SPECTRa-T: Machine-Based Data Extraction and Semantic Searching of Chemistry e-Theses

Jim Downing,[†] Matt J. Harvey,[‡] Peter B. Morgan,[§] Peter Murray-Rust,[†] Henry S. Rzepa,[||] Diana C. Stewart,[†] Alan P. Tonge,[*,†] and Joe A. Townsend[*,†]

Unilever Centre for Molecular Informatics, Department of Chemistry, Lensfield Rd., Cambridge CB2 1EW, U.K., Cambridge University Library, West Rd., Cambridge CB3 9DR, U.K., and Department of Chemistry and High Performance Computing Unit, ICT, Imperial College London, Exhibition Rd., London SW7 2AZ, U.K.

Received September 25, 2009

The SPECTRa-T project has developed text-mining tools to extract named chemical entities (NCEs), such as chemical names and terms, and chemical objects (COs), e.g., experimental spectral assignments and physical chemistry properties, from electronic theses (e-theses). Although NCEs were readily identified within the two major document formats studied, only the use of structured documents enabled identification of chemical objects and their association with the relevant chemical entity (e.g., systematic chemical name). A corpus of theses was analyzed and it is shown that a high degree of semantic information can be extracted from structured documents. This integrated information has been deposited in a persistent Resource Description Framework (RDF) triple-store that allows users to conduct semantic searches. The strength and weaknesses of several document formats are reviewed.

INTRODUCTION

We are engaged in a concerted and coordinated set of projects to develop the practice and tools for creating and using chemistry in machine-readable semantic form,[1] specifically using Chemical Markup Language[2] (CML, an XML language) and Resource Description Framework[3] (RDF), the syntax developed for describing and querying data on the Semantic Web. We have argued that if chemistry is published in this way, it will be possible for machines to take over much of the routine work of discovering, analyzing, and collating chemical information.

In this paper, we investigate the potential for extracting chemistry data from electronic theses, particularly those available in Adobe Portable Document Format (PDF), the principal legacy format for digital institutional thesis preservation,[4] and Microsoft Word, the principal word-processing format for document creation. The motivation for this springs from the following synergies:

- Students are increasingly required to submit and deposit their theses in open institutional repositories that are available to everyone.
- Much of the research in chemistry departments is carried out by graduate students.
- Much of the work in the departments as a whole, and especially the supporting chemical data, is never published in detail.
- The development of structured document formats such as Microsoft Word 2007/2008, which are based on XML, makes it easier and cost-effective to extract large volumes of material automatically.

Theses represent an important untapped mine of information. The majority of the experimental data contained within a graduate thesis is not submitted for peer-reviewed publication, and it has been estimated that 80% of chemical experiments and their associated data remain unpublished.[5] There is a major effort in many countries to provide Open Access repositories of academic theses. For example, EthOSnet[6] in the UK, the DAREnet[7] Promise of Science in the Netherlands, the DART Europe portal,[8] and ADT[9] in Australia/New Zealand provide search facilities and access to many thousands of e-theses. An increasing number of universities are developing and adopting the digital repository tools to enable the exposure of funded science. Chemistry theses have the special benefit that students are generally required to provide supporting information of synthetic preparations in great detail.

By a felicitous convergence of circumstances, this pressure for openness of content coincides with the development of the technologies for representing theses in structured XML-based format. Both Microsoft Office and Open Office have exposed high-quality structured descriptions of their documents, and we argue that this is a major positive development in publishing chemistry. Indeed, it may form a natural route to a complete semantic toolkit for chemists.[10] At the same time, Semantic Web RDF technology has come of age,[11] both in the general acceptance that it works and now has stable interoperable representations and also that several manufacturers produce working RDF databases ("triple-stores"). Its application in the integration and management of bioinformatics data, with the potential to improve drug discovery, has been reported.[12]

Here we show that chemical information in theses can be extracted automatically with acceptable precision and in

* To whom correspondence should be addressed. E-mail: alan@three-d.demon.co.uk, jat45@cam.ac.uk.

[†] Unilever Centre for Molecular Informatics.

[‡] High Performance Computing Unit, ICT, Imperial College London.

[§] Cambridge University Library.

[||] Department of Chemistry, Imperial College London.

10.1021/ci9003688
Published on Web 01/20/2010

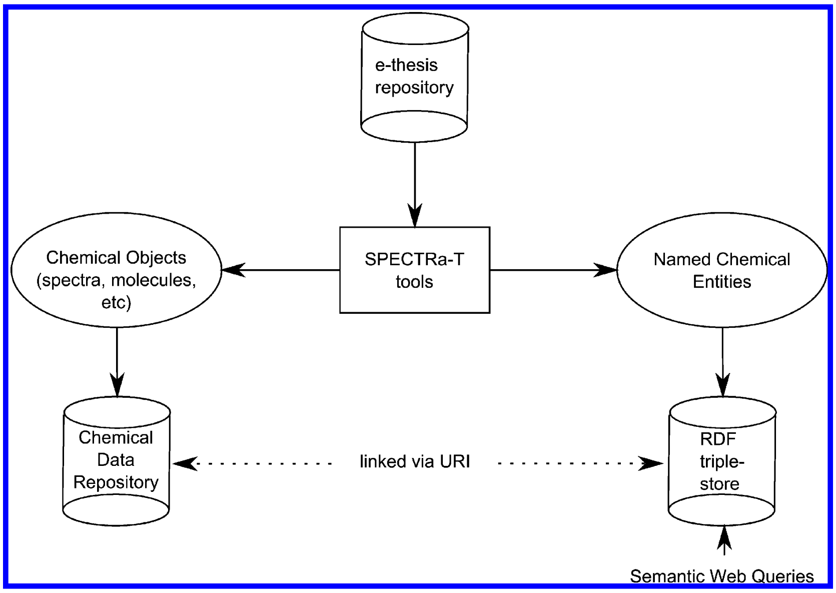


Figure 1. Overview of the SPECTRa-T data-mining architecture.

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 or term (especially an acronym) may need knowledge of its place in the document before it can be disambiguated. We also note that conventional classification and abstracting has required expensive human effort and in the work reported here we argue that machines can produce an acceptable alternative to this.

Chemical theses typically contain experimental preparative procedures, which are recipes for the chemical reaction and workup to create target molecules; spectral images and/or the associated spectral assignments; and physicochemical molecular properties. These are not routinely captured and exposed to search tools and are typically stored without being subjected to appropriate preservation techniques, with the likely irretrievable loss of data within a few years. Manual abstraction of this data, as undertaken by commercial publishers for journals and patents, is a resource intensive process and as such is not an option in an academic environment. The principal aim of the SPECTRa-T (Submission Preservation Exposure of Chemistry Teaching and Research data from Theses) project was therefore to develop text-mining tools to enable the automatic extraction of experimental data and chemical objects in high volumes from chemistry e-theses and facilitate their ingest into digital repositories, thus enabling semantic queries of the information (Figure 1).

We are aware of five common file formats currently used for thesis deposition in digital repositories:

(1) Adobe Portable Document Format (PDF)
(2) Microsoft Word (DOC/DOCX)
(3) OpenOffice (ODT)
(4) LaTeX
(5) PostScript

PDF is the de facto format for deposition of electronic text-based documents in digital repositories. It is primarily 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, readability, i.e., the physical view, not the content view. As a result, the text component of PDF presents a significant problem for text-mining purposes, as it is not a continuous text stream, unlike the originating word-processed document. In order to provide for reliable text searching and extraction of elements for data conversion, modifications to basic PDF format have been made:[13,14]

(1) 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 including, for example, the presence and meaning of significant elements such as figures, lists, tables, and so on (in a form not unlike HTML in appearance).

(2) PDF/A-1a (PDF/Archive Conformance Level 1a) was developed in support of ISO 19005−1:2005 (Document Management Standard—Electronic Document File Format for Long-Term Preservation). 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.

This new variation has been proposed as an improved basis for document preservation. However, standard versions of PDF (i.e., untagged v1.2−v1.6 which constitute the majority of electronic legacy documents stored in institutional digital repositories) cannot be reliably upgraded, as this requires Acrobat Professional and automated tagging is unlikely to produce satisfactory results. Also, it is still a page description format and any continuous text could still be broken by images, tables, or embedded objects (crucial components of chemistry e-theses). Although it is likely that PDF/A will in the future be considered by many institutions as the primary document archival format, it has been noted that its use is not a straightforward automatic process:

"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."[15]

and

"Organizations that choose to adopt PDF/A need to implement additional policies, procedures, and requirements on the processes used to generate conforming files to ensure the reliable rendering of the documents."[4]

Although our initial investigations focused on legacy PDF documents, we found that in order to facilitate better document structure analysis for chemical object (CO) identification it was necessary to use a marked up document format derived from the original word-processed thesis provided by the student. We chose to use Office Open XML, originally developed by Microsoft and now an ISO standard, rather than OpenDocument, as the former is more widely used for thesis authoring. However, our choice would not preclude the use of alternative XML standards. While LaTeX documents can be considered marked up, LaTeX it is not commonly used by synthetic organic chemists for thesis authoring; hence, we decided that such documents were out of the scope of the initial investigation.

XML Marked up Document Formats (DOCX and ODT). Microsoft has developed the Open Packaging Convention (OPC) specification[16] 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 (ISO) standard. The file-extension DOCX indicates an OPC document that should be edited using Microsoft Office Word 2007. The DOCX document is actually a zipfile (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. Open Document is an XML format developed by the Organization for the Advancement of Structured Information Standards (OASIS) consortium.[17] It is an open standard, i.e., freely available and implementable. It was not investigated further as no appropriate documents were available in this format. Converters between these two XML formats are available.

The ability to automatically identify and markup standard thesis document components, such as title page, table of contents, abbreviations, introduction and discussion, experimental sections, references, would have been an advantage, particularly as this would enable identification of standard Dublin Core metadata elements.[18] Some US institutions mandate the appearance and structure of their theses, e.g., California Institute of Technology (CalTech)[19] and Massachusetts Institute of Technology (MIT).[20] Such documents could potentially be processed using institution-specific rules, similar to those described for journal processing.[21] However, the observed (largely undocumented) format variability that arises between different institutions meant that development of machine-based analysis was beyond our resources.

The corpus of Ph.D. theses used in this work was composed of 34 in PDF and 10 in DOCX. Only "born digital" PDF theses (i.e., those derived electronically from the original word processed documents) were used; those with text streams created by optical character recognition (OCR) from scanned page images were omitted because of the difficulty in ensuring character identification in mixed English and Greek alphabets.[22] The PDF theses came from online digital repositories at CalTech (28), the Universities of Cambridge (1), St. Andrews (2), and Stirling (2); the DOCX theses were from the University of Cambridge (10). These theses cover the general range of chemistry domains defined in the Library of Congress classification (general, organic, biochemistry, physical, theoretical, photochemistry, and crystallography).[23] However, those subsequently chosen for specific analysis were restricted to synthetic organic chemistry. Synthetic procedures for organic molecules are presented in a fairly consistent format within both original theses and derived peer-reviewed publications: compound title, followed by a preparative recipe and analytical/spectroscopic data. This allows for potential machine-based identification of chemical objects within these documents on a large scale. The DOCX theses from the Chemistry Department at Cambridge are open and are available on the departmental Web site.[24]

METHODOLOGY

The development of a number of text-mining applications that enable the automatic identification and characterization of chemical terms within scientific documents has been described. These include the capture of chemical names within patents,[25,26] biomedical terms in document abstracts,[27] the current CheTA project[28] and Project Prospect.[29] OSCAR3 (Open Source Chemistry Analysis Routines) is an Open Source application that uses natural language processing to identify chemical terms and objects.[30] OSCAR3 analyzes chemistry texts and identifies chemical information contained within, including

- chemical names and terms, also known as named chemical entities (NCE's)
- details of compound synthesis
- spectra and analytical data (characterization of chemical structure)[31]

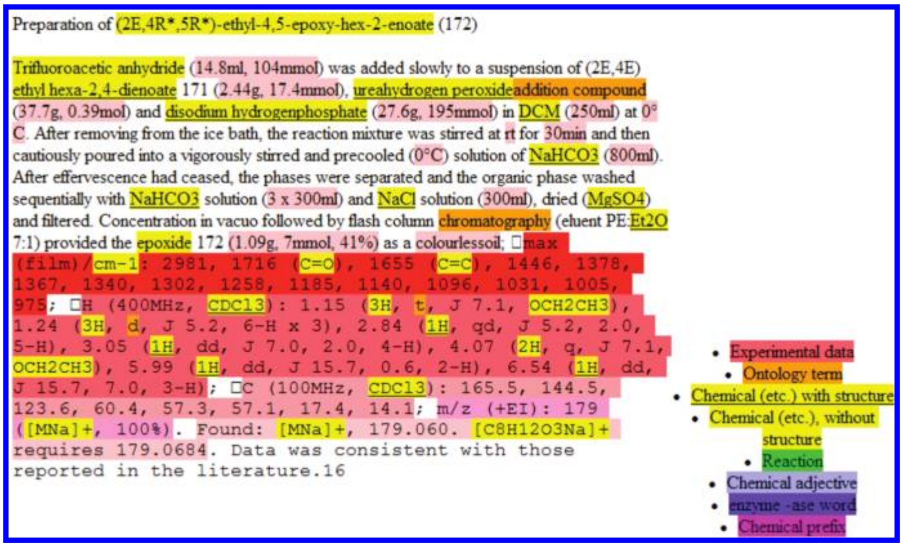
SciXML[32] is the preferred OSCAR3 input format (although plain text may be used) and this richer markup gives better opportunity for the association of chemical objects with the appropriate chemical entity. The conversion of all document formats to SciXML forms the first step in the SPECTRa-T workflow and allows all subsequent tools[33] to process a single format that first tokenize the input and identify chemical names using

- an internal lexicon (primarily based upon the ChEBI dictionary);[34] further names have also been manually curated as have the list of stop words[35]
- naïve-Bayesian rules
- a set of regular expressions to recognize chemical data and formulas

OSCAR3 breaks the captured named entities (annotations) into several classes, including

CM, chemical names
ONT, chemical (ontology) term
ASE, enzymatic name
RN, reaction type

An example of the highlighted OSCAR3 processing markup upon a sample synthetic preparative procedure is shown in Figure 2. The chemical type (CM) annotations were the primary link for the identification of chemical objects, and OSCAR3 attempts to assign structures to these either by lookup (from ChEBI or PubChem[36]) or an internal name-to-structure converter library (OPSIN). The OSCAR3 library output is contained in three files:



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 2. Highlighted experimental procedure created by OSCAR3.

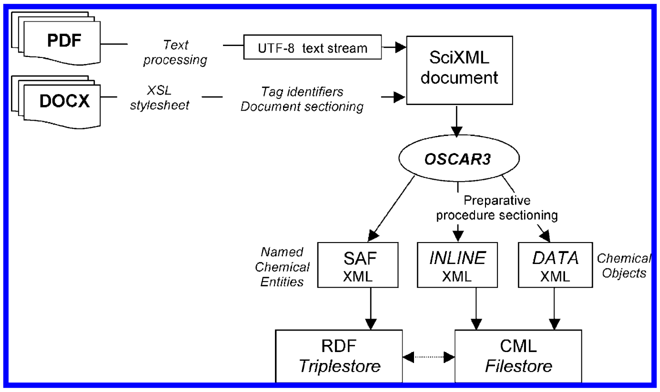


Figure 3. The overall flow of data in SPECTRa-T.

(1) SAF, a Stand-off Annotated Format XML[37] that contains all the annotations made and their associated confidences;

(2) INLINE, a SciXML file retaining all the original markup but with annotations added inline where possible (i.e., where the inline annotation does not cross existing elements);

(3) 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.

To increase the precision of the identification of experimental data, it was decided to only attempt to find it in those sections labeled as Experimental. The overall flow of data in SPECTRa-T is shown in Figure 3. Once the theses are converted to SciXML they follow the same path, but it was found that only those available in DOCX allowed sufficient sectioning for synthetic preparations and associated data to be identified.

CONVERSION OF PDF TO SCIXML

As noted previously, PDF is not ideally suited to text-mining. We identified at least five technical problem areas associated with this format:

(1) irregular word order

(2) improper line-breaks, i.e., loss of continuous text and difficult-to-identify paragraphs

(3) loss of subscripts and superscripts

(4) nonprinting characters

(5) erroneous character assignment from OCR

As PDF is a page description format, the continuous text stream that is found in the original word-processed document is no longer available and is cut, by the inclusion of additional linefeeds, into multiple individual lines. Systematic chemical names are frequently longer than a single line and the separated text strings cannot be rejoined in the native text output; as a result OSCAR3 cannot correctly identify the full chemical name. It is no longer possible to identify paragraph and section endings with any confidence; therefore, preparative sections (and the associated chemical objects) cannot be reliably identified. We therefore developed software to process the PDF text-stream in order to make it suitable for OSCAR3 analysis. The steps involved are

(1) conversion of PDF to Unicode text using PDFBox library[38]

(2) removal of unwanted characters (linebreaks, nulls), recreating much of the original text stream

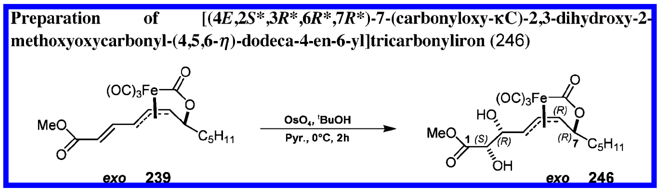
(3) removal of disconnected text derived from breakup of figures and tables

The following modifications to the PDF text-streams were made.

Linefeeds and Hyphen Fix and Nonprinting Characters. Conversion of continuous screen-wrapped text derived from an original word-processed document (e.g., from MS Word) into PDF requires the inclusion of additional newline breaks at the end of every displayed line. As systematic chemical names are frequently long, they often overlap onto the following line. For example, examine the preparative section heading

Preparation of [(4*E*,2*SR*,3*SR*)-2-(carbonyloxy-*κC*)-6-[(2′-methoxyethoxy)-methyloxy]-(3,4,5-η)-hex-4-en-3-yl]tricarbonyliron (**175**).

If this is not captured and corrected, OSCAR3 would be unable to correctly identify the full chemical name. Note that the ability to identify a string as a chemical does not mean that it is also able to convert the name to a structure. Our example of a preparative heading shows the chemical name split over two lines in the original thesis and, for the raw PDF-derived text, OSCAR3 would identify two independent chemical entities



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)

Figure 4. Original Microsoft Word document: embedded chemical objects.

(1) [(4E,2SR,3SR)-2-(carbonyloxy-*κ*C)-6-[(2′-methoxyethoxy)-

(2) methyloxy]-(3,4,5-*η*)-hex-4-en-3-yl]tricarbonyliron

Conversion back into a text stream suitable for analysis by OSCAR3 requires that the occurrence of hyphens coupled with a following linefeed character is trapped and the linefeed character removed. It is possible that use of PDF/A may avoid the majority of these linefeed problems. However, we were unable to source any examples deposited in a digital repository.

Unfortunately, the profusion of remaining linefeeds still overwhelms much of the structuring information, resulting in an inability to identify the start of sections or paragraphs using machine-based techniques. There are also a number of nonprinting characters (e.g., null values) that must be trapped and removed if text processing is to proceed correctly. This was performed by regular expression replacement of values with either an empty string or a single space character.

Removal of Text Input Derived from Chemical Structures and Tables. Chemical structures and reaction schemes shown in theses are typically drawn using commercial desktop drawing packages (e.g., ChemDraw[39] and ISIS/Draw[40]), and the resulting chemical objects (which retain the underlying chemical connection tables) are embedded into the word-processed document. Figure 4 illustrates how an embedded chemical object appears in a Word document. However, on conversion to PDF format, these chemical objects are destroyed and replaced by text, which identifies heteroatoms (i.e., atoms which are not hydrogen or carbon), atom identifiers, and a collection of disconnected lines that represent the chemical bonds that join them. The latter are not text and are lost from the text stream, while any heteroatom or substituent label is transformed into a disconnected fragment of text with no chemical context. Figure 5 shows the disconnected fragments produced in the PDF text stream for the same chemical object.

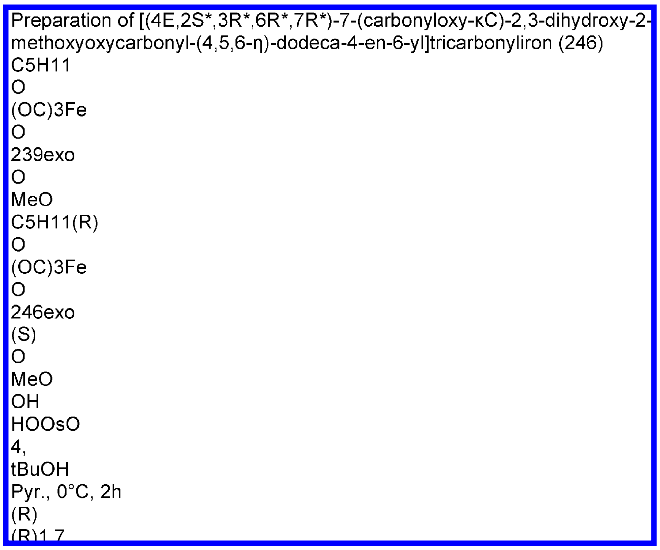
An empirical algorithm[41] was developed to remove these sections from the output stream, as these added to the number of false positive chemical terms identified by OSCAR3. Limited sections of the PDF text were erroneously identified by this technique: the title page, abbreviations sections, and those containing large sections of tabulated experimental values (e.g., crystallographic data). Subsequent analysis showed that these sections only contributed about 10% of the total number of positives identified.

Whitespace in Stereochemical Descriptors. OSCAR3 tokenizes on whitespace; encountering whitespace (a space, tab or newline) indicates that the end of the previous token (word) has been reached. In some cases (such as the term "dimethyl ester"), OSCAR3 is capable of recognizing that both words "dimethyl" and "ester" are chemical, and because certain chemical terms are known to be "connectors", in this case ester components, OSCAR3 looks to see if it can combine the two chemical terms to a single one, i.e., "dimethyl ester". A correcting regular expression was applied to systematic chemical names to correct a common syntax error introduced into the stereochemical descriptors where additional whitespace is incorrectly included by some authors. This whitespace must be removed:

$$(2E,\ 4R^*,\ 5S^*) \rightarrow (2E{,}4R^*{,}5S^*)$$

If this is not done, OSCAR3 will break the entity name at these whitespace points. Multiple whitespaces are currently not captured and this has continued to give rise to processing problems. This is a good example of how SPECTRa-T tools could be used as part of a thesis checker for chemistry. It should be noted that being able to correctly identify chemical names does not mean that OPSIN can construct the connection table.[42]

After these modifications had been made, the resulting PDF text-stream was wrapped in a set of top level SciXML tags. No further document structure analysis was attempted and following subsequent processing by OSCAR3 only the SAF XML file is populated.



Preparation of [(4E,2S*,3R*,6R*,7R*)-7-(carbonyloxy-κC)-2,3-dihydroxy-2-
methoxyoxycarbonyl-(4,5,6-η)-dodeca-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 5. Processed PDF document: resultant text output stream.

Conversion of DOCX to SciXML. Original thesis documents prepared in Microsoft Word 97−2003 were converted to DOCX format for this study by manual conversion in Word 2007. The OPC specifies that the main text of the document (i.e., not headers and footers) is contained in document.xml under the main directory of the unzipped DOCX file. A combination of XSL stylesheets and Java classes convert this into SciXML. There are two major obstacles to the conversion:

(1) non-UTF-8 character codes

(2) conversion from standoff style markup to inline style markup

The first of these was caused by the use of non-Latin characters (such as Greek) and required the creation of a lookup table that allowed the replacement character to be determined from the original character code. This is an ongoing project; the most common non-Latin characters encountered in chemistry theses, such as the Greek letters, double and triple bonds, and joining dots have been encoded, but the table is not exhaustive.

SciXML uses mixed content, that is, text elements may be children of styling elements, for instance, <B>**bold text**</B>, whereas the DOCX format separates the text content from the markup. The conversion between these representations is nontrivial, and we currently process the following styling information:

- bold
- italic
- superscript
- subscript
- underline

It is common for the assignments in NMR spectra to be shown using italics (as are stereochemical indicators in chemical names) or underlined text. Bold text often indicates the beginning of sections. Subscripts are used in formulas and superscripts for formulas and references. While other depictions do occur, these were found to be the most common and tractable.

Sectioning SciXML and OSCAR3. SciXML allows documents to be split into <DIV> sections. These are used within the main body to identify document sections such as introduction, results, and experimental. A very naïve deterministic algorithm (slightly modified from that present in OSCAR3) was used to find the paragraphs that indicated the beginnings of these sections. Identifying the ends of sections was much more difficult, and the heuristic used was that one section was ended when the start of a new section was encountered.

Once the SciXML had been through the sectioning process, an ID attribute was added to each element. The values of these IDs were nested XPoint-based pointers,[43] which use the XML DOM tree as the primary way of locating points in the document and are unaffected by processing with OSCAR3, thereby allowing sections and paragraphs to be linked between the different output files. XPoint pointers are series of integers separated by slashes (/). Each integer, *n*, locates the *n*th child element of the previously located element. Thus, the pointer "/1" refers to the root element of the document, while "/1/2/4" selects the fourth child element of the second child element of the root element.

OSCAR3 was run on each document using a default confidence cutoff value of 20%. This means that any annotation with a confidence less than 20% would not be included in the output. The confidence value is calculated from both the *n*-gram score and the context. The value used is the one recommended by the authors in the documentation.[44] As indicated previously, this resulted in three XML documents: SAF, INLINE, and DATA.

Identification of Preparative Sections. The preservation of new lines and paragraph structure in the experimental sections of DOCX documents enabled the reasonable identification of sections and hence preparative sections. It was felt that because we would be using heuristics to identify preparative sections, we would only look for them in the experimental section to reduce the number of false positives reported. Each new preparation is identified as beginning with a chemical name header followed by a number (optionally in brackets) (e.g., as shown in Figures 3 and 4). As before, identifying the end of a preparation 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 correct identification of preparative sections is vital to allow the association of NCEs to COs. The association implicitly assumes that all the data in a preparative section with a single compound in the title relate to the title compound. It is noted that this is not always true; very occasionally authors report a preparation for a compound and the accompanying data relates to a mixture of products rather than the single product intended. The ability to correctly associate extracted analytical data with NCEs from reports of this type and preparative sections with multiple NCEs in the title were out of scope for this project; however, the association of the preparation itself (i.e., the free text describing the recipe) with multiple NCEs was within our scope.

Extraction of Preparative Sections and Analytical Data. The identification of preparative sections 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, we were able to identify the relevant paragraphs in the DATA document (which contains the identified analytical data). The INLINE preparative sections and DATA preparative sections are then extracted into separate files with unique file names.

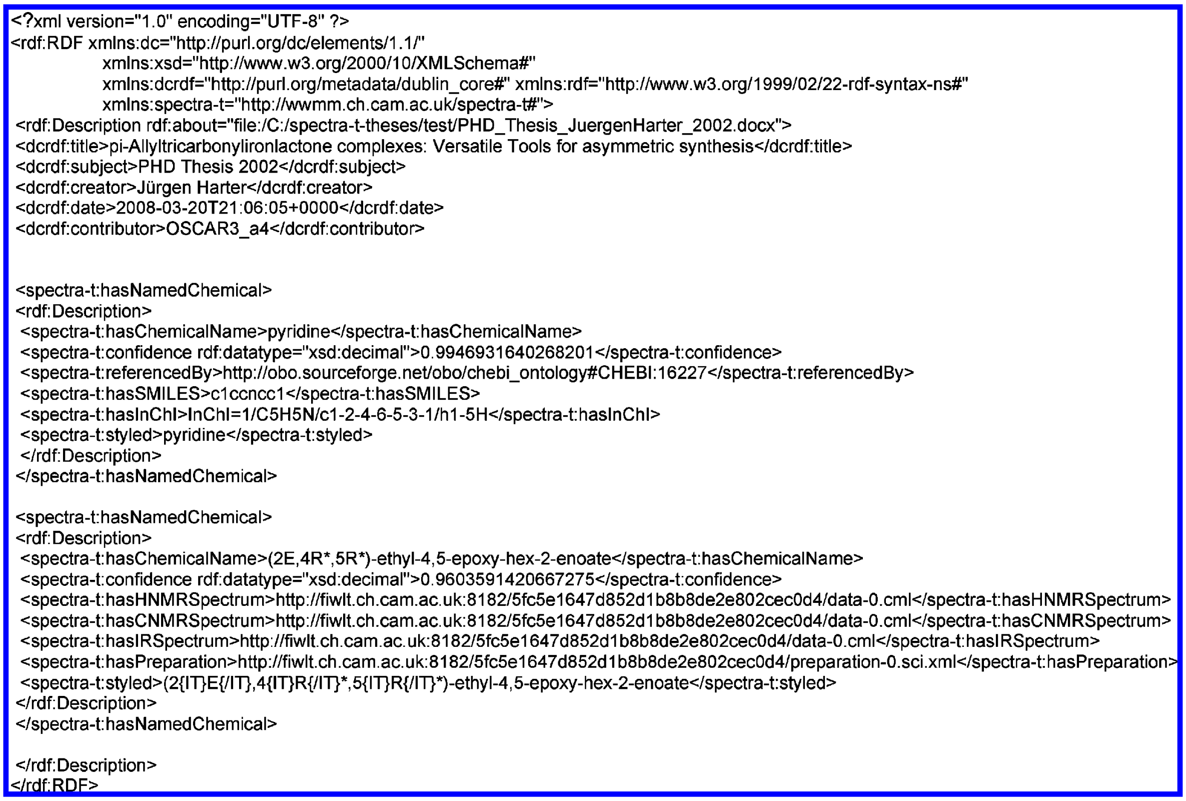
The analytical data sections relating to each synthetic procedure in the resulting DATA file are converted to CML and merged with the extracted INLINE preparative section. Each new preparation and the associated spectral assignment data is placed in a data repository with unique Uniform Resource Identifier (URI) directory-filename (i.e., a web-server 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.

DATA STORAGE AND ACCESS

We have extracted data in an XML form that can be converted for use by Semantic Web tools. In 2000, Tim Berners-Lee of the W3C consortium (the World Wide Web's governing body) presented his vision of the Semantic Web,[45] a next-generation Internet where the ability of computers to make "intelligent" deductions and decisions from machine-readable data would be possible, "creating a web that can be interpreted by machines". The current Web model, using pages marked up in HTML, is limited in providing methods in which the huge amount of unstructured data stored in static web pages may be usefully searched and indexed.

This new Semantic Web is about structured data, marked up with useful tagged elements, and the creation of relationships between these data. This should allow data reuse across application and enterprise boundaries and hence data capable of being processed automatically by machines as well as humans. The implication is that "inference" procedures can be used to generate new relationships based on the data and on some additional information in the form of an ontology or a set of rules. The components (technologies and standards) required to create such an Internet are being realized, principally

- XML—eXtensible Markup Language—a meta-language that allows people to define context-specific tags and the structural relationships between them.



```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
            xmlns:xsd="http://www.w3.org/2000/10/XMLSchema#"
            xmlns:dcrdf="http://purl.org/metadata/dublin_core#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:spectra-t="http://wwmm.ch.cam.ac.uk/spectra-t#">
 <rdf:Description rdf:about="file:/C:/spectra-t-theses/test/PHD_Thesis_JuergenHarter_2002.docx">
 <dcrdf:title>pi-Allyltricarbonylironlactone complexes: Versatile Tools for asymmetric synthesis</dcrdf:title>
 <dcrdf:subject>PHD Thesis 2002</dcrdf:subject>
 <dcrdf:creator>Jürgen Harter</dcrdf:creator>
 <dcrdf:date>2008-03-20T21:06:05+0000</dcrdf:date>
 <dcrdf:contributor>OSCAR3_a4</dcrdf:contributor>


 <spectra-t:hasNamedChemical>
 <rdf:Description>
  <spectra-t:hasChemicalName>pyridine</spectra-t:hasChemicalName>
  <spectra-t:confidence rdf:datatype="xsd:decimal">0.9946931640268201</spectra-t:confidence>
  <spectra-t:referencedBy>http://obo.sourceforge.net/obo/chebi_ontology#CHEBI:16227</spectra-t:referencedBy>
  <spectra-t:hasSMILES>c1ccncc1</spectra-t:hasSMILES>
  <spectra-t:hasInChI>InChI=1/C5H5N/c1-2-4-6-5-3-1/h1-5H</spectra-t:hasInChI>
  <spectra-t:styled>pyridine</spectra-t:styled>
  </rdf:Description>
 </spectra-t:hasNamedChemical>

 <spectra-t:hasNamedChemical>
 <rdf:Description>
  <spectra-t:hasChemicalName>(2E,4R*,5R*)-ethyl-4,5-epoxy-hex-2-enoate</spectra-t:hasChemicalName>
  <spectra-t:confidence rdf:datatype="xsd:decimal">0.9603591420667275</spectra-t:confidence>
  <spectra-t:hasHNMRSpectrum>http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852d1b8b8de2e802cec0d4/data-0.cml</spectra-t:hasHNMRSpectrum>
  <spectra-t:hasCNMRSpectrum>http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852d1b8b8de2e802cec0d4/data-0.cml</spectra-t:hasCNMRSpectrum>
  <spectra-t:hasIRSpectrum>http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852d1b8b8de2e802cec0d4/data-0.cml</spectra-t:hasIRSpectrum>
  <spectra-t:hasPreparation>http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852d1b8b8de2e802cec0d4/preparation-0.sci.xml</spectra-t:hasPreparation>
  <spectra-t:styled>(2{IT}E{/IT},4{IT}R{/IT}*,5{IT}R{/IT}*)-ethyl-4,5-epoxy-hex-2-enoate</spectra-t:styled>
 </rdf:Description>
 </spectra-t:hasNamedChemical>

 </rdf:Description>
</rdf:RDF>
```

Figure 6. Sample RDF XML created from a DOCX e-thesis showing marked-up chemical entities from preparative procedures and the associated analytical data files.

- RDF—Resource Description Framework[3]—a framework for data and metadata description and exchange. It is based upon the idea of making statements about resources in the form of a subject−predicate−object (or a resource−property−value) expression (called a *triple* in RDF terminology). The value of one property can in turn be used as the resource for another.
- OWL[46] and SKOS[47]—Web Ontology Language and Simple Knowledge Organization System—are machine-processable languages that can be used to create domain-area data models, which describe relationships between, and properties (attributes) of, objects within that domain.

Although some efforts are being made to develop and apply these components in the chemical area, their adoption is as yet far from being community-wide.[48] Nevertheless, Chemical Markup Language (CML) is now a mainstream scientific XML language, and its application in the fields of chemical structure, spectroscopy, crystallography, and polymers has been reported.[49] The development of RDF vocabularies has been reported in studies of chemical reactions[50] and aspects of chemistry publishing.[51] The major chemical ontology ChEBI (Chemical Entities of Biological Interest), developed at the European Bioinformatics Institute (EBI), assigns a hierarchy of chemical and biomedical relationships and properties to small molecules of biological interest.

RDF Model. RDF extends the linking structure of the Web by using URIs to name the relationship between things as well as the two ends (nodes) of the link; this is usually referred to as a "triple". For example, take the sentence in English: "Jürgen Harter is the author of thesis entitled π-Allyltricarbonyliron lactone complexes versatile tools for asymmetric synthesis". This can be represented in RDF as two statements, each containing three elements

Statement 1
subject (resource)—
http://wwmm.ch.cam.ac.uk/spectra_t/jh_thesis.pdf
predicate (property)—
http://purl.org/dc/elements/1.1/creator
object (value)—"Jürgen Harter"
Statement 2
subject (resource)—
http://wwmm.ch.cam.ac.uk/spectra_t/jh_thesis.pdf
predicate (property)—
http://purl.org/dc/elements/1.1/title
object (value)—"π-Allyltricarbonyliron lactone complexes"

The subject of a triple is the URI identifying the described resource. The object can either be a simple *literal value* (a text string, number, or date that describes the property) or the URI of another resource that is somehow related to the subject. The predicate, in the middle, indicates what kind of relation exists between subject and object. Literal values may only be used to describe an object, and such an object may not therefore be used as the subject of another RDF statement. The resulting data model enables you to make links between data from different sources.

The SAF created by OSCAR3 is converted to RDF XML, with the appropriate resource and property assignments, using XSL stylesheets. An example of the derived XML is shown in Figure 6, and Figure 7 shows some of this derived RDF as a directed graph. The "spectra-t" namespace is defined as xmlns:spectra-t = http://wwmm.ch.cam.ac.uk/spectra-t#. Note the presence of the blank node genid:A5803 in the graph; this indicates a node without a name, because one is either not wanted or needed. Instead, it is assigned a local identifier ID. Multiple examples of the hasNamedChemical predicate will occur, each assigned a unique ID. The RDF data model

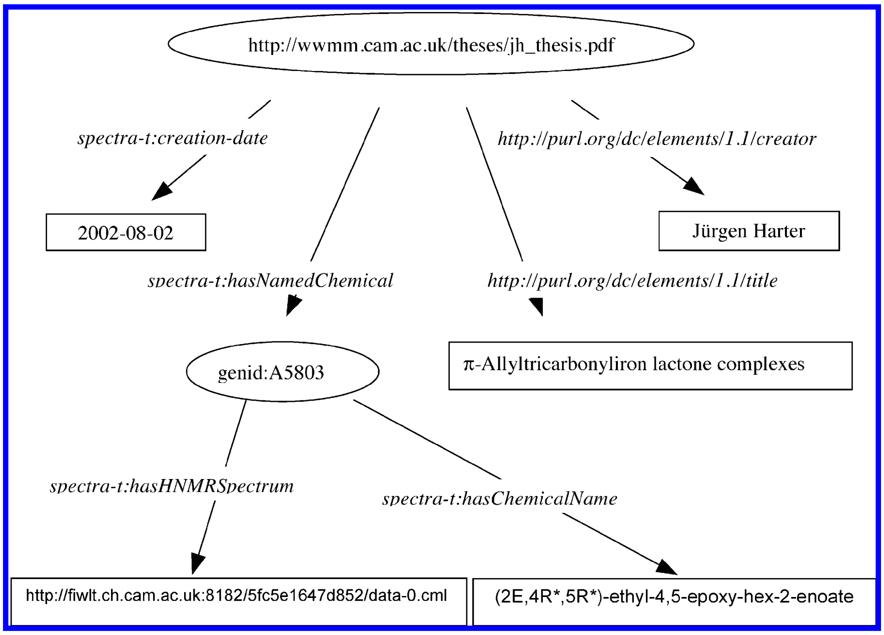
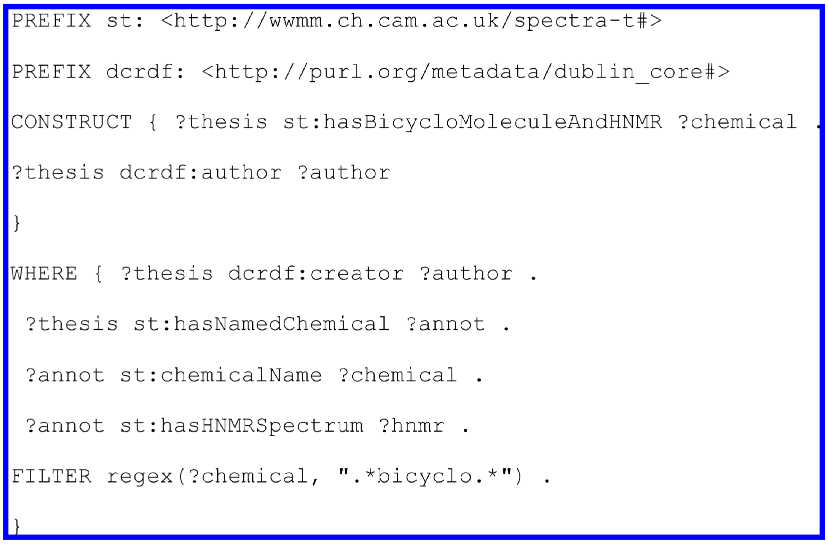

http://wwmm.cam.ac.uk/theses/jh_thesis.pdf
spectra-t:creation-date
http://purl.org/dc/elements/1.1/creator
2002-08-02
Jürgen Harter
spectra-t:hasNamedChemical
http://purl.org/dc/elements/1.1/title
genid:A5803
π-Allyltricarbonyliron lactone complexes
spectra-t:hasHNMRSpectrum
spectra-t:hasChemicalName
http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852/data-0.cml
(2E,4R*,5R*)-ethyl-4,5-epoxy-hex-2-enoate

Figure 7. A graph of the RDF derived from the SAF and linked data.



```
PREFIX st: <http://wwmm.ch.cam.ac.uk/spectra-t#>

PREFIX dcrdf: <http://purl.org/metadata/dublin_core#>

CONSTRUCT { ?thesis st:hasBicycloMoleculeAndHNMR ?chemical .

?thesis dcrdf:author ?author

}

WHERE { ?thesis dcrdf:creator ?author .

 ?thesis st:hasNamedChemical ?annot .

 ?annot st:chemicalName ?chemical .

 ?annot st:hasHNMRSpectrum ?hnmr .

FILTER regex(?chemical, ".*bicyclo.*") .

}
```

Figure 8. An example SPARQL query.

implemented does not use Containers (e.g., rdf:Bag) for collecting similar entity types, as it was thought that this would lead to increased complexity of querying through the introduction of additional blank nodes in the graph model.

RDF data may be queried using SPARQL,[52] which defines queries in terms of graph patterns that are matched against the directed graph representing the RDF data. An example query is shown in Figure 8.

Strings beginning with a question mark are dummy variables that allow constraints to be put on the search query and the returned hits. The CONSTRUCT keyword tells the query engine to create and return a new user-defined set of triples. The WHERE keyword is used to define what values the dummy variables take.

When this query is run over the data in Figure 6, the dummy variables take the following values:

?annot = genid:A5803
?thesis = http://wwmm.cam.ac.uk/theses/jh_thesis.pdf
?author = “Jürgen Harter”
?chemical = (2E,4R*,5R*)-ethyl-4,5-epoxy-hex-2-enoate
?hnmr = http://fiwlt.ch.cam.ac.uk:8182/5fc5e1647d852/data-0.cml

The presence of the *?annot st:hasHNMRSpectrum ?hnmr* restriction on the query means that there must be a proton NMR spectrum associated with the chemical name, even though its location is not explicitly returned in the CONSTRUCT. The FILTER keyword allows additional restrictions to be put on the values of the dummy variable; in this example, that somewhere in the chemical name there must be the string “bicyclo”. This restriction would prevent the query matching the data shown in Figure 6.

RESULTS AND DISCUSSION

Data extracted from synthetic chemical procedures are likely to be the most valuable asset extracted by our tools from e-theses. Only ca. 5% of identified nonduplicated NCEs are actually systematic chemical names associated directly with synthetic procedure (e.g., as in the preparative titles in

Figures 2 and 4). We have shown that semantics can be extracted from theses (especially those using XML based formats); however, there are several issues that deserve to be highlighted.

Indexed Nomenclature. OSCAR3 fails to correctly identify some chemical names that use the indexed form of nomenclature, as used by Chemical Abstracts. The indexed form puts the root entity first followed by additional separate words with whitespace, for example: (*R*)-2-(5-methylfuran-2-yl)-4-oxobutyric acid methyl ester rather than this typical alternative systematic name: methyl (*R*)-2-(5-methylfuran-2-yl)-4-oxobutyrate.

The OSCAR3 parser is probabilistic, which makes it difficult (or impossible) to predict which chemical names will be incorrectly identified; however, it has been anecdotally reported that the indexed form is more susceptible to misidentification. The current version of OSCAR splits the first example into two NCEs:

(1) (*R*)-2-(5-methylfuran-2-yl)-4-oxo

(2) butyric acid methyl ester

Whether or not these should be interpreted as false positives or true positives is often a matter of personal opinion, although either way, OSCAR3 has not been able to make the association between the two fragments.

False Positives. As well as capturing chemical names and terms as CM- and ONT-type annotations, OSCAR3 is also identifying

- molecular/empirical formulas
- functional group/solvent/reagent abbreviations
- spectral assignments (atoms or functional groups/peak multiplicities)

For example, the following identified entities (typically strings of 3−8 characters) generally fit into these classes and make up approximately 30% of the identified hits.

- HEP-G2 AD 2.5 h −40 °C ∼50 °C 983G 1.6 M 0.5H 4H C(1) J13CP 6A 4B
- COOR R2 R3 CH2O
- FAB
- H-1 11-H C-7 14-C g-1 2-C′-(CH2 3-Ha 7-OH 1,7-syn aryl-C Me(CH2)4
- C═C HC═CH C−H C(1)−C(5) 1H-13C
- MW=122.7 bp=89 °C *p* = 0.5 mmHgT = 120 °C
- −[MNa+] [C13H15O2ClNa] + [M−3CO]+
- 1 h ddd dt ca.4 d.e.'s sl0208

These false positives are mainly a result of the nondeterministic *n*-gram parsing used. While it is possible that these may be reduced by increasing the cutoff confidence, this would result in the loss of some of the true positives. High numbers of CM-type hits were recorded in theses possessing tables of crystallographically determined bond lengths and angles in multiple undivided columns, resulting in lines with a 30+ character length, which could not be captured and removed.[41] We additionally found a smaller number that we regard as true false positives, and a regular flow of hits found when author initials (found in abundance in literature references) were mistaken for chemical elements and any word written in capital letters

- Trost Jpn A-Z Ph.D.
- B,C,F,H,I...
- DOCTOR CONTENTS ABSTRACT
- OTBSDIBAL-H toluene

These arise from the current mixture of methods that OSCAR3 uses for chemical name recognition: *one man's junk is another man's data*; i.e., what is perceived as a false positive in this project may well be regarded as a true positive by others.

Additional false positives are formed from incomplete names and terms that result from dissociation and fragmentation of larger entities:

- 2,6-dimethylbenzoate
- (*S*)-2-Chloro-6-oxo
- 4S*)-3-(tert-butyldiphenylsiloxy)methyl-2,4-dimethyl-cyclohexene
- 2E,4E,6S*,7R*)-7-(carbonyloxy

OSCAR3 could have been retrained and the regular expression terms optimized in order to meet requirements for e-thesis data extraction. Alternatively, a simple method of preventing these NCE's from appearing in the current RDF output was to use a string-length filter in the SAF to RDF stylesheet. Examination of 184 preparative procedures identified in the three PDF theses studied in detail showed that all systematic chemical names in the preparative titles are greater than 15 characters in length, and only 6 are less than 20 characters in length. The choice of which string-length cutoff to use is a balance between recall and precision; higher recall by removal of NCE's only shorter than 15, rather than 20, characters also results in recovery of some additional false positives (e.g., ACKNOWLEDGMENTS, tetracarbonyliron, Cyclohexenyl Core105, 2-sulfonylmethyl).

NCE Duplication and Distribution. Between 2000 and 8000 NCEs per thesis were identified by OSCAR3 in the corpus of theses studied. This number is dependent upon thesis size and subject area. Of the identified NCE hits, approximately 80−90% were found to be duplicates, leaving between 400 and 1500 unique NCE values per thesis. The significance of such duplicates is debatable; multiple occurrences of a particular word might be indicative of thesis subject area but could simply be a function of document length and also add a CPU overhead to searching.

Duplicates were readily removed using XSL stylesheets. Analysis of the deduplicated lists showed those theses from the organic/biochemical/polymer areas to have a frequency of hits for the chemical term (ONT) and chemical name (CM) subclasses ranging from 140 to 200 and 280 to 820, respectively, while theses from other areas (physical/theoretical/spectroscopic) had values in the range 70−160 and 150−600. Modest hit rates of 10−30 for the reaction (RN) subclass were found in the set of organic theses.

PDF Documents: Systematic Chemical Names Associated with Preparative Procedures. Three synthetic organic theses[53] in standard PDF format were examined in detail to discover the percentage of the systematic names used as titles of preparative procedures that were captured in the processed RDF (Table 1). All the systematic names identified were found in the chemical name (CM) subclass. Unlike the DOCX procedure outlined above, we were unable to programmatically determine whether these identified NCEs were actually part of a preparative experimental procedure because of the multiple linefeed problem. However, by using an empirical string cutoff length of 20 characters, we were able to filter the list of hits to help remove false positives. Although the resulting recall values are acceptable, showing that the PDF text-stream processing to reunite broken chemical name fragments was largely suc-

Table 1. Analysis of Systematic Chemical Names Associated with Preparative Sections in PDF Processing

		chemical names identified (CM-type)				precision (%)	
	A: synthetic preparations[a]	B: total	C: word-length restricted[b]	D: preparative chemical names identified	recall (%) (D/A)	total (D/B)	word-length restricted (D/C)
thesis 1[c]	60	1329	218	57	95	4.2	26
thesis 2[d]	55	575	141	40	72	6.9	28
thesis 3[e]	69	790	176	45	65	5.7	25

[a] Note that some preparative sections (e.g., in Thesis 1) may contain the synthesis of more than one compound (such as two or more stereoisomers). [b] Using additional string-length cutoff criteria: ≥20 characters. [c] Many preparative names are duplicated in a Table of Contents, giving rise to additional fragmented names that partly account for the high number of identified NCEs. [d] There are six 'indexed' nomenclature examples that are not fully recognized. [e] Many preparative procedures in this thesis have multiple whitespace errors in the stereochemical descriptors for the systematic names. These break the OSCAR3 processor.

Table 2. Metrics for the Identification of Preparative Sections in 10 Theses[a] in DOCX Format

	preparative sections	true positives	false positives	conditional false positives[b]	false negatives
thesis 1	30	5	66	0	25
thesis 2	73	55	4	3	18
thesis 3[c]	52	48	80	29	4
thesis 4[c]	52	42	260	10	10
thesis 5	64	44	1	0	20
thesis 6	65	55	6	0	10
thesis 7[d]	0	0	7	0	0
thesis 8	62	46	3	0	16
thesis 9	95	84	2	0	11
thesis 10[e]	39	0	0	0	39
total	532	379	429	42	153

[a] The following theses were analyzed: Adams, N. Combinatorial-type Approaches to Transition Metal Chemistry and Catalysis; Oxford University, 2003. Mitchell, C. Microencapsulation and Organocatalysis in Organic Synthesis; University of Cambridge, 2005. Harter, J. π-Allyltricarbonyliron Lactone Complexes: Tools for Asymmetric Synthesis; University of Cambridge, 2002. Cowburn, C. Studies Toward the Total Synthesis of Thapsigargin; University of Cambridge, 2003. Stepan, A. New Methodologies and Their Application in a Synthesis toward Leustroducin B; University of Cambridge, 2006. Taylor, S. The Design, Synthesis and Application of Solid Supported Reagents; University of Cambridge, 2001. Ulgot, B. Electron Trafficking in Molecular Assemblies; University of Cambridge. Polara, A. New Strategies for the Synthesis of Polyketide Natural Products; University of Cambridge, 2006. Talbot, A. Toward a Unified Strategy for the Spongistatins; University of Cambridge, 2006. Tait, M. The Synthesis of New Analogues of the Natural Product Thapsigargin; University of Cambridge, 2006. [b] The conditional FP measure, i.e., the number of false positives which would have occurred if all experimental sections had been correctly identified. [c] Theses 3 and 4 both contained extracts from crystallographic information files that accounted for the majority of the false positives identified. [d] This is a nonsynthetic chemistry thesis. [e] In this thesis, the experimental section was not algorithmically identified.

cessful, the observed precision achieved even with filtering (ca. 25%) is significantly lower than that seen in DOCX processing.

Chemical Object Identification in DOCX documents. Previous analysis[32] of experimental sections in Royal Society of Chemistry (RSC) journal articles to identify COs using OSCAR3 tools showed high figures for both recall and precision: 88% recall for CO's (IR, proton-NMR, carbon-NMR, MS spectra) and 95% precision.

While we would not expect the values to be as high for theses because the OSCAR3 expressions have been optimized for data in the RSC-specified format, anecdotal evidence has shown that it is the recall that falls far more quickly than the precision. A more in-depth analysis was beyond the scope of this project, requiring the human annotation of a complete corpus by different domain experts to measure interannotator agreement and the comparison with the computed results.[54]

Sections Describing Preparations. A sample of 10 DOCX theses was analyzed to calculate recall and precision metrics for the identification of preparative sections and hence those NCEs deemed to have associated preparations and analytical information. The identification of COs in preparations is entirely dependent on the ability to identify the experimental section (Table 2).

Table 2 shows the metrics for the identification of the preparations. These values give a recall of 71% and precision of 47% and conditional precision 90%. This highlights the importance of good tools and algorithms for identifying sections of documents. The vast majority of the false positives are caused by OSCAR3 identifying fragments of crystal structure analyses (such as atom names in bond length and bond torsion angle tables) reported in the theses as chemical names, as each line of the atom coordinate file contains what OSCAR3 considers to be a chemical name. The associated preparations extracted are of effectively zero length and may be trivially removed if required.

CONCLUSIONS

We have shown that it is possible to identify and extract chemistry from legacy formats into semantically rich representations with reasonable recall and precision. We have highlighted several areas that require further improvement of the heuristics used to identify sections. It is hoped that in the future institutions will mandate the deposition of data in richer formats as well as the version encapsulated in the final manuscript. The development of XML-based authoring tools is likely to drastically improve the availability of research data for reuse.

ACKNOWLEDGMENT

SPECTRa-T was funded by the JISC Repositories and Preservation Programme[55] as a joint project between Cambridge University Library and the Chemistry Departments of the University of Cambridge and Imperial College London.

REFERENCES AND NOTES

(1) Murray-Rust, P.; Rzepa, H. R.; Tyrell, S. M.; Zhang, Y. Representation and Use of Chemistry in the Global Electronic Age. *Org. Biomol. Chem.* **2004**, *2*, 3192–3203.

(2) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757–772.

(3) Manola, F.; Miller E. (Eds.) *RDF Primer*, World Wide Web Consortium W3C, 2004; http://www.w3.org/TR/rdf-primer/ (accessed November 26, 2009).

(4) Fanning, B. A. Preserving the Data Explosion: Using PDF. Digital Preservation Coalition and The Association for Information & Image

Management (AIIM), 2008; http://www.dpconline.org/docs/reports/dpctw08-02.pdf (accessed November 26, 2009).

(5) de Laet, A.; Hehenkamp, J. J.; Wife, R. L. Finding Drug Candidates in Lost/Emerging Chemistry. *J. Heterocycl. Chem.* **2000**, *37*, 669–674.

(6) Electronic Theses Online Service (EthOSnet). http://www.ethos.ac.uk/ (accessed November 26, 2009).

(7) Narcis, the Gateway to Dutch Scientific Information: Promise of Science. 2009; http://www.narcis.info/index/tab/publication/Language/en/ (accessed November 26, 2009).

(8) DART-Europe E-theses Portal (DEEP). 2007; http://www.dart-europe.eu/index.php/index (accessed November 26, 2009).

(9) ADT Australasian Digital Theses Program. 2009; http://adt.caul.edu.au/ (accessed November 26, 2009).

(10) Murray-Rust, P.; Downing, J.; Townsend, J. Chem4Word. 2009; http://www.chem4word.com/ (accessed November 29, 2009).

(11) Daconta, M. C.; Obrst, L. J.; Smith, K. T. *The Semantic Web: A Guide to the Future of XML, Web Services and Knowledge Management*; Wiley: Indianapolis, 2003.

(12) Stephens, S.; La Vigna, D.; DiLascio, M.; Luciano, J. Aggregation of Bioinformatics Data. *Web Semantics* **2006**, *4*, 216–221.

(13) Walker, F. L.; Gallagher, M. E.; Thoma, R.; PDF File Migration to PDF/A: Technical Considerations. http://archive.nlm.nih.gov/pubs/ceb2007/2007020.pdf (accessed November 26, 2009).

(14) ISO 19005−1:2005, Document Management—Electronic Document File Format for Long-Term Preservation—Part 1: Use of PDF 1.4 (PDF/A-1). 2006; http://www.aiim.org/documents/standards/19005-1_FAQ.pdf (accessed November 26, 2009).

(15) What is Tagged PDF? http://www.planetpdf.com/mainpage.asp?webpageid=1269 (accessed November 26, 2009).

(16) Davis, J.; Shur, A. OPC A New Standard For Packaging Your Data. 2009; http://msdn.microsoft.com/en-us/magazine/cc163372.aspx (accessed November 29, 2009).

(17) OASIS: Advancing the Standards for the Open Information Society. 2009; http://www.oasis-open.org/who/ (accessed November 26, 2009).

(18) Dublin Core Metadata Initiative. 2009; http://www.dublincore.org/ (accessed November 26, 2009).

(19) Ph.D. Thesis Regulations, California Institute of Technology, 2008; http://www.gradoffice.caltech.edu/documents/PHD-Thesisregulations.pdf (accessed November 26, 2009).

(20) Specifications for Thesis Preparation, Massachusetts Institute of Technology, 2009; http://libraries.mit.edu/archives/thesis-specs/ (accessed November 26, 2009).

(21) Lewin, I. Using Hand-Crafted Rules and Machine Learning To Infer SciXML Document Structure. *Proceedings of the 7th E-Science All Hands Meeting (AHM2007)*; Nottingham, UK.

(22) Le, X. L.; Straughan, S. R.; Thoma, G. R. Greek Alphabet Recognition Technique for Biomedical Documents. *Proceedings of the 5th. International. Workshop on Document Analysis*; Berlin (2002), pp 423−42.

(23) Library of Congress Classification Outline Class Q—Science. 2009; http://www.loc.gov/aba/cataloging/classification/lcco/lcco_q.pdf/ (accessed November 26, 2009).

(24) Downing, J.; Murray-Rust, P. TheOREM Marked-up Theses, 2009; http://wwmm.ch.cam.ac.uk/projects/theorem/theses/ (accessed November 29, 2009).

(25) Rhodes, J.; Boyer, S.; Kreulen, J.; Chen, Y.; Ordonez, P. Mining Patents Using Molecular Similarity Search. *Pacific Symp. Biocomput.* **2007**, *12*, 304–315.

(26) Grego, T.; Pezik, P.; Couto, F. M.; Rebholz-Schuhmann, D. Identification of Chemical Entities in Patent Documents. *IWANN*; Omatu, S., et al., Eds.; Springer-Verlag: Berlin, 2009; pp 941−948.

(27) Tsuruoka, Y.; Tsujii, J; Ananiadou, S. FACTA: A Text Search Engine for Finding Associated Biomedical Concepts. *Bioinformatics* **2008**, *24*, 2559–60.

(28) JISC Projects, CheTA (Chemistry using Text Annotations). 2009; http://www.jisc.ac.uk/whatwedo/programmes/inf11/cheta.aspx (accessed November 29, 2009).

(29) (a) Kidd, R. Project Prospect from the RSC: The Evolving Journal Article and Chemical Education. *Abstracts of Papers*; 235th ACS National Meeting, New Orleans, LA, April 6−10, 2008. (b) Batchelor, C. R. Project Prospect and the InChI. *Abstracts of Papers*; 237th ACS National Meeting, Salt Lake City, UT, March 22−26, 2009.

(30) Corbett, P.; Murray-Rust, P. High-Throughput Identification of Chemistry in Life Science Texts. *Computational Life Sciences II*; Springer: Berlin, 2006; pp 107−118.

(31) Townsend, J. A.; Adams, S. E.; Waudby, C. A.; de Souza, V. K.; Goodman, J. M.; Murray-Rust, P. Chemical Documents: Machine Understanding and Automated Information Extraction. *Org. Biomol. Chem.* **2004**, *2*, 3294–3300.

(32) Rupp, C. J.; Copestake, A.; Teufel, S.; Waldron, B. Flexible Interfaces in the Application of Language Technology to an eScience Corpus. *Proceedings of the 4th UK E-Science All Hands Meeting*; Nottingham, UK, 2006.

(33) Command line-driven software to enable the described workflows and interface with OSCAR3 has been written in Java and is available as Open Source on the SourceForge website: https://spectra-chem.svn.sourceforge.net/svnroot/spectra-chem/spectrat-textmining/trunk/ (accessed November 29, 2009).

(34) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2008**, *36*, D344–D350 (database issue).

(35) The lexicons used for this project are available as part of the OSCAR3 package: http://sourceforge.net/projects/oscar3-chem/ (accessed November 29, 2009).

(36) PubChem. 2009; http://pubchem.ncbi.nlm.nih.gov/ (accessed November 29, 2009).

(37) Waldron, B.; Copestake, A. A Standoff Annotation Interface between DELPH-IN Components. *NLPXML-2006 (Multi-Dimensional Markup in Natural Language Processing)*, Trento, Italy, 2006.

(38) Apache PDFBox is an open source Java library for working with PDF documents: Apache Software Foundation, 2008; http://pdfbox.apache.org/ (accessed November 26, 2009).

(39) CambridgeSoft, 100 Cambridge Park Drive, Cambridge, MA 02140 [http://www.cambridgesoft.com (accessed November 26, 2009)].

(40) Symyx Technologies, 2440 Camino Ramon, San Ramon, CA 94583 [http://www.symyx.com (accessed November 26, 2009)].

(41) The algorithm searches for three or more consecutive lines of text in the output stream which contain less than 30 characters. These values were set empirically. Manual analysis of one thesis[53a] showed ca. 170 such sections to be present, which were identified with 95% success rate.

(42) There are a number of unresolved name-to-structure issues with OPSIN (including *R*/*S* stereochemistry): Murray-Rust, P. http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=691 (accessed November 29, 2009). In the three PDF theses studied in detail,[53] 95% of the preparative procedures were of chiral structures.

(43) XML Pointer Language (XPointer), W3C, 2001; http://www.w3.org/TR/WD-xptr (accessed November 29, 2009).

(44) Corbett, P.; Copestake, A. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinf.* **2008**, *9* (Suppl 11), S4; http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2586753 (accessed November 29, 2009).

(45) Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, 29–37.

(46) Web Ontology Language OWL, W3C, 2004; http://www.w3.org/TR/owl-features/ (accessed November 26, 2009).

(47) SKOS, W3C, 2008; http://www.w3.org/TR/2008/WD-skos-reference-20080609/ (accessed November 26, 2009).

(48) Adams, N.; Semantic Chemistry, Semantic Technology Conference, **2009**. http://semanticuniverse.com/articles-semantic-chemistry.html (accessed November 29, 2009).

(49) Adams, N.; Winter, J.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *J. Chem. Inf. Model.* **2008**, *48*, 2118–2128, and references cited therein.

(50) Taylor, K. R.; Gledhill, R. J.; Essex, J. W.; Frey, J. G.; Harris, S. W.; De Roure, D. C. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.* **2006**, *46*, 939–952.

(51) Casher, O.; Rzepa, H. R. Semantic Eye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing. *J. Chem. Inf. Model.* **2006**, *46*, 2396–2411.

(52) Dodds, L. Introducing SPARQL: Querying the Semantic Web. 2005; http://www.xml.com/pub/a/2005/11/16/introducing-sparql-querying-semantic-web-tutorial.html (accessed November 26, 2009).

(53) (a) Harter, J. π-Allyltricarbonyliron Lactone Complexes: Versatile Tools for Asymmetric Synthesis; Dept. of Chemistry, Cambridge, 2002 (converted to PDF from the original Word document)[24]. (b) Brown, S. B. Iminium and Enamine Activation Methods for Enantioselective Organocatalysis; CalTech, 2005;http://etd.caltech.edu/etd/available/etd-02242005-174252/ (accessed November 26, 2009). (c) Lambert, T. H. Development of the Lewis Acid Catalyzed Allenoate−Claisen Rearrangement. Investigations of Enantioselective Catalysis of the Allenoate−Claisen Rearrangement. Studies towards the Total Synthesis of Erythrolide E; CalTech, 2004; http://etd.caltech.edu/etd/available/etd-12112003-091509 (accessed November 26, 2009).

(54) Corbett, P.; Batchelor, C.; Teufel, S. Annotation of Chemical Named Entities. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational and Clinical Language Processing*; Association for Computational Linguistics: Morristown, NJ, 2007; pp 57−64.

(55) JISC Repositories and Preservation Programme; 2009; http://www.jisc.ac.uk/whatwedo/programmes/reppres.aspx (accessed November 29, 2009).

CI9003688