in documents (often several hundred pages) into semantic form. This arises because the process of converting to "print" loses much of the explicit structure that the author uses in creating the document.

"Converting PDF to XML is a bit like converting hamburgers into cows. You may be best off printing it and then scanning the result through a decent OCR package." [5]

The feasibility and precision/recall depends critically on the nature of the legacy documents. This paper describes the application of generic heuristics and domain specific (chemical) tools to the structuring and information extract of over a hundred chemical theses from a wide variety of sources and several hundred articles.

We produced a system that allows information in chemical theses and articles to be extracted automatically with acceptable precision and in sufficient quantity to make it valuable. We note that until recently it was difficult to extract more than isolated words and phrases from scientific text. The context is important and a phrase (especially an acronym) may need knowledge of its place in the document before it can be disambiguated.

A. PDF

PDF is the de facto format for electronic document deposition in digital repositories. It is a page description format – that is, it describes the graphical appearance of page content (such as text, images and tables). As such it is optimized for human, not machine, understandability – the physical view, not the content view. Consequently the text component of PDF presents a significant problem for text-mining purposes as it is not a continuous text stream, unlike the original document. In order to provide for reliable text searching and extraction of elements for data conversion, modifications to basic PDF format have been made [6].

Tagged PDF, originally introduced by Adobe with PDF v1.4, provides the basis for applications that need the linear text stream as an input: each page of a Tagged PDF document contains the text, graphics, and images with a set of tags that bind the content elements together – in the correct reading order and, for example, the presence and meaning of significant elements such as figures, lists, tables, and so on (not unlike HTML in appearance).

PDF/A-1a (PDF/Archive Conformance Level 1a) was developed in support of ISO 19005-1:2005 [7]. Level 1a uses "Tagged PDF" and Unicode character maps to preserve the document's logical structure and content text stream in natural reading order, and has been recommended for archival purposes. However, it has been noted that:

"Adobe's Acrobat 6.0 will add tags to a PDF file, but human intelligence is still required to ensure the tagging process is performed correctly. There is little room for error in document tagging. Even seemingly small errors in document structure can easily render a file completely incomprehensible." [8]

B. XML based document formats

Microsoft developed the Open Packaging Convention (OPC) specification as a successor to its binary Microsoft Office file formats and it was handed over to ECMA International to be developed as the ECMA 376 standard, which was published in December 2006. It has also received approval as an International Organization for Standardization standard. The file-extension DOCX indicates an OPC document which should be edited using Microsoft® Office Word 2007 (as opposed to the XSLX file extension for example which are OPC documents editable using Excel). A DOCX document is actually a zip-file (the package) which contains the original text as a marked-up XML component (document.xml), with images and other embedded objects stored as separate files (Fig. 1). It is possible to convert from the old binary DOC format to DOCX although this typically is a manual process.

Open Document is an XML format developed by the Organization for the Advancement of Structured Information Standards (OASIS) consortium [9]. It is an open standard – *i.e.* freely available and implementable. The "reference" implementation is by OpenOffice.org. It was not investigated further as no appropriate documents were available in this format. Converters between these two XML-based formats are available [10]. The ability to automatically identify standard document components (for theses or articles) such as: Table of Contents, Abbreviations, Introduction, Experimental sections and References would have been an advantage, particularly as this would enable identification of standard Dublin Core metadata elements [11]. Some institutions mandate the appearance and structure of their theses, *e.g*., CalTech and MIT [12]. Such documents could be processed using institution-specific rules, similar to those described for journal processing [13].

C. HTML and XHTML

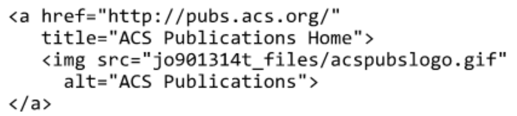
Many publishers allow users to access online articles in either PDF or HTML format. The use of HTML allows publishers to incorporate structure within a document and allow hyperlinked navigation (much of which is also provided in the PDF version). Although each publisher uses different templates for their articles there is much which is similar *e.g.* the use of sectioning elements such as `<div>` to indicate sections such as the abstract and the introduction.

The National Library of Medicine (NLM) created a set of elements and attributes to describe both the content and metadata of journal articles [14] which is being used by some publishers, *e.g.* the American Chemical Society, to outline document structure in the HTML. Hence, it could be expected that by using HTML many of the problems associated with extracting data from PDF would be avoided and more resemble that of XML documents. However, the HTML provided is frequently not well-formed (see Fig. 2).

Programs such as HTML Tidy [15] can create well-formed documents from those which are not well-formed but the process frequently results in the incorrect structure owing to ambiguity in the source document. A typical example of such incorrect restructuring is the inclusion of the entire article's text in the abstract section. This caused by an unclosed element in the `<div class='abstract'>` meaning that the correct closing `</div>` tag is not recognized.

In many ways the tidied HTML is now harder to work with than PDF because although the text stream is continuous the document now contains absolute statements of structure which may be completely wrong. At least with PDF it is expected that the task of recreating the original text will be difficult and ambiguous.

We created preprocessors for several journal templates which improve the conversion of the original HTML to well-formed documents by removing or mending certain common non-conforming elements. However, the process of creating such preprocessors is time consuming as each is specially (human) crafted. The preprocessor may ultimately have a short lifetime as publishers frequently change the document format. Declarations in the HTML may specify that the documents conform to certain definitions (such as XHTML 1.0 Transitional) and the World Wide Web Consortium provides validation services [16] to verify this conformance. No documents were encountered in the course of this work or since which have validated.



```
<a href="http://pubs.acs.org/"
   title="ACS Publications Home">
   <img src="jo901314t_files/acspubslogo.gif"
     alt="ACS Publications">
</a>
```

Figure 2. Not well-formed XML – the <img> tag is not closed.

III. Chemical Document Structure

The structure of chemical documents (specifically those reporting chemical syntheses) is usually fairly consistent, as are the names of the main sections. Typically the following main sections are found: Abstract, Introduction, Results and Discussion, Experimental, and References. A similar structure is also found in theses reporting chemical synthesis. Substructure is present in the experimental section. This consists of a section outlining all the equipment used, general synthetic methods (if any), then a series of "preparations" (see Fig. 3 for a typical example).

A preparative section contains not only a description of how to recreate the substance but also data collected about the substance to verify that the correct product was produced. The language and format of the data follow conventions and as such are semi-structured and most attractive for semantic extraction.

A. OSCAR

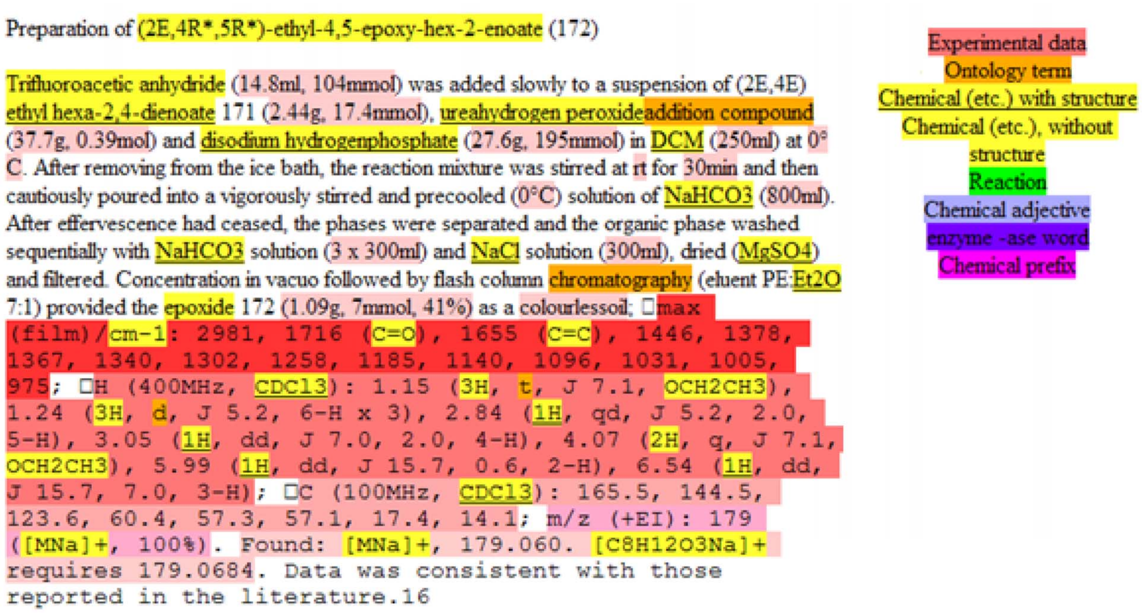
OSCAR3 is an Open Source application which identifies chemical terms and objects [17]. The following is a subset of the identified data:

- chemical names and terms – named chemical entities (NCEs)
- details of compound synthesis (quantities *etc.*)
- spectra and analytical data (characterization of chemical structure) [18]
- OSCAR3 also assigns a particular class to each NCE identified. These classes include chemical names (CM), chemical term (ONT), enzymatic name (ASE) and reaction type (RN).

Fig. 3 shows a typical preparative section as it has been annotated by OSCAR3. OSCAR3 attempts to assign structures to CM-type annotations either by lookup (from ChEBI [19] or PubChem [20]) or an internal name-to-structure converter (OPSIN).

B. SciXML and Stand-off Annotation

Regardless of input format (HTML, text, *etc.*) OSCAR3 internally uses SciXML [21] to retain the overall document structure throughout processing. Section and paragraph level formatting are retained. In some cases (full text) it is important to retain styles as these often convey implicit semantic information, in other cases (data rich) styles are omitted as they can interfere with the heavy use of regular expressions.



Preparation of (2E,4R*,5R*)-ethyl-4,5-epoxy-hex-2-enoate (172)

Trifluoroacetic anhydride (14.8ml, 104mmol) was added slowly to a suspension of (2E,4E) ethyl hexa-2,4-dienoate 171 (2.44g, 17.4mmol), ureahydrogen peroxideaddition compound (37.7g, 0.39mol) and disodium hydrogenphosphate (27.6g, 195mmol) in DCM (250ml) at 0° C. After removing from the ice bath, the reaction mixture was stirred at rt for 30min and then cautiously poured into a vigorously stirred and precooled (0°C) solution of NaHCO3 (800ml). After effervescence had ceased, the phases were separated and the organic phase washed sequentially with NaHCO3 solution (3 x 300ml) and NaCl solution (300ml), dried (MgSO4) and filtered. Concentration in vacuo followed by flash column chromatography (eluent PE:Et2O 7:1) provided the epoxide 172 (1.09g, 7mmol, 41%) as a colourlessoil; □max (film)/cm-1: 2981, 1716 (C=O), 1655 (C=C), 1446, 1378, 1367, 1340, 1302, 1258, 1185, 1140, 1096, 1031, 1005, 975; □H (400MHz, CDCl3): 1.15 (3H, t, J 7.1, OCH2CH3), 1.24 (3H, d, J 5.2, 6-H x 3), 2.84 (1H, qd, J 5.2, 2.0, 5-H), 3.05 (1H, dd, J 7.0, 2.0, 4-H), 4.07 (2H, q, J 7.1, OCH2CH3), 5.99 (1H, dd, J 15.7, 0.6, 2-H), 6.54 (1H, dd, J 15.7, 7.0, 3-H); □C (100MHz, CDCl3): 165.5, 144.5, 123.6, 60.4, 57.3, 57.1, 17.4, 14.1; m/z (+EI): 179 ([MNa]+, 100%). Found: [MNa]+, 179.060. [C8H12O3Na]+ requires 179.0684. Data was consistent with those reported in the literature.16

Experimental data
Ontology term
Chemical (etc.) with structure
Chemical (etc.), without structure
Reaction
Chemical adjective
enzyme -ase word
Chemical prefix

Figure 3. Highlighted experimental procedure created by OSCAR3.

Since several tools process the same document we retain the definitive structural markup as SciXML and add extracted information through stand-off annotation [22]. OSCAR3 output is in three files (Fig. 4):

- **SAF** – a Stand-off Annotated Format which contains all the annotations made and their associated confidence. SAF uses XPointer [23] to refer to a precise point in the document. An example of a SAF which has been further annotated is shown in Fig. 5.
- **INLINE** – a SciXML file retaining all original markup but with annotations added inline where possible (*i.e.* where the inline annotation does not cross existing elements).

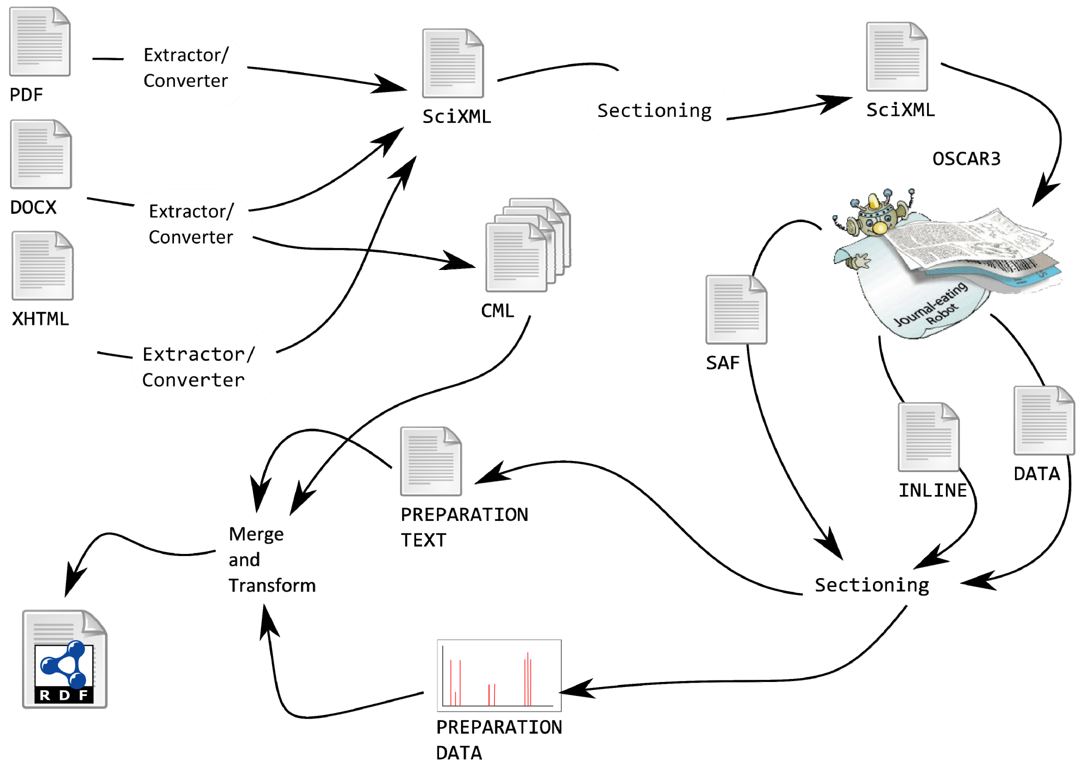


Figure 4. High-level overall workflow.

- **DATA** – a SciXML file with all the text styling elements removed (*i.e.* no bold, underline, superscript, strikethrough, italic *etc.*) but with the experimental data marked up inline.

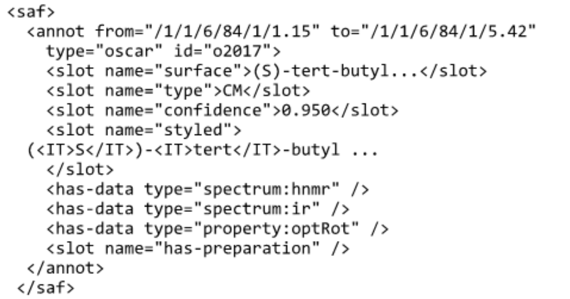
IV. Image Conversion

Traditionally images have been difficult to interpret, often because they were hand-drawn or present as bitmaps and because it was extremely hard to find their bounding boxes. The increase in computer-authoring of diagrams leads to a more tractable problem. When a vector format (such as PDF, WMF, SVG or EPS) is used we can often extract the primitives and with considerable labour reconstruct some of the chemistry. When XML-based formats are used, the bounding boxes are also discoverable.

The Windows Metafile (WMF) and the more recent Windows Enhanced Metafile (EMF) are graphics file formats on the Windows platform. Although capable of storing bitmap images they are more commonly used to store vector graphics. It is not trivial to reconstruct chemistry from vector based formats although it is easier than having to interpret pixel maps (recently the OSRA program [24] offers hope for interpreting chemical bitmaps). Common problems encountered in the conversion were:

- Is a line chemical or not?
- What is the role of text?
- Where do lines cross/intersect and is there an implied atom there?
- Where do lines end and is there an implied atom?
- What do crossing lines mean?
- What do short lines mean? Parallel lines?
- Is "2" an atom number or a multiplier?

These are hard to interpret and the heuristics depend on the precise software used to generate the graphics. Some tools have careful mitering of bonds, while others simply butt them. Crossing bonds, multiple bonds, and truncated bonds are difficult (the latter can be confused with minus signs or other obscured primitives). Nonetheless, with EMF files retrieved from OPC (DOCX) packages we have been able to achieve over 95% conversion in favourable cases. Conversion from PDF can be harder because it is difficult to determine where text and graphics are separated but as we add heuristics to the analysis of domain specific PDFs we can get very high recall from modern documents.



```
<saf>
  <annot from="/1/1/6/84/1/1.15" to="/1/1/6/84/1/5.42"
    type="oscar" id="o2017">
    <slot name="surface">(S)-tert-butyl...</slot>
    <slot name="type">CM</slot>
    <slot name="confidence">0.950</slot>
    <slot name="styled">
  (<IT>S</IT>)-<IT>tert</IT>-butyl ...
    </slot>
    <has-data type="spectrum:hnmr" />
    <has-data type="spectrum:ir" />
    <has-data type="property:optRot" />
    <slot name="has-preparation" />
  </annot>
 </saf>
```

Figure 5. An example of a Stand-off annotation file.

Chemistry benefits from having multiple independent representations of the same information. Thus a compound may be expressed as a graphic, a name or in some cases an embedded binary file (*e.g.* from the ChemDraw™ software). Our OPSIN software can now translate at least 85% of systematic chemical names. Similarly the numeric experimental data (Fig. 3) must be internally consistent with the extracted molecular formula. This allows a Rosetta-stone-like check on information extraction in that if two approaches agree then it is highly likely that the information extracted represents the author's intention. There are also other heuristics such as accepted ranges of experimental values (*e.g.*, melting point) where outliers may indicate deficiencies in the extraction procedure.

V. The CHIC Workflow

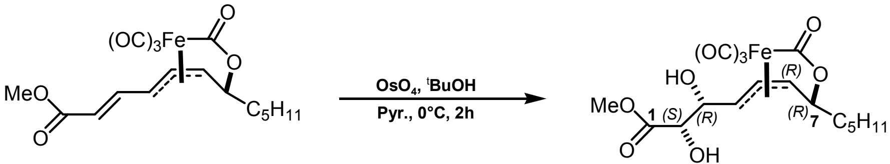
As shown in Fig. 4 the first step in the CHIC workflow is to convert all the documents into SciXML and it is emphasized that each legacy format goes through a different initial converter. This normalization allows all subsequent processors to consider only a single format.

A. PDF Input

The conversion of PDF to text-stream was achieved using the PDFBox library [25]. Although this has facilities for both extracting text and images and some support for vectors this could not be automated. In CHIC we concentrated on its text extraction capability. Bitmapped chemistry can be ignored but the graphic sections containing vectors were not interpreted and in fact introduced additional noise into the text. Fig. 7 shows the automatic translation of the text and images of Fig. 6 by PDFBox.

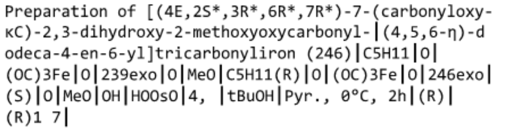
An empirical algorithm was developed to remove these sections from the output stream, to reduce the number of

Preparation of [(4*E*,2*S**,3*R**,6*R**,7*R**)-7-(carbonyloxy-κC)-2,3-dihydroxy-2-methoxyoxycarbonyl-(4,5,6-*η*)-dodeca-4-en-6-yl]tricarbonyliron (**246**)



***exo* 239** ***exo* 246**

Figure 6. Embedded chemical objects as they appeared in the Word document.



```
Preparation of [(4E,2S*,3R*,6R*,7R*)-7-(carbonyloxy-
κC)-2,3-dihydroxy-2-methoxyoxycarbonyl-|(4,5,6-η)-d
odeca-4-en-6-yl]tricarbonyliron (246)|C5H11|O|
(OC)3Fe|O|239exo|O|MeO|C5H11(R)|O|(OC)3Fe|O|246exo|
(S)|O|MeO|OH|HOOsO|4, |tBuOH|Pyr., 0°C, 2h|(R)|
(R)1 7|
```

Figure 7. Text stream extracted from the PDF. Vertical bars | indicate new lines.

false positive chemical terms.

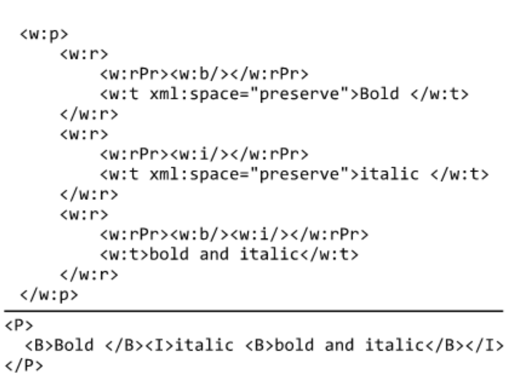
B. DOCX Input

Conversion of DOCX to SciXML was achieved by a combination of Java code and XSL stylesheets. There were two main obstacles; the presence of non UTF-8 character codes and moving from sibling style markup (used in DOCX) to inline style markup (used by SciXML). Fig. 7 shows the DOCX equivalent of the simpler SciXML. A lookup table was created which included the most common non-Latin characters in chemistry (*e.g.* Greek letters, double and triple bonds and middots).

The conversion to inline style data is non-trivial and becomes more complex the more styling elements are considered. For CHIC we have only considered the following formatting: bold, italic, superscript, subscript and underline. This particular subset was of interest because bold text often indicates the beginning of sections and is used to label molecules for quick reference, italics or underlining are often used in atom assignments in nuclear magnetic resonance and infra-red spectra, and superscripts and subscripts are used in formulae, isotopic labeling, references *etc*. Whilst other formatting is used it is markedly less prevalent.

C. Sectioning SciXML

SciXML uses `<DIV>` elements to define document structure. These are used within the main body to identify sections such as `Introduction`, `Results` and `Experimental`. A deterministic algorithm was used to find the paragraphs which indicated the beginnings of these sections. Identifying the ends of sections is much more difficult and the heuristic used was that a section was ended when a start of a new section was encountered. Identifiers (IDs) were added to the resultant SciXML. The IDs were XPointer-based pointers which uses the XML document object model tree as the primary way of defining position in the document and are unaffected by processing with OSCAR3 thereby allowing sections and paragraphs to be linked between the different output files. OSCAR3 was run on each document using a minimum cutoff value of 20% (as recommended in the documentation). The confidence value is calculated from both the quad-gram [26] score and the context.



```
<w:p>
    <w:r>
        <w:rPr><w:b/></w:rPr>
        <w:t xml:space="preserve">Bold </w:t>
    </w:r>
    <w:r>
        <w:rPr><w:i/></w:rPr>
        <w:t xml:space="preserve">italic </w:t>
    </w:r>
    <w:r>
        <w:rPr><w:b/><w:i/></w:rPr>
        <w:t>bold and italic</w:t>
    </w:r>
</w:p>
```

```
<P>
  <B>Bold </B><I>italic <B>bold and italic</B></I>
</P>
```

Figure 8. Sibling style data (top) and inline style data.

D. Sectioning Preparative Narrative

Preservation of new lines and paragraph structure from DOCX allowed reasonable identification of main sections (in SciXML). To improve the precision of the identification of preparations these were only considered to exist in the experimental section. A preparation is identified as beginning with a chemical name header followed by a (bold) number optionally in brackets (*e.g.* Fig. 3 and Fig. 6). Identifying the end of a preparative section is difficult and therefore the heuristic used was that a preparation should end when either the start of a new preparation is encountered or when the end of the experimental section is reached.

The sectioning of preparations is performed on the INLINE document using the paragraph and section information in that document and the annotations in the SAF. Using the IDs added before the OSCAR3 process allows the identification of the relevant paragraphs in the DATA document (containing identified analytical data). The INLINE preparative sections and DATA preparative sections are then extracted into separate files with unique file names.

The chemical-analytical data sections relating to each synthetic procedure in the resulting DATA file are converted to CML and merged with the extracted INLINE preparative section. Each new preparation and the associated spectral assignment data is placed in a data repository with unique URI directory-filename (*i.e.* a webserver with an associated filestore). Using the annotation of the chemical name as a reference, the SAF is updated with the location of these files, so that the derived RDF can be linked to extracted CML files.

VI. Current Developments

A number of groups have been working on heuristic and machine learning extraction of scientific information. The Penn State group has shown that machine learning methods can identify tables and diagrams with good precision. We are now combining forces with them and Southampton University in the ORECHem project [27] to develop automatic high-through put extraction of chemistry.

We have used the CHIC methodology to produce semantic and searchable linked open data from unstructured legacy science and have been able to perform computational chemistry directly on the results. The work presented in [2] used the CHIC workflow to convert the information in legacy documents into marked up documents before applying natural language processing techniques to extract further information. We are now looking at refactoring the CHIC pipeline and placing it within the Lensfield framework as described in [28]. This will allow us to maintain a record of the improvements in accuracy of the extraction processes at each stage as well as a fuller definition of how the final extracted information was obtained.

References

[1] A J. G. Hey and A E Trefethen, "The Data Deluge: An e-Science Perspective," in *Grid Computing - Making the Global Infrastructure a Reality*, F Berman, G C Fox, and A J G Hey, Eds.: Wiley and Sons, 2003, ch. 36, pp. 809-824.

[2] Lezan Hawizy, Nico Adams, Peter Murray-Rust, "Identification and visualisation of objects and their relationships in the scientific literature using natural language processing", IEEE Fifth International Conference on eScience 2009, in press.

[3] Aaron Griffiths, "The Publication of Research Data: Researcher Attitudes and Behaviour," *International Journal of Digital Curation*, vol. 4, no. 1, pp. 46-56, 2009.

[4] Joint Information Systems Committee. (2009, August) Grant Funding Call 07/09: Data Management Infrastructure : JISC. [Online]. http://www.jisc.ac.uk/fundingopportunities/funding_calls/2009/05/grant0709.aspx

[5] Michael Kay. (2009, August) xml-dev - RE: [xml-dev] How we can convert pdf data into xml?. [Online]. http://lists.xml.org/archives/xml-dev/200607/msg00509.html

[6] Frank L Walker, Marie E Gallagher, and George R Thoma. (2009, August) PDF File Migration to PDF/A: Technical Considerations. [Online]. http://archive.nlm.nih.gov/pubs/ceb2007/2007020.pdf

[7] International Organization for Standardization. (2009, August) ISO 19005-1:2005 - Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1). [Online]. http://www.iso.org/iso/catalogue_detail?csnumber=38920

[8] Duff Johnson. (2009, August) Classic Planet PDF - Learning Centre - What is tagged PDF? [Online]. http://www.planetpdf.com/mainpage.asp?webpageid=1269

[9] OASIS: Advancing the Standards for the Open Information Society. [Online]. http://www.oasis-open.org/home/index.php

[10] DIaLOGIKa, Sonata-Software, Microsoft, Novell, Clever Age, Aztecsoft. (2009, July) OpenXML/ODF Translator Add-ins for Office. [Online]. http://odf-converter.sourceforge.net/index.html

[11] (2009, July) Dublin Core Metadata Initiative. [Online]. http://dublincore.org/

[12] MIT Libraries. (2009, August) Specifications for Thesis Preparation 2008-2009. [Online]. http://libraries.mit.edu/archives/thesis-specs/

[13] I Lewin, "Using hand-crafted rules and machine learning to infer SciXML document structure," in Proceedings of the 6th UK e-science All Hands Meeting, Nottingham, 2007 [Online]. http://www.allhands.org.uk/2007/proceedings/papers/820.pdf.

[14] National Center for Biotechnology Information, U.S. National Library of Medicine. (2009, August) Journal Publishing Tag Set. [Online]. http://www.dtd.nlm.nih.gov/publishing/

[15] (2009, August) HTML Tidy Project Page. [Online]. http://tidy.sourceforge.net/

[16] World Wide Web Consortium. (2009, August) The W3C Markup Validation Service. [Online]. http://validator.w3.org/

[17] Peter Corbett and Peter Murray-Rust, "High-Throughput Identification of Chemistry in Life Science Texts," in Lecture Notes in Computer Science II, W Huisinga et al., Eds.: Springer Berlin / Heidelberg, 2006, pp. 107-118.

[18] Joseph Andrew Townsend et al., "Chemical documents: machine understanding and automated information extraction," Org. Bio. Chem., vol. 2, pp. 3294-3300, 2004.

[19] K Degtyarenko et al., "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic Acids Res.*, vol. 36, pp. D344-D350, 2008.

[20] E E Bolton, Y Wang, P A Thiessen, and S H Bryant, "PubChem: Integrated Platform of Small Molecules and Biological Activities," in Annual Reports in Computational Chemistry , David Spellmeyer and Ralph Wheeler, Eds. Amsterdam, The Netherlands: Elsevier, 2008, vol. 4, ch. 12.

[21] C J Rupp, Ann Copestake, Simone Teufel, and Benjamin Waldron, "Flexible Interfaces in the Application of Language Technology to an eScience Corpus," in Proceedings of the 4th UK E-Science All Hands Meeting, Nottingham, UK, 2006.

[22] Benjamin Waldron and Ann Copestake, "A Standoff Annotation Interface between DELPH-IN Components," in Proceedings of the fifth workshop on NLP and XML, Trento, Italy, 2006.

[23] Eric Wilde and David Lowe, Xpath, Xlink, Xpointer, and Xml: A Practical Guide to Web Hyperlinking and Transclusion. Boston, MA: Addison-Wesley Longman Publishing Co., 2002.

[24] I V Filippov and M C Nicklaus, "Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution," J. Chem. Inf. Model., vol. 49, no. 3, pp. 740-743, February 2009.

[25] The Apache Software Foundation. (2009, August) Apache PDFBox - Java PDF Library. [Online]. http://www.pdfbox.org/

[26] William B Cavnar and John M Trenkle, "N-gram-based text categorization," in Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 1994, pp. 161-169.

[27] Carl Lagoze. (2009, August) The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. [Online]. http://journal.webscience.org/112/2/websci09_submission_10.pdf

[28] Nick Day, Jim Downing, Lezan Hawizy, Nico Adams and Peter Murray-Rust, " Towards Lensfield: data management, processing and semantic publication for vernacular e-science", IEEE Fifth International Conference on eScience 2009, in press.