ARTICLE

OBC
www.rsc.org/obc

Chemical documents: machine understanding and automated information extraction†

Joe A. Townsend, Sam E. Adams, Christopher A. Waudby, Vanessa K. de Souza, Jonathan M. Goodman* and Peter Murray-Rust

Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, UK CB2 1EW. E-mail: pm286@cam.ac.uk; Fax: +44 1223 763076; Tel: +44 1223 763069

Received 21st July 2004, Accepted 8th October 2004
First published as an Advance Article on the web 20th October 2004

Automatically extracting chemical information from documents is a challenging task, but an essential one for dealing with the vast quantity of data that is available. The task is least difficult for structured documents, such as chemistry department web pages or the output of computational chemistry programs, but requires increasingly sophisticated approaches for less structured documents, such as chemical papers. The identification of key units of information, such as chemical names, makes the extraction of useful information from unstructured documents possible.

Introduction

Scientific information is global and many disciplines now recognise the need to make their data widely and freely accessible. The development of the Semantic Web[1] and the Grid (typified by the UK eScience program) is based on instant access to raw and processed scientific data. In biosciences, for example, web-based databases are commonplace and in many cases are seen as the first place for finding and re-using information. Examples are Ensembl,[2] the Protein Data Bank[3] and SwissProt,[4] all of which contain highly structured data with varying degrees of curation and annotation. These data are "machine-understandable"—a computer can not only read the characters ("machine-readable") but also has semantics and metadata which allow it to take autonomous actions such as aligning sequences or discovering binding sites.

Much modern bioscience ("systems biology") is multidisciplinary and relies on integrating data from different disciplines. Many of these have less structured data, and the cost of abstracting this in a traditional human manner is often too large. There is, therefore, a considerable effort in machine extraction of information from the primary literature and other related sources. It is noticeable that biosciences have a great need for structured chemical information and this is not currently available in open, machine-accessible, and understandable form.

This article highlights the need for machine-based information extraction in chemistry. We distinguish *information retrieval* (the process of identifying a document or subdocument by its associated concepts) from *information extraction* (obtaining structured information from the document). Information extraction can be used for many purposes:

- populating a structured database (*e.g.* of chemical names and connection tables)
- compiling a lexicon or dictionary of commonly used terms (*e.g.* solvents)
- building an ontology for semantic processing and machine reasoning, for example by the OWL language[5]
- data-mining (building predictive models from data)
- re-input into computational chemistry programs
- proof-checking (*e.g.* for self-consistent data).

We note that chemical information is micro-published and that few if any chemical projects publish collections of structured data. Nor, apart from chemical and protein crystallography,[6] is there any standard method of publishing structured chemical information either in primary or secondary publications.

Chemical information is available in a huge quantity and diverse quality. The search for particular data may begin with an index or a database, but ultimately it is necessary to read the papers themselves in order to be sure that the right information is available. If it were possible to set a computer to read the literature on our behalf, this major task could be removed, or at least reduced. Machine understanding of this vast resource of data is not currently possible.

Chemistry is one of the most fruitful disciplines for information extraction as there is considerably more *de facto* uniformity than other disciplines:

- concepts are very well understood (many have survived for over 100 years)
- terms are often well formalised (*e.g.* through IUPAC).
- many articles are, by convention, highly structured and relatively homogenous between publishers
- in some areas (*e.g.* chemical diagrams) the number of tools in common use is small, so there is a *de facto* uniformity of approach
- much chemistry occurs in regulated processes (patents, drug regulatory) which require highly structured documents
- much information is computer-generated or mediated (computational chemistry, spectra, *etc.*).

Information extraction uses many aspects of document structure and content. Simple and important examples are the commonly used words and phrases (entities) that identify instances of essential concepts. In chemistry these include:

- bibliographic components (authors, journals)
- molecular identity (name, connection table, synonym)
- properties (units, physical properties, colours, form/nature)
- procedures: (solvents, amounts, colours, reagents, techniques)
- instruments (manufacturer, specification)

These can be used to give context or to classify documents or subdocuments.

It is possible to automatically extract information from chemical papers, cross check it and assemble it into searchable databases. This is only possible because the chemical literature has a reasonably rigid structure, which is centred on molecules.

Results

Web pages

Most university chemistry departments maintain web pages with information about their staff and their activities. This is a

† This is one of a number of contributions on the theme of molecular informatics, published to coincide with the RSC Symposium "New Horizons in Molecular Informatics", December 7th 2004, Cambridge UK.

DOI: 10.1039/b411033a

resource that is freely available and reasonably simple. The title page for each department will have some fundamental information about the institution, but is not usually a document of the complexity of an academic paper.

An index of chemistry departments and journals around the world would, without constant revision, become dated and useless very rapidly.

The web site: http://www.ch.cam.ac.uk/c2k/ is an index of chemistry departments that is regularly checked and corrected. Every month, the pages indexed are downloaded (either one page or one frameset, depending on how the site is set up) and checked for chemical content. This is not a spider that searches through web sites, but a targeted program that downloads just one browser screen full of information, and then attempts to parse it to check that it is still a department of chemistry. Each month, about 1% of the database is highlighted as requiring further checking. Of these, the majority are sites for which nothing can be downloaded, either because the address is outdated, or, more usually, because the server is temporarily unavailable. In the latter case, the program automatically attempts to download the page several times and reports failure only when all approaches have been unsuccessful. A few of the pages will contain information, but information that is no longer appropriate for a list of chemistry departments.[7]

Chemistry department web pages take a wide variety of forms, and relevance is checked by searching for the presence of some keywords and the absence of others. Websites that have information in languages other than English are checked in a similar way. The program contains words relevant to chemistry from a number of languages. This works effectively, but can occasionally make mistakes—for example when the web page has a notice resembling: "The chemistry department web page has moved...". The program correctly picks up key words about chemistry, and is unable to parse the sentence to discover that the link is no longer relevant.

The program would be fairly easy to deceive, should anyone attempt to do so. It also relies on universities having standard web addresses - .edu in the USA, .edu.[countrycode] in some areas, .ac.[countrycode] ('ac' stands for 'academic') in others.

This approach is effective for such simple documents, even though they are diverse in structure. More complex data sources require more complex analysis.

Computational chemistry programs

There are standard ways of calculating molecular properties that are implemented by many different programs. For each method there are many options that can be varied to give different results from calculations which are similar, and which may appear identical on a superficial inspection. For example, an MM2 force field calculation on a molecule will give slightly different results if the original MM2 version is used[8] which uses bond dipoles to calculate electrostatic interactions, or if the widely-used MacroModel modification[9] which uses atom-centred point charges is used. Both may be described as MM2 (the latter should be called MM2*), both give similar results, but sometimes the many small differences in electrostatic interactions add up to a big effect. The problem is even more pronounced for quantum chemistry packages, where small differences in the convergence criteria for any of the parts of the calculation may lead to very different final results for calculations which might both be correctly labelled in an identical way. Sometimes different groups do what appears to be the same calculation, and reach different conclusions, because different conformation analyses or convergence criteria were selected.

Program input is designed to be machine-parsable and errors render it invalid. Normally it is interpreted by bespoke procedural code but in principle it can be represented by a grammar and, therefore, be processed by standard, compiler-like, tools.[10] Program output is more variable, but is generated by a machine and thus consists of a finite vocabulary and syntax.

An example GROMACS[11] input file is shown in Fig. 1. A pre-processor was used to remove all the comments (everything that follows a semicolon). The data contained in the [pairs] block can then be identified and parsed by the structure in Fig. 2, where we have only shown the parsing process for PairLine1. SPACE, EOL, String, INT, PAIR and FLOAT are all terminal tokens and PairBlock, PairLineBlock, PairLine, PairLine1, LongPairLine1, PairLine2 and BLANKLINES are non-terminals.

```
#include "ffgmx.itp"
#include "spc.itp"

 [ moleculetype ]
 ;name nrexcl
DRG      5
[ atoms ]
 ;  nr  type resnr resid  atom  cgnr charge
     1     O     1 DRG      O8     1    0.000
…
     8     H     1 DRG     HAB     1    0.280
 [ bonds ]
 ;ai  aj  fu    c0          c1
   1   2   1 0.123    502080.0 0.123    502080.0
;    O8   C3
                …
   7   8   1 0.100    374468.0 0.100    374468.0
;    N4  HAB
 [ pairs ]
 ;ai  aj  fu    c0          c1
   1   4   1                    ;    O8   C7
                …
   5   8   1                    ;    HA   N6
 [ angles ]
 ;ai  aj  ak  fu    c0          c1
   1    2    3    1  120.0           418.4  120.0
418.4 ;    O8   C3   N2
…
   4    7    8    1  120.0           418.4  120.0
418.4 ;    C5   N4  HAB
 [ dihedrals ]
 ;ai  aj  ak  al  fu    c0    c1 m  c0   c1 m
   2   5   3   1   2   0.0 1673.6 0   0.0 1673.6
0 ; IDI    C3   N4   N2   O8
                …
   7   6   4   8   2   0.0 1673.6 0   0.0 1673.6
0 ; IDI    N6   C5   N4   C3
 [ system ]
PRODRG in water
 [ molecules ]
 DRG   2

 SOL   2747
```

Fig. 1 An example of a GROMACS input file.

```
PairBlock ::= PAIR:t1 BLANKLINES PairLineBlock
            ;
PairLineBlock ::= PairLine
                | PairLineBlock PairLine
                ;
PairLine ::= PairLine1
           | LongPairLine1
           | PairLine2
           ;

PairLine1 ::= SPACE INT:i1 SPACE INT:i2 SPACE INT:i3
SPACE FLOAT:f1 SPACE FLOAT:f2 EOL
{: System.out.println("<pair atom1='"+i1+"'
atom2='"+i2+"' function='"+i3+"' param1='"+f1+"'
param2='"+f2+"' />");
:}
   ;
```

Fig. 2 The structure used to process the [pairs] section of pre-processor parsed GROMACS output.[17]

An overall document structure was defined in a similar manner, which allowed the order of the sections to be arbitrary and some sections were allowed to be optional. This means that program input can be read, validated and reused independently of the code required to run it.

We tried to apply a similar approach to output files from MOPAC,[12] (Fig. 3) but even small, regular, sections proved too complex to be dealt with using this approach. Typical problems are the high proportion of numeric data without other terminal tokens.

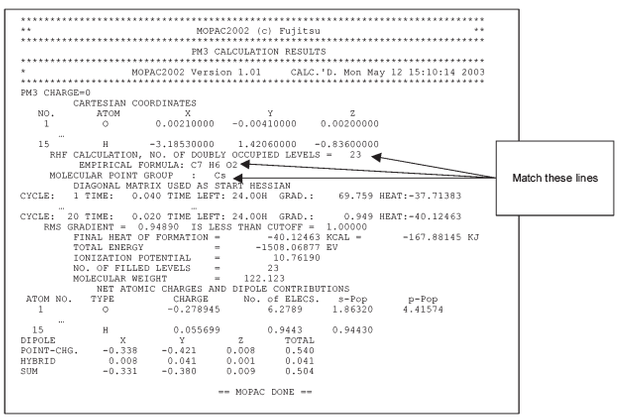

Fig. 3 MOPAC output file fragment. The lines highlighted would be matched by the template presented in Fig. 4.

```
<primitiveGroup name="chemistryData.head"
id="chemistryData.head.id">
  <primitive id="calcTypeAndLevels"
regexp="\s{6}(.*)\s+CALCULATION\.\sNOz\.\sOF\sDOUB
LY\sOCCUPIED\sLEVELS\s=\s+([0-9]+)">
    <scalar calcType="{$1}" />
    <scalar nDoubleFilledLevels="{$2}" />
  </primitive>
  <primitive id="chemFormula"
regexp="\s{11}EMPIRICAL\sFORMULA:\s+(.*)">
    <scalar formula="{$1}" />
  </primitive>
  <primitive id="pointGroup"
regexp="\s{6}MOLECULAR\sPOINT\sGROUP\s{3}:\s+(.*)"
>
    <scalar pointGroup="{$1}" />
  </primitive>
</primitiveGroup>
```

Fig. 4 A JUMBOMarker MOPAC template.

Program output often has a well-defined structure but irregular annotations and error messages make perfect parsing difficult. The output is designed to be human-readable sometimes at the expense of being easily machine-understandable. We have developed a parser (JUMBOMarker[13]) to convert machine-produced output to XML. This was tested on 750 000 jobs (MOPAC, GULP,[14] GROMACS, GAMESS[15]). We found correct conversion occurred at a rate greater than 99%.

JUMBOMarker contains relatively simple concepts of syntactic and lexical analysis but was written as a deterministic one-pass application that recognises 99% of the information in computational output. The program is based on structured templates containing regular expressions[16] defining the target document and syntax of a parsed document. (For example, the MOPAC output, Fig. 3, can be processed by the template in Fig. 4 to give the XML in Fig. 5). The parsing process may fail when unusual phrases or paragraphs are present (*e.g.* for errors or reports of unusual input). This can sometimes result in much of the output being unrecognised, and it may be useful to extend JUMBOMarker to a two-pass system (a pre-processor which structures the input into smaller chunks fed into a lexical analyser). Our current experience is that such a pre-processor will work well for many programs.

The template approach has many strengths for use in parsing computational chemistry output files.

• The XML can be further processed by Stylesheets so that context is maintained.

• They clearly indicate the final structure of the document.

• They can deal with a wide variety of different levels and options of output.

Table 1 A list of the most commonly found analytical data types

Name	IR	Optical rotation
Yield	UV	Refractive index
Boiling point	H-NMR	tlc
Nature	C-NMR	Elemental analysis
Melting point	Mass spectroscopy	High-resolution mass spectroscopy

```
<scalar calcType="RHF" />
<scalar nDoubleFilledLevels="23" />
<scalar formula="C7 H6 O2" />
<scalar pointGroup="Cs" />
```

Fig. 5 The XML structure of the parsed MOPAC data.

• They are easy to maintain and document.

• They are fast and can work with very large files (300 Mb).

In principle, the templates can have the full power of the XML Schema content model but we have not implemented all of the constructs. As a result, unexpected elements or unusual data order can lead to match failures, and so some of the information is not captured. However, the parser is capable of "catching up" with later sections it can match.

Synthetic and mechanistic chemistry papers

References. Most chemical papers contain a list of references at the end. This is very structured information, which cites a fairly small number of journals, and a rather larger number of people. It is possible to check references by comparison with a database of journals. A program to do this, *Cyril*,[18] was written in 1992, on the Apple Macintosh program, *HyperCard.* Although it worked quite well, and could check theses and highlight variant spellings and abbreviations for journals (*Angew. Chem., Int. Ed. Engl.* showing more variation than any other) the program was not widely used, as it was restricted to the Apple Macintosh, was slow to run, and there was no good distribution mechanism. All of these problems are now straightforward to overcome, using the WWW.

Experimental data. All synthetic chemistry papers have an experimental section describing how the experiments were performed, and listing analytical data for the molecules that were synthesised. These analytical data are presented in a very structured way that may be parsed automatically. The necessity of conforming to predefined specifications makes this data amenable to analysis by regular expressions. These have been incorporated into a program (OSCAR, an experimental data checker[19]) that was developed in collaboration with the RSC, and is available on the RSC's web site: http://www.rsc.org/is/journals/checker/run.htm.

Regular expressions have been developed to identify and extract the most common analytical data types (see Table 1), as found in the example corpus of *ca.* 100 articles. The test set comprised seven articles randomly selected from *Organic and Biomolecular Chemistry* 2003 and three documents that the RSC had received for submission to *Organic and Biomolecular Chemistry.* The recall and accuracy statistics for the recall and precision obtained from this sample are given in Table 2. The analysis was designed to be as critical as possible—thus, if a comma were omitted from an NMR spectrum (causing OSCAR to recognise a partial spectrum) this would be classified as a false positive. Table 3 shows some sample regular expressions.

In general, the less tight the specifications on the data, the lower the recall rate. Thus, identification of the nature of a compound (which has the least controlled vocabulary) presents more problems to parse than the NMR spectra. Although creating larger lists of regular expressions to match the state, colour and colour-modifiers of a compound would increase the recall, an entirely different approach, such as machine learning, would be preferable because this would remove the reliance on human authoring and curation.

Table 2 Breakdown of recall and precision rates for data identification by OSCAR

Data type	TP	FN	FP	Recall (%)	Precision (%)
Overall	1554	240	96	86.62	94.18
C NMR	187	14	5	93.03	97.40
Elemental analysis	103	15	0	87.29	100.00
H NMR	212	23	4	90.21	98.15
HRMS	126	1	0	99.21	100.00
Infra red	186	19	8	90.73	95.88
Mass spectroscopy	145	20	0	87.88	100.00
Melting point	151	11	2	93.21	98.69
Chemical name	171	72	47	70.37	78.44
Nature	100	22	21	81.97	82.64
Yield	173	43	9	80.09	95.05

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

TP	The number of X that the system correctly identified and were present in the corpus
FN	The number of X that the system failed to recognise
FP	The number of X that were recognised by the system which were not in the corpus
X	A particular data type

Table 3 Some examples of regular expressions used in OSCAR

Melting points	\b(m\.?p\.?)(\s*(?:\((?:[^\(\)]|\((?:[^\(\)]|\([^\(\)]+\))+\))+\)))?\W+([+−\−\?]?\<\d+(?:\.\d+)?(?!\d)\s*[−\−\?]\s*[+−\−\?]?\<\d+(?:\.\d+)?(?!\d)|[+−\−\?]?\<\d+(?:\.\d+)?(?!\d))\s*.C(?:(\W*lit.*?\d\W*.C\s*\d*\s*[\]\)]?)|(\W*dec\w*\.?[\]\)]?)|\W*\(from[^\(\)]+\))*
Hydrogen NMR	\b(?:(NMR\W*)?(d||δ|ä)H|1H(?!\))\W*NMR|\b(?−i:H)\b(?![\)\]−\−\?]))(?:\W*for\s+\w+(?:(![\(\);]).)*?)?(((?![\(\[\{])\W)*(?:\((?:[^\(\)]|\((?:[^\(\)]|\([^\(\)]+\))+\))+\)))?(\W+(d||δ|ä|H)+\b)?[\s:=]+?(?:\W*ppm\W*?)?(?:peaks\s+at\s+)?((?:about\s+)?[+−\−\?]?\d{1,3}(?:\.\d+)?(?:\s*[−\−\?]\s*\d{1,3}(?:\.\d+)?)?(?:\s**)?(?:\s*(?:(?:\((?:[^\(\)]|\((?:[^\(\)]|\([^\(\)]+\))+\))+\)|\[(?:[^\[\]]|\[[^\[\]]+\])+\]|\{(?:[^\{\}]|\{[^\{\}]+\})+\})(?:\s**)?))(?!\w|\s*=)(?:(?:\s*(,|and|including|&|;|\/)\s*)+((?:about\s+)?[+−\−\?]?\d{1,3}(?:\.\d+)?(?:\s*[−\−\?]\s*\d{1,3}(?:\.\d+)?)?(?:\s**)?(?:\s*(?:(?:\((?:[^\(\)]|\((?:[^\(\)]|\([^\[\]]|\[[^\[\]]+\])+\)|\{(?:[^\{\}]|\{[^\{\}]+\})+\})(?:\s**)?))?)(?!\w|\s*=))+(?!−)

Chemical names. Chemical structures form the basis of most organic chemistry. The two dimensional representations of molecules (sometimes with an indication of the three dimensional structure included) are the form which chemists use when describing a molecule, or a reaction mechanism.

A connection table can usually describe the structure of an organic molecule. Systematic chemical names are often avoided, until the molecule is to be included in a formal report. Chemical names are often abbreviated or non-systematic versions are used because systematic names are often lengthy, difficult to interpret and less memorable.

Although commercial programs to parse chemical names are available, they are not perfect and do not reveal the details of their processing algorithms. We have developed the approach of Kirby *et al.* to create a parser that converts IUPAC-like nomenclature to an XML based parse tree.[20] The process reveals both the reasoning and any errors or ambiguities and can also be used to generate a conventional connection table. These tools can be used for batch processing and for the exhaustive search for chemical entities in running text. Currently, we are populating its lexicon with lexemes and morphemes from the IUPAC nomenclature specifications.[21] As this will not match the many non-IUPAC lexemes, we are investigating machine-learning routines for expanding and updating the lexicon.

Chemical names in running text. We have also investigated the feasibility of identifying chemical entities in general scientific publications (particularly in biosciences). The goal was to markup publications such that chemical concepts (molecules) could be linked to online resources thus providing readers and authors with immediate access to rich chemical information. A corpus of 295 letters from *Nature* was used to create a lexicon of the most common chemical names. The world-wide-web was explored to see what percentage of these could be found by chemical robots and further contained some or all of physical properties, connection table or 3D coordinates. A thorough search was made of sites offering some or all of this information (Table 4). Forty two sites were identified, but in many of these the information was sparse or fragmented and so only five sites were finally used (Table 5).

An analysis of the 295 letters revealed that 15% of the letters contained a significant amount of chemical nomenclature (at least 10 names) and a further 19% useful molecular information. The papers containing chemistry were further analysed and 695 chemical entities were recognised of which 368 were unique and 268 could be easily identified using the top three resources.

The lexicon was transformed into a database by the addition of the web based data and structures. When reading a paper the scientist is presented with known names but can also use the regular expressions from OSCAR to identify other possible chemicals. The scientist can follow these as HTML links to search the online databases. If a hit is found the results can be automatically added to the local database. In this way every reader contributes almost painlessly to the aggregation of high quality information. Databases can be customised to support different classes or uses of compounds. The recall and precision can be high as shown in Fig. 6.

Descriptions of synthetic chemistry procedures. Analytical data is usually preceded by a less structured paragraph, but with a high density of technical terms. This semi-structured human-authored text should yield to several complementary approaches:

• Processing to recognise common punctuation (units, amounts, sentence boundaries)

• Shallow parsing using language-sensitive tools. This discourse uses a small subset of sentence structures and can often be completely parsed without chemical knowledge. This is particularly useful for parts-of-speech tagging which highlights the role of unknown lexemes. Fig. 7 shows automatic parsing and parts-of-speech tagging for a typical sentence from chemical synthesis.

• Entity recognition. This is critical for many abbreviations and may provide additional context for structure and recognition.

Table 4 A list of non-subscription, open access web sites that hold molecular information[a]

Antoine	http://www.mitchellscientific.com/antoinequery.html
ATSDR (toxicity faqs)	http://www.atsdr.cdc.gov/cgi-bin/search-tox?words=&scope=ToxFAQs+and+Public+Health+Statements
ChemACX	http://chemacx.cambridgesoft.com/chemacx/index.asp
chemcompass (suppliers)	http://www.chemcompass.com/
chemcyclopedia (suppliers)	http://www.mediabrains.com/client/chemcyclop/BG1/search.asp
ChemDat-Merck	http://chemdat.merck.de/
ChemExper	http://www.chemexper.com/
ChemFinder	http://chemfinder.cambridgesoft.com/
Chemicals with pharmaceutical activity	http://www.chem.ox.ac.uk/mom/chemical-database/
ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/
ChemIndex	http://ccinfoweb.ccohs.ca/chemindex/search.html
Chemnet	http://www.chemnet.com/suppliers/
DBGET(genomenet)	http://www.genome.ad.jp/dbget-bin/www_bfind?compound
Fisher catalog-some products use ChemExper database	https://www1.fishersci.com/catalogs/root.jsp
HIC-Up	http://xray.bmc.uu.se/hicup/
Imperial MOTM (history)	http://www.ch.ic.ac.uk/motm/
liqcryst-online	http://liqcryst.chemie.uni-hamburg.de/lc/fc_lolas_e.html
Matweb-materials (polymers or properties)	http://www.matweb.com/index.asp?ckck=1
MDPI	http://www.mdpi.net/search.html
Molecular database without transition elements	http://www.faidherbe.org/site/cours/dupuis/banque.htm
Molecule of the month	http://www.bris.ac.uk/Depts/Chemistry/MOTM/motm.htm
molecules R Us	http://molbio.info.nih.gov/cgi-bin/pdb
MSD ligand chemistry	http://www.ebi.ac.uk/msd/Services.html
NCI	http://dtp.nci.nih.gov/docs/dtp_search.html
ncms-solvDB	http://solvdb.ncms.org/solvdb.htm
NIAID-Anti-HIV/OI Chemical Compound Search	http://apps1.niaid.nih.gov/struct_search/
NIOSH Pocket Guide to Chemical Hazards	http://www.cdc.gov/niosh/npg/npgd0000.html
NIST	http://webbook.nist.gov/chemistry/name-ser.html
NTP Chemistry H&S	http://ntp-server.niehs.nih.gov/Main_Pages/Chem-HS.html
NYU Mathmol library	http://www.nyu.edu/pages/mathmol/library/
organic compounds database	http://www.colby.edu/chemistry/cmp/cmp.html
Oxford MOTM	http://www.chem.ox.ac.uk/mom/
PDB	http://www.rcsb.org/pdb/index.html
reciprocalNet	http://www.reciprocalnet.org/recipnet/search.jsp
SDBS	http://www.aist.go.jp/RIODB/SDBS/sdbs/owa/sdbs_sea.cre_frame_sea
Sigma-Aldrich	http://www.sigmaaldrich.com/
smell database	http://mc2.cchem.berkeley.edu/Smells/index.html
The MSDS Hyperglossary glossary index	http://www.ilpi.com/msds/ref/
TheMSDS.com (MSDS)	http://www.TheMSDS.com
Thermogalactic spectra online	http://spectra.galactic.com/SpectraOnline/Default_ie.htm
Vermont SIRI	http://hazard.com/msds/index.php
Wellesley-alphabetical listing of molecules	http://www.wellesley.edu/Chemistry/Flick/molecules/newlist.html

[a] This list was assembled and checked in September 2003. Some of these sites are no longer available.

Table 5 The number of molecules and properties that can be associated with them given in five open sites[a]

	ChemExper	ChemIDplus	NCI	NIAID	NIST
Number of molecules	100000	367447	270000	50000	~70000
2D structure	y	y	y	y	y
3D structure	y	y	y		
2D-coordinates	y	y	y		y
3D-coordinates	y		y		y
Molecular formula	y	y	y	y	y
Molecular weight	y		y	y	y
Synonyms	y	y	y	y	y
SMILES			y		
Density	y				
Melting point	y				y
Boiling point	y				y
MSDS	y				
IR	y				y
Mass spec.					y
Other physical data				y	y
Suppliers	y				

[a] These data were collected from the websites in September 2003.

Table 6 A list of some of the terms used for part of speech tagging

S	Sentence
NP	Noun phrase
VP	Verb phrase
PRN	Parenthetical
AUX	Auxiliary verb phrase
DT	Determiner
NN	Noun, singular or mass
VBD	Verb, past tense
PP	Prepositional phrase
IN	Preposition/subordinating conjunction
CD	Cardinal number
NNP	Proper noun, singular

Buffers and pH Experiments were carried out in a physiological salt solution (PSS) containing 130 mM *NaCl*, 0.9 mM *NaH2PO4*, 5.4 mM *KCl*, 0.8 mM *MgSO4*, 1.0 mM *CaCl2*, 25 mM *glucose*. This solution was buffered either with *HEPES* alone (20 mM) or HEPES/EPPS/*MES* (8 mM each; HEM-PSS), to cover a wider pH range. HEPES-buffered PSS was used in all experiments unless HEM-PSS is specifically mentioned. *HEPES* is *4-(2-hydroxyethyl) piperazine-1-ethanesulphonic acid*, EPPS is N\'-*(2-hydroxyethyl)piperazine-N*\'-3-propanesulphonic acid, *MES* is *2-(N-morpholino)ethanesulphonic acid*. The pH of all solutions was adjusted using a carefully calibrated pH meter (Metrohm). All data in this report are referenced to pH at room temperature. To obtain pH at 37 °C, 0.15 pH units should be subtracted for *HEPES* buffers in the range of pH 6.8–7.8 according to our calibration experiments. IP formation assay Confluent cell cultures grown in 24-well plates were labelled with *myo[3H]inositol* (100 MBq ml-1; ART/Anawa Trading) for 24 h in serum-free DMEM medium. Where indicated, cells were pretreated with PTX (100 ng ml 1; Sigma) during the last 4 h of *inositol* loading. Cells were then incubated at 37 °C in PSS with indicated buffer. *Lithium* (20 mM) was added to block inositol monophosphatase activity, leading to accumulation of IP1,5, 6. Where indicated, bovine thrombin, SPC, *carbachol* or bradykinin (Sigma) was added 1 min before *lithium* addition. Unless otherwise stated, incubation was continued for 20 min. Cells were then extracted with ice-cold *formic acid* and total IPs separated from free *inositol* using batch column chromatography25. Data are shown as means s.e.m. for *triplicate* determinations.

Fig. 6 Markup of a paragraph from an article in the test set.[22] Words identified as chemical names have been underlined and italicised. Note that most chemical entities are recognised (false negatives include "PTX", "bradykinin" and "IPs"; a false positive is "triplicate").

• Limited regular expressions for stock phrases (*e.g.* instrumental parameters).

If sentences can be completely parsed in this manner, their context-independent meaning may be inferred.

Unstructured text. Chemical documents often show consistency in their structure, and many paragraphs and sentences can be recognised or classified. Tools include entity recognition and analysis of co-occurrence (*e.g.* by Bayesian or similar methods). It is however unrealistic to interpret the deep meaning of other components of chemical papers.

Tables, graphs and chemical diagrams. There are a number of common uses for tables such as co-reporting compounds and their properties and data. They are more variable than the analytical data parsed by OSCAR but it is worth continuing to extend the approach to include the parsing of tables, which are tractable but harder to parse than ordinary text. Images are almost invariably not tractable. It is hard even to discover whether a diagram represents a single compound. The styles used and the location of identifying numbers or names is highly variable.[24] In the immediate future, we expect chemical name analysis to be more fruitful than diagram analysis.

Representation of output

XML offers great advantages over conventional methods for representing the results of parses, and machine understandable documents. There are many emerging tools

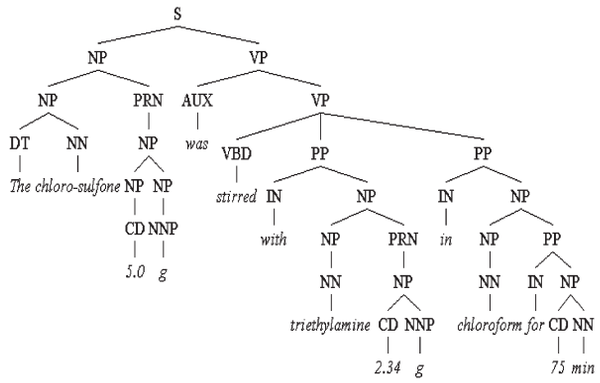

Fig. 7 The result of shallow parsing of a typical Chemical Synthetic paragraph. Parse performed by the Linguist's Search Engine[23] For an explanation of the terms see Table 6.

(RDF,[25] OWL, RSS,[26] *etc.*) that are designed to add value to an XML file. As XML preserves context, it is possible to re-use preliminary parses for further analyses. JUMBOMarker already has the ability to add metadata to the parsed output, for example in Dublin Core and CML formats. This is normally limited to program name and date. Author names and institutions are rarely included in the output.

Conclusions

Searching for particular words, or patterns of characters, a process that may conveniently be carried out using regular expressions, is an effective way of searching structured text. This has been used to analyse computational chemistry output and web pages (http://www.ch.cam.ac.uk/c2k/).

The analytical data of synthetic chemistry papers can be analysed in a similar way (OSCAR[19]). However, the regular expressions now need to be linked to a dictionary of key words, in order to pull out the key information, and some analysis is done in addition to the results from the regular expressions. Systematic and semi-systematic chemical names are susceptible to a similar approach, as they can be built up from a library of fragments in a useful proportion of cases. This works well in cases where it is clear which string of characters represents a chemical name.

The identification of chemical names in a block of text is a harder problem, as many non-chemical words have fragments in common with chemical names.

Descriptions of chemical procedures are less structured, and so are not effectively analysed using this approach, despite the high density of technical terms.

Unstructured text is currently impossible to analyse. Diagrams are hard to convert to connection tables, but are very powerful if this initial step can be done.

The future of chemistry depends on the automated analysis of chemical knowledge, combining disparate data sources in a single resource, such as the World-Wide Molecular Matrix,[27] which can then be analysed using computational techniques to assess and build on these data.[28] We have made substantial progress towards the goal of complete automation.

Acknowledgements

The RSC, Unilever, the EPSRC, and Nature Publishing Group are thanked for their support of this work. We particularly thank Richard Kidd and the RSC for making available documents from *Organic and Biomolecular Chemistry* in a variety of formats.

References

1 Semantic Web: http://www.w3.org/2001/sw/.
2 Ensembl: http://www.ensembl.org/.
3 Protein Data Bank: http://www.rcsb.org/pdb/.
4 SwissProt: http://www.ebi.ac.uk/swissprot/.

5 OWL language: http://www.w3.org/TR/owl-guide/.
6 Crystallographic Information File (CIF) http://www.iucr.org/iucr-top/cif/; S. R. Hall, F. H. Allen and I. D. Brown, *Acta Crystallogr., Sect. A: Fundam. Crystallogr.*, 1991, **47**, 655–685.
7 J. M. Goodman, *Molecules*, 2000, **5**, 33–36.
8 N. L. Allinger, *J. Am. Chem. Soc.*, 1977, **99**, 8127.
9 F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson and W. C. Still, *J. Comput. Chem.*, 1990, **11**, 440–467.
10 A. V. Aho, R. Sethi and J. D. Ullman, *Compilers: Principles, Techniques and Tools*, Prentice Hall International, Upper Saddle River, New Jersey, 2003.
11 GROMACS: http://www.gromacs.org/.
12 J. J. P. Stewart, *J. Comput. Aided Mol. Des.*, 1990, **4**, 1–45.
13 JUMBOMarker: http://wwmm.ch.cam.ac.uk/moin/JumboMarker/.
14 J. D. Gale and A. L. Rohl, *Mol. Simul.*, 2003, **29**, 291–341.
15 GAMESS-US: M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, *J. Comput. Chem.*, 1993, **14**, 1347–1363.
16 J. E. F. Friedl, *Mastering Regular Expressions*, O'Reilly & Associates Inc., Sebasopol, CA, USA, 2002, 2nd edn.
17 The GROMACS output was parsed using CUP (http://www.cs.princeton.edu/~appel/modern/java/CUP/): J. Levine, T. Mason and D. Brown, *CUP, a Java implementation of yacc: lex & yacc*, O'Reilly & Associates Inc., Sebasopol, CA, USA, 1992, 2nd edn.
18 Cyril: http://www.ch.cam.ac.uk/MMRG/cyril/.
19 S. E. Adams, J. M. Goodman, R. J. Kidd, A. D. McNaught, P. Murray-Rust, F. R. Norton, J. A. Townsend and C. A. Waudby, *Org. Biomol. Chem.*, 2004, **2**, DOI: 10.1039/b411699m.
20 D. I. Cooke-Fox, G. H. Kirby and J. D. Rayner, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 101–105; D. I. Cooke-Fox, G. H. Kirby and J. D. Rayner, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 106–112; D. I. Cooke-Fox, G. H. Kirby and J. D. Rayner, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 112–118.
21 R. Panico, W. H. Powell, and J.-C. Richer, *A Guide to IUPAC Nomenclature of Organic Compounds (recommendations 1993)*, Blackwell Science, Oxford, 1993, [ISBN 0-63203-4882]; Corrections published in: H. A. Favre, K.-H. Hellwich, G. P. Moss, W. H. Powell and J. G. Traynham, *Pure Appl. Chem.*, 1999, **71**, 1327.
22 M. G. Ludwig, M. Vanek, D. Guerini, J. A. Gasser, C. E. Jones, U. Junker, H. Hofstetter, R. M. Wolf and K. Seuwen, *Nature*, 2003, **425**, 93–98.
23 http://lse.umiacs.umd.edu:8080/.
24 Graphical representation standards for chemical structure diagrams: IUPAC project 2003-045-3-800. http://www.iupac.org/projects/2003/2003-045-3-800.html.
25 RDF: http://www.w3.org/RDF/.
26 RSS: http://www.w3.org/2002/01/rss/rss1_namespace/.
27 WWMM: http://wwmm.ch.cam.ac.uk/wwmm.html.
28 J. M. Goodman, *Philos. Trans. R. Soc. London, Ser. A*, 2000, **358**, 387–398.