Chapter 3 Automated Identification and Conversion of Chemical Names to Structure Searchable Information

Antony J. Williams, ChemZoo Inc. , 904 Tamaras Circle, Wake Forest, NC-27587. email: antony.williams@chemspider.com

Andrey Yerin, Advanced Chemistry Development Inc., Moscow Department, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation. email: erin@acdlabs.ru

Abstract

The communication of chemistry-related information occurs both via print and electronic media and chemical entities can appear as structure depictions or, more commonly, as systematic names (commonly either IUPAC[1] or CAS[2] names), as trade names or of one of a plethora of registry numbers (CAS[3], EINECS/ EC-number[4] or others). The preferable form of communication for a chemist is via a depiction of the chemical structure with an electronic molecular connection table as its basis. Electronic representations of chemical structures are one of the informatics underpinnings for any organization operating in the domain of chemistry or biology and enable the creation of a structure/substructure searchable database of chemical structures and associated data and knowledge. There is an enormous wealth of information embedded inside both print and electronic documents in the form of chemical names and a means by which to convert those alphanumeric text descriptors into a more rich chemical structure representation has long been the mission of a large group of investigators. The challenges and hurdles to success are quite profound in their nature. We will review the present state of this research and the efforts underway to recover the value of information textually trapped in publications, patents, databases and Internet pages across the multiple domains of chemistry.

Introduction

Chemical names have been in use as textual labels for chemical moieties even since the days of the alchemist. With increasing understanding of chemistry and the graphical representation of chemical structures came the need for an agreed upon language of communication between scientists. Eventually systematic nomenclature was established and then extended as deeper knowledge and understanding of molecular structures grew. One would hope for a single agreed upon international standard for systematic nomenclature adopted and understood by all chemists. Despite the efforts of IUPAC[5] such an ideal still does not exist, exists in many variations, has changed over time, can be organizationally specific, is multilingual and is certainly complex enough that the majority of chemists would struggle with even the most general heterocyclic compounds. The application of nomenclature by scientists of different skill levels is far from pure and chemical names for a single species will be heterogeneous. This does not bode well for clear communication in chemistry.

Chemical nomenclature is a specific language for communication between people with an understanding of chemistry. The language facilitates the generation of chemical names that are both pronounceable and recognizable in speech. The ability to communicate via systematic names collapses fairly quickly based on the complexity of the chemical structure and the associated name. Simple and short names are easily interpreted but in general most systematic names are rather long, complex and include non-linguistic components such as locants and descriptors made up of obscure numbers and letters. A chemical nomenclature system must continuously follow the increasing complexity and diversity of chemical structures as new chemistries are pursued. The majority of chemical names are rather complex and a chemist will

need a reasonable knowledge of the nomenclature rules to interpret a chemical name and convert back to a graphical structure representation. Chemical nomenclature rules and recommendations for IUPAC are now captured online in a series of volumes with several thousand pages[6].

Despite the limitations and challenges associated with chemical names graphical chemical structure representations on the other hand can easily be interpreted by humans even with the most rudimentary chemistry knowledge. Chemical structure representations were, of course, in use well before the advent of software programs for the generation of such figures. Structure drawing software was developed to provide a manner by which to store, transfer and homogenize molecular structure representations. The ability to both represent and transfer chemical structures electronically provided a significant boost to communication between chemists and structure images became the preferred medium for human recognition. Despite the availability of software tools for the graphical representation of chemical structures, chemical names, labels or abbreviations must still remain in order for us to converse. They remain as valuable terms of communication in patents and publications and essential to the process of chemical registration for a number of bodies. The generation of appropriate systematic nomenclature remains a challenge to even the most skilled chemist but since systematic nomenclature is rules-based the development of software tools to speed the process has been possible. The opposite is also true whereby the conversion of systematic names to the original chemical structure also remains just as much a challenge. By providing software tools for the conversion of differing chemical nomenclatures into universally recognized chemical structures, chemists can more easily review the chemical structure of interest and the data can be migrated to database technologies. This facilitates the integration of disparate forms of chemical information with the intention of enabling the discovery process.

There are numerous sources of chemical names. Commonly, chemical databases would not include chemical structures but be made up of lists of chemical names. Nowadays, thanks to the availability, cost and ease-of use of chemical structure databases many of these "text databases" have been converted into a structure format and most chemical databases are now structure searchable. A simple search of the internet will show that there are still many databases lacking chemical structures and therefore not searchable by structure in the original format, for example an online HTML page. These pages however can contain valuable information and, with the application of the appropriate name to structure (N2S) conversion tools can be made searchable.

Electronic documents exist in a plethora of formats, the most common being Microsoft Word, Portable Document Format (PDF) and web-based HTML formats, as well as a number of others. Electronic documents in general do not embed information regarding chemical structures, but do include chemical names that are extractable. It is likely that nearly all modern documents of interest to chemists are now made available in electronic format. Published both before and after the early stages of computerization, such documents might be considered as lost for chemical information. However, scanning and optical character recognition[7] (OCR) into electronic files provides a means for conversion by software tools. Of course, even without such tools, scientists commonly read print documents and manually convert the chemical names to structures. It should be noted that it is also possible to identify chemical images and convert them to structure searchable information using optical structure recognition (OSR). This is discussed in detail in Chapter 4 of this book.

The conversion of chemical names and identifiers into appropriate chemical structure representations offers the ideal path for chemists and organizations to mine chemical

information. Since chemical names are not unique and a multitude of labels can map to a single chemical entity, the facile conversion of alphanumeric text identifiers to a connection table representation will enable superior data capture, representation, indexing and mining. The industry need to mine more information from both the historical corpus as well as new sources is obvious and a number of researchers have initiated research into the domain of chemical identifier text mining and conversion. There have been a number of efforts in the field of bioinformatics research[8] and, while interesting as a parallel, we will focus the efforts in this chapter on the extraction and conversion of identifiers related to chemical entities rather than, for example, genes, enzymes or proteins.

Our intention in this book chapter is to examine the challenges related to extracting identifiers from chemistry-related documents and the conversion of those identifiers into chemical structures. The authors of this work each have well over a decade of experience in chemical structure representation and systematic nomenclature. We have been deeply involved in the development of software algorithms and software for the generation of systematic names and the conversion of chemical identifiers into chemical structures[9]. While we have our own biases in regards to approaches to the problem of N2S conversion, we have done our utmost to be objective in our review of the subject and comparison of approaches and performance.

Existing Structure Mining Tools and Projects

It is likely that ever since systematic nomenclature was introduced, chemists have wished for a simple way to convert a systematic name to a graphical representation of the associated structure. A number of organizations have built business models around the extraction and conversion of chemical names from different materials (e.g. publications, patents and chemical

vendor catalogs) to build up a central repository of chemical structures and links to associated materials. The Chemical Abstracts Service (CAS) is recognized as the premier database and presently contains over 33 million compounds[10]. Other offerings include those of Beilstein[11], Symyx[12] (previously MDL), Infochem[13] and VINITI[14]. These organizations manually curate, nowadays with the assistance of software tools, chemical structures and reactions from the respective publications and documents.

The delivery of new chemical entities of commercial value can clearly be constrained by the coverage of patent space. Chemical structure databases linked to patents are available (e.g. CAS[2], Elsevier[15] and Derwent[16]) and deliver high value to their users. Some of these organizations utilize both text-mining and N2S conversion tools prior to manual examination of the data. Two free-access services utilizing text-mining and conversion of chemical names to structures are those of SureChem[17] and IBM[18].

Both approaches use proprietary entity extraction tools developed and customized specifically for the recognition of chemical names[19,20]. The chemistry-specific entity extractors use a combination of heuristics for systematic names and authority files for entities that are less amenable to rules-based recognition, specifically drug and chemical trade names. During the extraction and conversion processes chemical entities are run through one or more N2S conversion tools to generate chemical structure data. A set of post-processing routines are applied to remove spelling and formatting errors that often cause N2S conversion failure but nevertheless their experiences have shown that due to the poor quality of many chemical names in patents and other text sources, not all of the names can be converted by commercially available tools.

SureChem offers a free access website for searching the world patent literature via text-based or structure/substructure searching[17], as well as commercial offerings based on the same chemical patent data (for example, they supply the data in formats to allow importing of the data into organizational databases). They have utilized a series of different N2S conversion tools under their system supplied by three commercial entities: ACD/Labs[21], Cambridgesoft[22] and Openeye[23]. They provide ongoing updates of the patent literature within 24 hours of release to the public and update their homepage accordingly with the latest statistics of extracted chemical names, details regarding each of the patent classes and the number of unique structures extracted to date. SureChem report the extraction of over half a *billion* chemical structures[17] from various patent granting bodies and these have been deduplicated to almost 9 million unique structures. They offer online access to various forms of patent literature including US and European granted and applications as well as WO/PCT documents[26] and Medline[27]. All of these sources are updated within a day of release of the updates from the patent offices to SureChem. Name to structure conversion results vary among the patent databases due to different levels of original text quality among the patent issuing authorities. SureChem reports[28] that in their latest database build they observed improvements of as much as 20 percent in name-to-structure conversion rates following application of new post-processing heuristics and expect further incremental increases in future builds.

IBM[18] also has a free access online demonstration system for patent searching via text or structure/substructure and presently exposes data extracted from US Patents (1976-2005) and Patent Applications (2003-2005). The work has been described in detail by Boyer *et al.* [24] and a brief overview of the technology is provided on the website[25] . They report using the Cambridgesoft Name=Struct[22] algorithms for their work. The IBM team have also analyzed both

granted and patent applications to present day for all sources listed above but these data are not yet exposed at their website and the exposed data is presently limited to USPTO patents and Medline articles issued up to 2005. IBM reports the extraction of over 4.1 million unique chemical structures. Caution should be taken with the comparison of unique chemical structures reported by SureChem and IBM as the methods of deduplication are not necessarily comparable and are not reported in detail. At present SureChem is the most mature free access online service, updated on a regular basis and covering a number of patent granting bodies.

Accelrys[29] have also developed text analytics capabilities for the purpose of extracting and converting chemical names. Using their Scitegic pipelining tools as the platform, the ChemMining[30] chemical text mining and conversion system has been developed. This software uses text-mining algorithms to extract chemical names and then feeds these to one or more of the commercial N2S conversion algorithms licensed by the user. After processing one or more documents a report is created showing the examined document(s) highlighted with all found the structures as live chemistry objects.

Murray-Rust *et al*[31,32] have examined the challenges associated with mining data from text and have encouraged the adoption of appropriate architectures, molecular identifiers and a shift towards more open data in order to facilitate information exchange in the sciences. They have appropriately espoused the virtues of their OSCAR system[32], a chemical data checker in an Open XML architecture, in terms of its benefits to authors, publishers and readers alike. In this work, compounds were identified by connection table links to open resources such as PubChem[33]. Originally a part of the OSCAR system, OPSIN[32,34] (Open Parser for Systematic Identification of Nomenclature) has been released as an Open Source Java library for parsing IUPAC nomenclature. OPSIN is presently limited to the decoding of basic IUPAC nomenclature

but can handle bicyclic systems, and saturated heterocycles. OPSIN does not currently deal with stereochemistry, organometallics and many other expected domains of nomenclature but since the source code is open it is hoped that this work can provide a good foundation technology for others to enhance and develop.

TEMIS and Elsevier-MDL[35] worked together[36] to develop the Chemical Entity Relationship Skill Cartridge™ to identify and extract chemical information from text documents. The software identifies chemical compound names, chemical classes and molecular formulae then translates them into chemical structures. They use a N2S translation service to match textual information with proprietary chemical libraries and provide a unique fingerprint is provided for de-duplication purposes. The cartridge integrates chemical name recognition software developed and used by Elsevier MDL to identify chemical names and extract reaction schemes from scientific literature and patents. This software was proven for more than two years in the production of the MDL® Patent Chemistry Database, including processing a backfile of more than 20 years of patents. Unfortunately these authors cannot locate any further details regarding the details of the software or performance. Research into text-mining continues to expand and a national center of text-mining, with a focus on the sciences, has been founded in the United Kingdom[37].

The projects outlined above all point to their focus on the extraction of chemical identifiers from text but there is a clear dependence on the N2S conversion algorithms in regards to the overall output of the various approaches. The remainder of this chapter will review the challenges associated with the development of N2S algorithms and how these can be addressed.

The General Approach for Mining Chemical Structures in Chemical Texts

The scheme by which chemical structures are mined from chemical documents is shown in Figure 1..

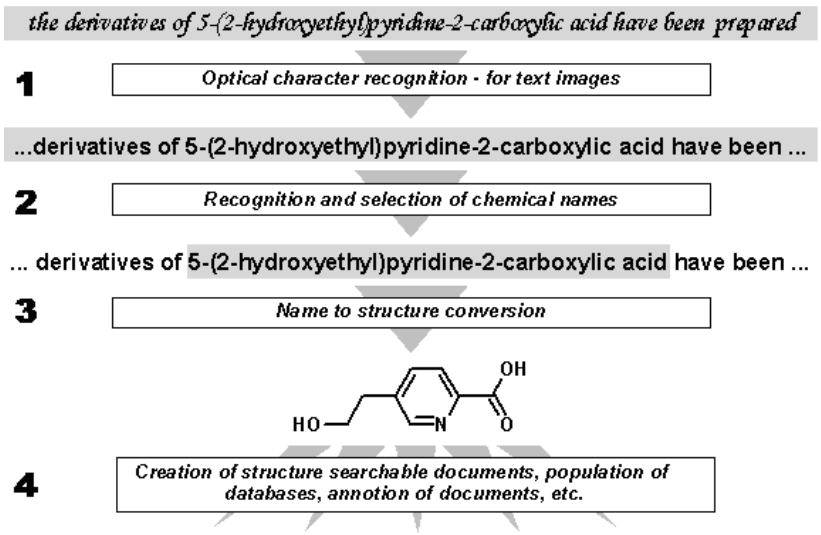


Figure 1 The general scheme for mining chemical structures from text

The greatest hurdle associated with successful mining of chemical structures via chemical name to structure conversion is the quality and complexity of the chemical names themselves. Thus, a significant part of this chapter is devoted to the consideration of the quality of names and its contributions to the procedure of conversion.

Text recognition in images – OCR of chemical texts.

Starting from the very beginning of OCR technologies a huge amount of resources was invested in the development of computer-based systems. For general language-based texts this problem has been efficiently solved and the success rate of recognition is higher than 99% for Latin-script texts[38]. The basic challenges of OCR have been reviewed elsewhere and will not be repeated here[39]. While OCR can efficiently handle generic text they experience fairly significant

limitations in the recognition of chemical names. In the same way that general OCR programs use language specific dictionaries to assist in the process of recognizing text, a chemically intelligent OCR program needs to use a dictionary of appropriate chemical text fragments and use a series of specific algorithms to recognize chemical names. Figure 2 illustrates the recognition of chemical name images captured with different settings. A standard software package was utilized for these test procedures[40]. Each of the examples shows the graphical image of the chemical name as well as that extracted by the software.

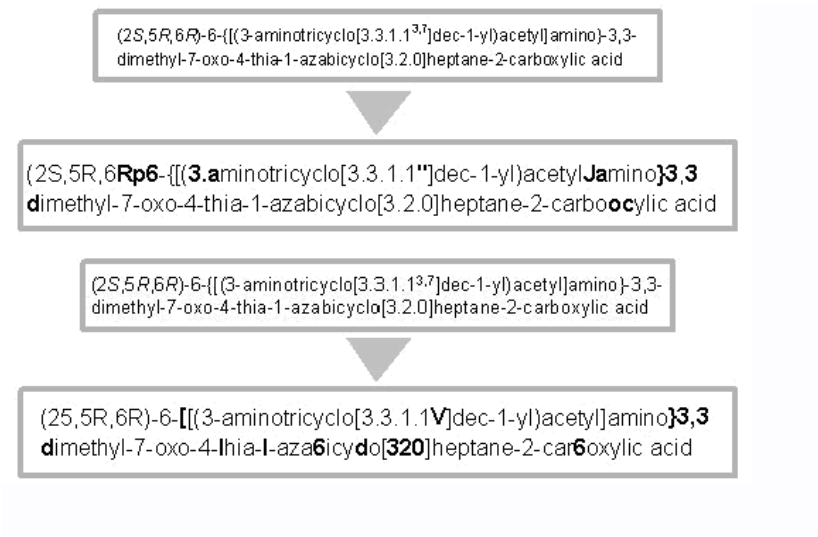


Figure 2. Problems with character recognition in chemical names.

While it is clear that recognition of this example can be easily improved by enhanced resolution of the initial image, this example is given to demonstrate the most common problems associated with chemical name recognition and possible errors introduced in chemical names. The problems include:

- Super and subscript recognition, especially in combination with italics;
- Introduction of additional spaces, often instead of paragraphs;

- Lost spaces mainly at line breaks;
- Dashes are often lost or mistaken as hyphenation marks;
- Incorrect recognition of punctuation marks – comma versus period;
- Misinterpretation of enclosing marks;
- Incorrect recognition of some letters and numbers, e.g. l, i, 1;
- Lost formatting, e.g. normal text vs. sub- or superscripted characters

As a result of these considerations we can conclude that OCR of chemical names can be improved by:

- utilizing higher resolution text images;
- usage of chemical dictionaries;
- modification of OCR algorithms for chemical name recognition and specifically retaining dashes and avoiding added spaces.

Chemical names selection/extraction.

When text analysis is required as a result of either OCR conversion or simply from direct electronic formats, the selection or recognition of chemical names becomes the challenge. As stated earlier, the nature of chemical names can vary widely and be represented either by single words or as a set of grammatically linked words. Another difficulty is that text within a chemistry context can include various terms derived from chemical names that serve as verbs, adjectives or plural forms describing processes, chemical relations or groups of chemical substances. For example, in the phrase "acetylation of isomeric diethylnaphthyridines with acetic anhydride" only one distinct chemical name can be selected: "acetic anhydride", though clearly the conversion of “diethylnaphthyridines” as a class of compounds could lead a reader to a text

of interest.

The first publications in this area were from the ‘80s and ‘90s[41,42]. This area of research now uses the general principles of Natural Language Processing (NLP) and specifically Named Entity Extraction (NER) enhanced with specific developments for chemical and biochemical name recognition[43,44]. Chapter 7 of this book is devoted to NLP and NER approaches applied to the extraction of chemical information and we will not discuss these approaches in more detail here.

The specific problems and potential solutions associated with chemical name recognition have been reviewed in a recent work describing the OSCAR3 software[32]. The general approach is the recognition of chemistry related terms whereby chemical names are identified by the appropriate algorithms. Chemical name identification uses several steps and procedures that may include:

- splitting words with common separators such as spaces and punctuation marks with spaces according to natural language and chemical name rules,
- recognition of chemical words using dictionaries of chemical lexemes,
- syntax and semantics analysis of relationships between words to recognize chemical names that include spaces.

Following the chemical name recognition process, annotated documents are created with specific tags to provide a reference to the specific part of the document where the specific chemical is mentioned. The extracted chemical names are then provided as inputs to the N2S algorithms and form the basis of the next section of this work.

Generating Chemical Structures from Chemical Names

Algorithmic Name to Structure Conversion and Related Software Applications

The first publication regarding the computer translation of chemical names was by published by Garfield in 1961. In that article he described the conversion of names into chemical formulae and initiated the path towards name to structure algorithm development[45]. Developments in 1967 at CAS provided internal procedures for the automatic conversion of CAS names into chemical diagrams[46,47]. The first commercially available software program was CambridgeSoft's Name=Struct released in 1999[48], now patented[49], and then followed shortly thereafter by ACD/Labs' Name to Structure product released in 2000[50]. There are two more commercial products available: ChemInnovation's NameExpert[51] and OpenEye's Lexichem[23] and ChemAxon[52] have announced the imminent release of their own product early in 2008. As mentioned earlier, an Open Source Java library for the interpretation of IUPAC systematic names[34], OPSIN, has also been made available. For this book chapter most examples are based on CambridgeSoft Name=Struct and ACD/Name to structure programs. We judge these programs to be the most advanced products in this area at present, but all considerations are general in nature and relevant to all of the conversion routines presently existing or still under development.

The vision for all name to structure conversion algorithms is likely consistent. Convert as many chemical names as possible to the correct chemical structure. While this is the general target, the approaches to arrive at the conclusion can differ. ACD/Labs have maintained an approach of caution in terms of name conversion initially focusing only on the translation of fully systematic names, controlling ambiguity to as high a level as possible yet supporting the conversion of trivial names using a dictionary lookup. Cambridgesoft has approached the

problem with the intention of converting as many names as possible and being fairly neutral in terms of name format and strict systematic nomenclature format. For many test comparisons, both approaches have their failings. ACD/Labs' product sometimes fails to successfully convert names yet with Cambridesoft will commonly convert a much larger proportion of the test set but with a larger number of inappropriate conversions. Many of the larger companies have chosen to support both approaches licensing both tools and performing intersecting comparisons and examining the results outside of the intersection for appropriateness. This approach has been taken by the groups analyzing the patent literature as discussed earlier and SureChem uses three name to structure products for their work as discussed earlier.

General scheme of name to structure conversion

The conversion of chemical names into chemical structures can be represented as two intersecting schemes – that of utilizing a look-up dictionary and of using syntax analysis. A combination of these two approaches is definitely needed for the analysis of chemical names experienced in the real world.

The figure below illustrates the simplest approach of using lookup tables. In this approach the N2S engine utilizes the relationship between a large database of chemical names and the corresponding chemical structures.

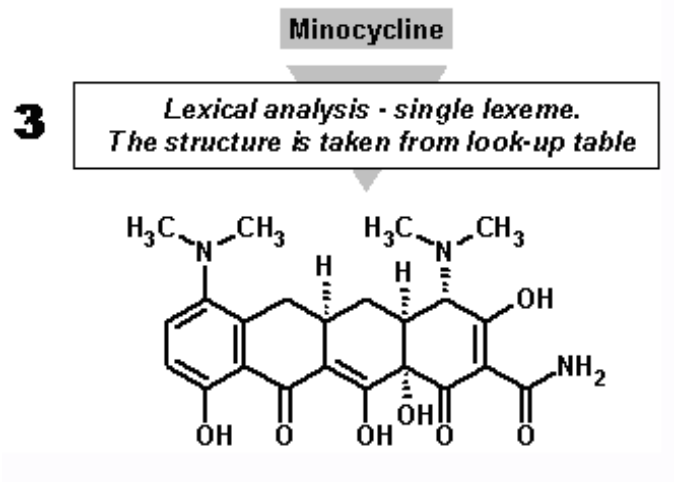


Figure 3 Single step conversion of trivial name

The rather restricted nature of this approach is obvious – the potential number of chemical structures and their associated chemical names is very large and cannot be included in a computer program of a reasonable size. Clearly, significant resources would be needed to create such a database of names and structures and keep it updated and distributed to users at the appropriate pace of chemical development. When the diversity of name formatting resulting from human intervention is taken into consideration then simply this factor will make N2S conversion essentially intractable. A lookup table approach is nevertheless very useful and InfoChem utilize their inhouse IC_{N2S} program for the purpose of chemical structure mining from texts using an internal file of 27 million names[53]. A lookup algorithm and associated databases is unavoidable for the treatment of trivial names and other structure identifiers such as registry numbers.

For the conversion of systematic names a more powerful and flexible approach must be based on the parsing of the chemical names and the application of syntax analysis. The scheme below illustrates the principle steps of this procedure.

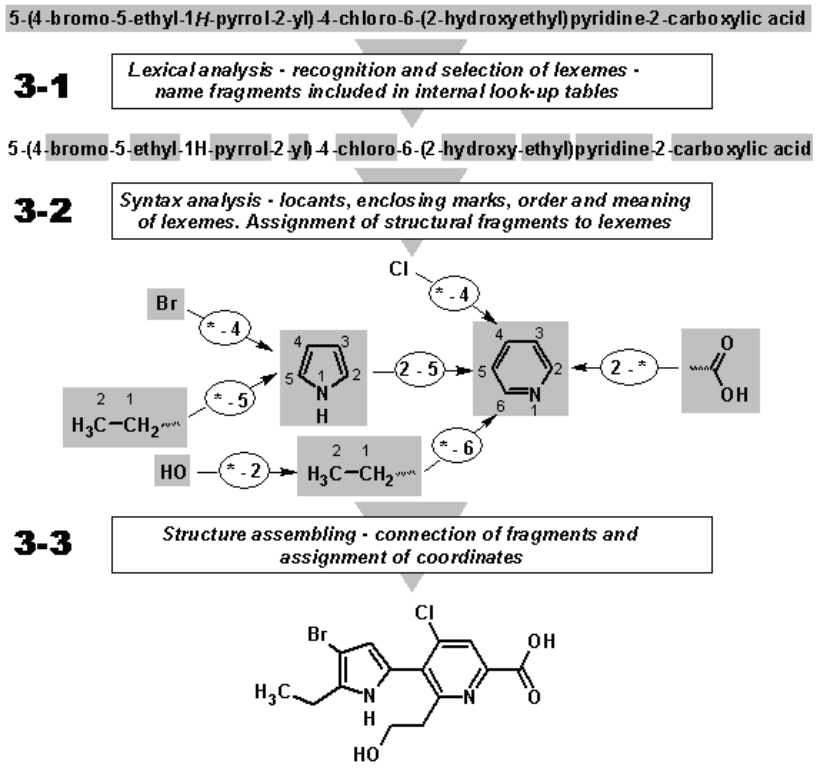


Figure 4 The general steps of conversion of an unambiguous systematic name

The **first step** in the process, lexical analysis, splits the whole chemical name into a series of name fragments, known as lexemes, that have structural and/or grammatical meaning. Also split out from the name are the locants, the enclosing marks and the punctuation marks. If any part cannot be recognized by the program then structure generation will normally fail or an attempt to continue generation by applying a rules-based spelling correction or ignoring a part of the input name can be performed. The lexical dictionary used at this stage is related to that described earlier to find the chemical names in the text.

The **second step** shown in the figure is the syntax analysis of the chemical name. At this stage the chemical name is analyzed according to chemical nomenclature grammar, assigns to each fragment its structural meaning, and attempts to derive a connection between the various structural fragments. In the simplest case of an unambiguous systematic name, all name parts can

be interpreted in only one way allowing the determination of a single chemical structure. This step is the primary component of an N2S engine. There are many challenges and problems associated with this engine and these are discussed below for specific chemical names.

During the **last step,** all structural name fragments are assembled into a chemical structure and atom coordinates are assigned to provide an attractive representation of the chemical structure for storage or exporting into various chemical formats including line notations, such as InChI (International Chemical Identifier) and SMILES (Simplified Molecular Line Entry System). The basic principles and problems of N2S conversion have been discussed previously by Brecher[54] in his description of the CambridgeSoft Name=Struct program. We will discuss here further challenges of N2S conversion concerning specific types of chemical names in relation to the mining of chemical structures from texts.

Conversion of trivial names.

As illustrated in Figure 1, the simplest N2S engine may be fully based on a look-up table and does not require the parsing of chemical names. As discussed above, while it is necessary to have large dictionaries of chemical names and structures, this approach is unavoidable for the conversion of names and structure identifiers where parsing cannot help in the process of structure generation. Such an algorithm can be used to convert trivial, trade and retained names together with registry numbers like CAS, EINECS and vendor catalog numbers.

One important aspect of this approach deserves mention – the support of stereoisomerism requires caution. There are many cases in the literature and in many databases where a specific stereoisomer is represented without definition of the configurations and the specific stereoisomer is simply implied. Figure 5 shows several examples of such cases.

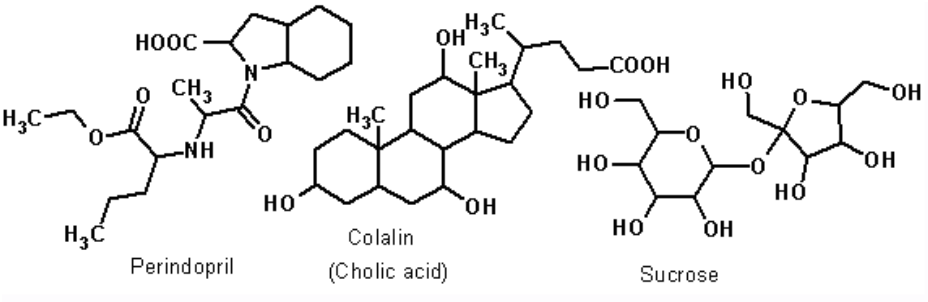
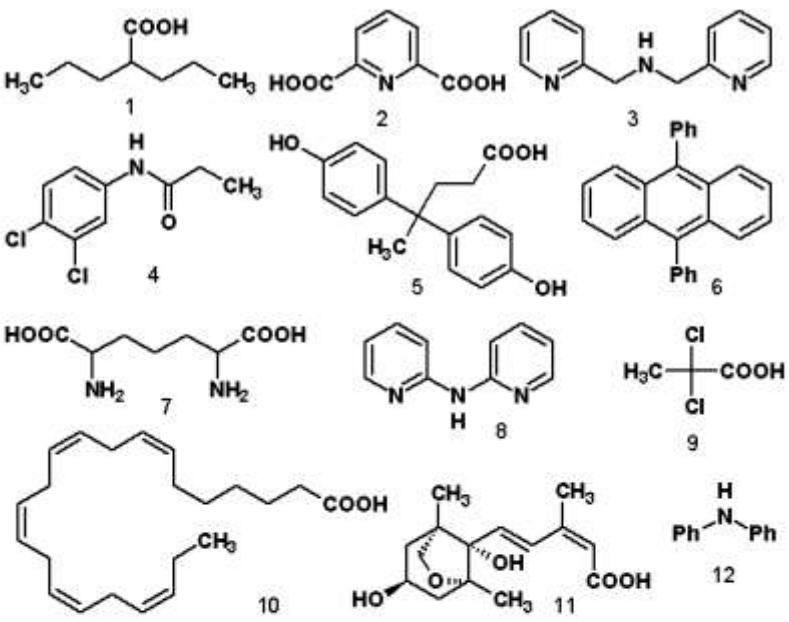


Figure 5 Stereoisomers represented without indication of configurations.

The structures shown in Figure 5 can give rise to 32, 2048 and 512 different stereoisomers and certainly do not accurately represent the chemical names displayed below. It is very common that the representation of stereochemistry for both amino acids and steroids is not reported in publications. Caution must be taken with the generation of the N2S engine structure dictionary from representations that omit stereo configurations, for example, from non-stereo SMILES notation.

In most cases, the N2S conversion of indivisible/elementary identifiers is safe and the quality of conversion depends only on the internal dictionary quality. One important exception is that of chemical abbreviations. While they can be treated as “trivial names” they are very context dependent and highly ambiguous since such a limited number of letters cannot be treated as a unique identifier.



1	DiPropylAcetic acid	7	2,6-DiaminoPimelic Acid
2	DiPicolinic Acid	8	Di(2-Pyridyl)Amine
3	Di(2-Picolyl)Amine	9	DichloroPropionic Acid
4	3',4'-DichloroPropionAnilide	10	DocosaPentaenoic Acid
5	DiPhenolic Acid	11	DihydroPhaseic Acid
6	9,10-DiPhenylAnthracene	12	DiPhenylAmine

Figure 6 Twelve structures that may correspond to DPA abbreviation. The letters used in the abbreviation are capitalized.

Figure 6 shows twelve structures that may correspond to the abbreviation of DPA. Six of them can be output by the ACD/Name to Structure software package and six more were found by browsing the Internet. Note that even a specific context cannot guarantee an exact meaning. For example, both structures 3 and 8 were found in publications about coordination compounds. In general, chemical abbreviations are *not* unique and can rarely be distinguished from other trivial

names except for the rather weak criterion that all letters are capitalized. We can conclude that conversion of any trivial name shorter than about 5-6 characters is not safe. A few rarer exceptions do exist but this is a very short list. Examples include reserved abbreviations such as those for dimethyl sulfoxide, DMSO and ethylenediaminetetraacetic acid, EDTA.

Conversion of systematic names

The lexical and syntax-based analysis of systematic names illustrated in Figure 2 depends directly on the algorithms underlying the name conversion engine. The set of lexemes that can be recognized by an algorithm are a critical characteristic of the program since it defines what type of names can be treated. However, the number of elementary lexemes is not the defining limitation of the program. The integration of the appropriate set of lexemes with the appropriate treatments for handling complex nomenclature grammar are superior to an extended set of lexemes. For example, the treatment of all fused system names requires the support of specific nomenclature grammar and approximately 100 specific lexemes. This approach is far more powerful than the support of a thousand fused system names represented as elementary lexemes such as furo[3,2-b]pyridine, cyclobuta[a]naphthalene, and so on.

Chemical nomenclature has a very large number of specific procedures to create chemical names and many of these are not easily amenable to algorithmic representation , requiring significant investments in both development and validation time to develop automated procedures. Software developers of N2S engines prefer to support just the basic operations for conversion at least at the early stages of development.

One of the largest challenges is that many chemical names, even when generated appropriately and without errors, are created according to different nomenclature systems.

Specifically, the two most-common nomenclature systems, those of IUPAC and CAS, have a number of differences and can lead to potential ambiguity of the names. The situation becomes even more complex when we take into account the fact that chemical names have mutated through history with the development of the nomenclature systems and so, for example, many chemical texts, will follow old nomenclature procedures thereby significantly expanding the number of nomenclature operations requiring support.

The conversion of systematic names to their chemical structures is a time-consuming, skill intensive process and is not a minor undertaking. Such a project is guaranteed to take many years of development to cover the most important nomenclature operations.

Quality of published chemical names

The main problem of name conversion is the rather low quality of published systematic names. It may be considered as one of the reasons for the appearance of N2S programs and the paper describing CambridgeSoft's Name⇔Struct program has a very symbolic title "Name⇔Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature"[54]. Most chemists have limited nomenclature knowledge and, therefore, their ability to resolve chemical names of fairly nominal complexity is a non-trivial task. The contrary is also true, the generation of systematic names for complex chemical structures can be a challenge and as a result there has been a proliferation of incorrect structure-name pairs not only to the Internet but also into peer-reviewed publications. A recent review of systematic nomenclature on Wikipedia chemicals by one of our authors (AJW), demonstrated significant gaps in quality to the point where the names represented very different structures to those discussed on the Wikipedia pages. The quality of published systematic names is rather low and

this is true not only of publications but also of patents. In a recent paper Eller[55] randomly selected about 300 names of organic chemicals cited to be systematic in nature. The names were extracted from four chemical journals and analyzed and compared to the corresponding names generated by a number of systematic nomenclature generation software packages. The results of this comparison are given in Table 1.

Table 1. Comparison of computer generated names with published names

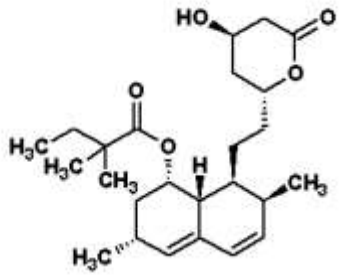
Results of analysis of 303 systematic names						
	unambiguous		**intolerable**		**no name**	
Published names	74%	224	26%	79		
AutoNom 2000	86%	260	1%	3	13%	40
ChemDraw 10.0	88%	267	1%	2	11%	34
ACD/Name 9.0	99%	300	1%	3	0%	0

Software for generating a systematic name from a structure has been available for well over a decade. Whether the issue is one of access to software or trust that software can produce high-quality systematic nomenclature, it is clear that papers still contain far too many errors in their systematic names. The data in the table reflect the situation in 2006. While this is not exactly a statistical sampling of data (only 300 names from 4 journals) the data suggest that about a quarter of published chemical names do not accurately represent the associated structures. There are two specific issues: 1) the chemical name does represent the structure and

can be converted back to the intended structure, but the name does not follow systematic nomenclature guidelines; 2) the chemical name does not represent the structure and when converted generates a *different* structure from that originally intended. The data in the table clearly demonstrate that algorithmically generated names are of dramatically higher quality and reliability than manually generated names and that wider adoption of software programs for this purpose will significantly improve the quality of published nomenclature. The barriers to this shift are likely threefold: awareness of the availability of such software applications, price and technology barriers to accessing such applications and trust in the ability of the software to produce an appropriate systematic name. Attention must be given to improved generation of systematic nomenclature as soon as possible since the proliferation of poor quality and the contamination of the public records can now occur at an outstanding rate with new software platforms.

Thielemann[56] recently commented that the number of mistakes in systematic names is far higher than that of trivial names. He provided examples as a result of his examination of patents regarding the cholesterol lowering drug Simvastatin. He observed that out of 141 patents examined, not one contained the correct IUPAC name of Simvastatin. He also pointed out in his presentation what the correct IUPAC name, in his opinion, was: *6(R)-[2-[8(S)-(2,2-dimethylbutyryloxy)-2(S),6(R)-dimethyl-1,2,6,7,8,8a(R)-hexahydronaphthyl]-1(S)ethyl]-4(R)-hydroxy-3,4,5,6-tetrahydro-2H-pyran-2-one*". Unfortunately this "correct name" is far from appropriate according to IUPAC rules, primarily due to the incorrect citation of stereodescriptors. Neither the CambridgeSoft Name⇔Struct nor the ACD/Labs Name to Structure software can convert the systematic name suggested by Thielemann back to the original Simvastatin chemical structure. In our judgment none of the commercially available

name to structure conversion algorithms can convert this name to the structure. The structure of Simvastatin with an appropriate IUPAC name is given in Figure 7.



(1*S*,3*R*,7*S*,8*S*,8a*R*)-8-{2-[(2*R*,4*R*)-4-hydroxy-6-oxotetrahydro-2*H*-pyran-2-yl]ethyl}-3,7-dimethyl-1,2,3,7,8,8a-hexahydronaphthalen-1-yl 2,2-dimethylbutanoate

Figure 7 The chemical structure and IUPAC name of Simvastatin.

This example demonstrates that one of the main challenges for a name to structure conversion algorithm applied to data mining is the conversion of chemical names that are not strictly systematic, are ambiguous or include typographical errors or misprints.

Ambiguous systematic names

It is not difficult to identify many ambiguities in chemical names in chemical catalogs, publications, patents, and Internet pages. Even the simplest structures can be given ambiguous names and cause confusion. Figure 8 shows a series of examples of names with missing locants or parentheses that very often, but not necessarily, lead to name ambiguity.

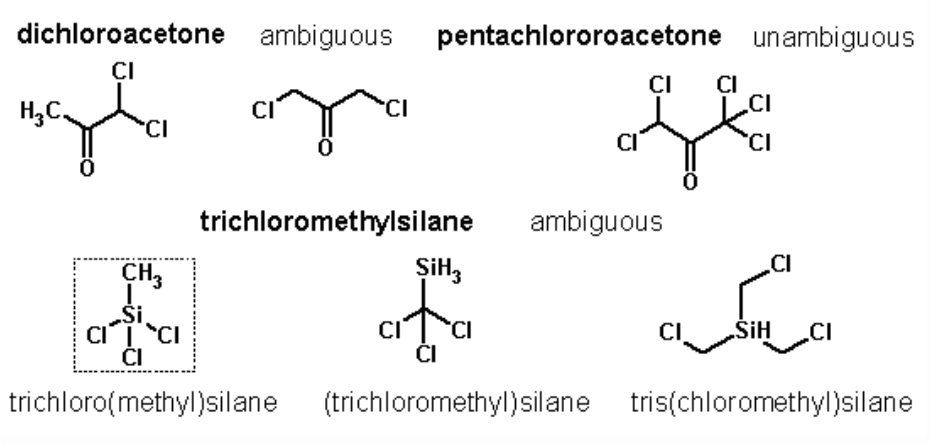


Figure 8 The potential ambiguity of names with missing locants or parentheses.

It should be noted that the name *trichloromethylsilane* is the correct CAS name for the framed structure that provides legal status to some ambiguous names. A more complex example of ambiguity introduced by missing parentheses is shown in Figure 9. In this case the recognition of ambiguity requires support of a specific nomenclature procedure, functional modification of trivial acid names.

In an example such as this, there are a number of ways to proceed: 1) convert the name to a single acceptable structure matching the ambiguous name; 2) do not convert the name to a structure but fail because of the ambiguous nature of the name; 3) convert the name to all possible structures in order to demonstrate potential ambiguity. For the example above, the commercial software providers take different paths. ACD/Name to Structure generates two structures for this name, while CambridgeSoft Name=Struct outputs only the second structure, since it is the most probable match given that the correct systematic name of the first structure is **4-(methylthio)benzoic acid**. For the >550 hits returned by a search in Google™ most, but not all, refer to the first structure.

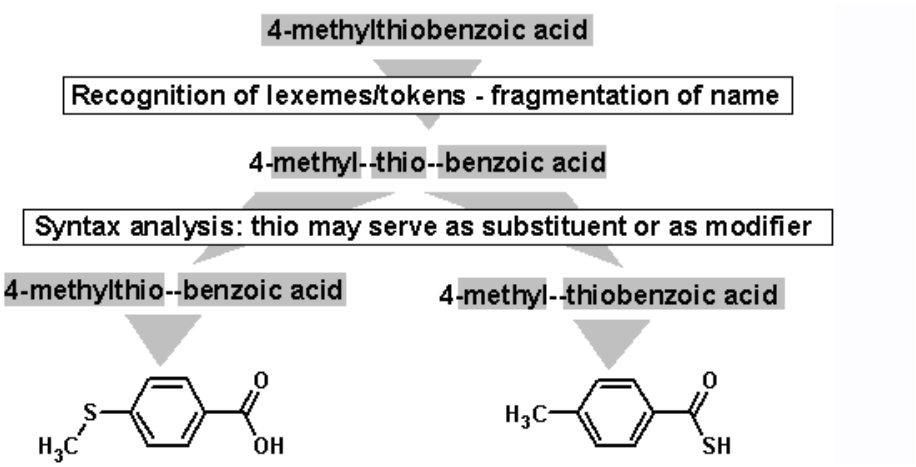


Figure 9 Different meanings of "thio" in an ambiguous name

A similar example is "4-methylthiophenol". This name also allows the generation of two structures but here the situation is reversed and in most cases refer to 4-methyl(thiophenol) (or 4-methylbenzenethiol according to current naming conventions).

This short overview with simple examples provides evidence for the need of warnings regarding ambiguity in the names. Clearly, the more complex a chemical structure is, the more potential there is for miscommunication. It is our belief that the recognition and reporting of ambiguities in chemical names and the associated structures generated by software programs must be implemented as part of any N2S engine in order to ensure some level of caution to provide reliable results.

Ambiguous vs. trivial names

One of the primary issues with systematic nomenclature is that some names can appear systematic in nature but, in fact, are not. They can have the expected structure of a chemical name generated according to a rules-based system but are false systematic names at least in their

specific context. When the name to structure conversion algorithms are too flexible in their implementation, for example, when the name is not present in a lookup dictionary or ambiguity is not reported, then such labels can be erroneously interpreted as systematic..

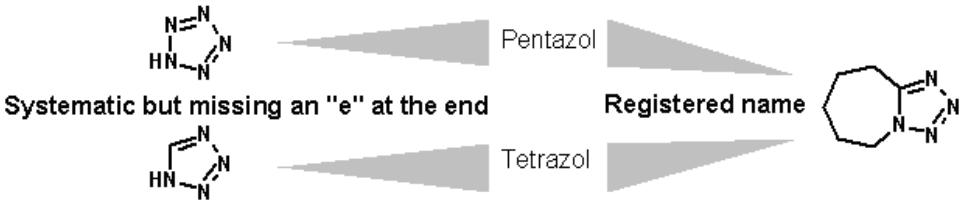


Figure 10 Alternative treatments of registered names

While the two names shown in Figure 10 are incorrect according to English IUPAC guidelines for the two structures on the left, they are only *almost* systematic. In fact, in German language nomenclature where the terminal "e" is not cited they are correct. However, both of them are listed as registered names for the structure shown on the right side and can be found on the ChemIDplus website[57]. There are many such examples that have proliferated this problem across the literature and other sources of chemical information. Thus, the support of trivial names is very important even in terms of helping to distinguish real systematic names from false systematic names. On the other hand, it would be highly desirable to discontinue the assignment of registered names that mimic systematic names and can therefore be misleading.

Spelling correction and treatment of punctuation

In previous sections we examined problems arising as a result of errors in nomenclature. Another very significant area is naming errors resulting from misprints or OCR misinterpretation as reviewed earlier in this article. Table 2 lists the most common naming errors and the reasons for their occurrence.

Table 2 Typical name errors and their reasons (The errors in examples are shown in bold and underlined.)

Error type	Main reason	Example
Missed character	misprint	**Bn**zene
Character replacement	OCR, misprint	B**c**nzene
Addition of a character	misprint	Benzene
Inversion of a pair of characters	misprint	b**ne**zene
Lost space or dash	OCR, misprint	**1c**hloropropane
Added space	OCR, misprint	1-chloro propane
Punctuation replacement	OCR, misprint	1.2-dichloroethane

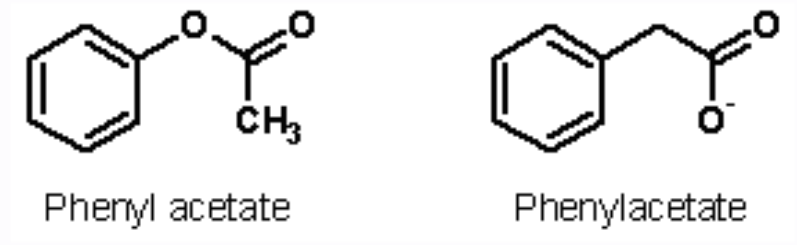
Automatic recognition and correction of these errors is a very important component of the chemical name conversion process. Based on available information, this procedure is implemented in the most flexible way in the CambridgeSoft Name⇔Struct program[48].

Table 3 Supported and unsupported automatic error recognition in Name>Struct.

Supported errors		**Unsupported errors**	
benz**io**c acid	*pair inversion*	benzoic ac**di**	*inversion - end or beginning*
benz**x**oic acid	*letter addition*	benzoic aci**de**	*addition - end or beginning*
benz**oo**ic acid	*double letter*	benzoic ac**i_**	*missed - end or beginning*
benzoic **a c**id	*space*	benzoic acif	*replaced - end or beginning*
bnzoic acid	*missed letter*	be**nn**zoic a**cc**id	*two errors in name*
benzoic ac**l**d	*replaced letter*		

Table 3 shows that Name⇔Struct supports four main types of errors inside chemical names – addition, deletion, replacement and pair inversion. For the conversion of names generated by OCR, the most common error is character replacement. For example, the name "heptane-2-car**6**oxylic acid" shown in Figure 1 and resulting from OCR cannot be converted to a structure.

Other common mistakes are due to the handling of punctuation and enclosing marks. While their presence is important, the replacement of one type by another generally does not affect the name analysis procedures. The same situation exists with the recognition of enclosing marks where the actual type of enclosing mark has no specific grammatical sense. A well-known exception is that a space is very important for the names of esters, as is shown in the simple example below.



The formatting of chemical names is generally not important. Whereas capitalization or italicization are essentially senseless, both sub- and superscripts are helpful in name analysis and in most cases the absence of formatting can be resolved simply by grammatical implementation. For example, Name⇔Struct successfully converts polycyclic names like **Tricyclo[3.3.1.11,5]decane** that according to nomenclature rules must be written as **Tricyclo[3.3.1.1^{1,5}]decane**. A good N2S engine therefore needs to be to be insensitive to both chemical name formatting and punctuation. This can generally be handled very efficiently using

name normalization procedures converting all punctuations into one type of separator and all enclosing marks into parentheses.

7. Problems Associated with Assembling Chemical Structures.

It could be assumed that the conversion of chemical names to their associated structures would conclude the task to provide the necessary data to a chemist for them to peruse. Unfortunately, the output from N2S engines can be in various formats including SMILES strings, InChI strings or one of a number of connection table formats. In order for a chemist to examine a structure, it must be represented in an interpretable graphical format with appropriate spatial configurations including bond angles, bond lengths, E/Z displacements and stereochemical centers. While the majority of chemical structure drawing packages integrated with N2S algorithms do include a "cleaning" algorithm, this process is extremely complex and there is no perfect procedure[58,59].

One specific issue that should be noted is the problem of over-determination of a structure, a circumstance that can arise when the generated structure is more specific than the initial chemical name. Part of this problem was described previously in the discussion regarding the conversion of ambiguous names. A particular problem concerns the assembly of chemical structure with the appropriate configuration of double bonds. As shown in Figure 11, the configuration of the nitrogen-nitrogen double bond is a *trans*-orientation but the source name did not contain this information. Most N2S engines generate such structures in this situation. In many cases omitted stereoconfigurations in the chemical name means that either the configuration is unknown or the sample contains a mixture of isomers. The most appropriate result would be to follow the IUPAC guideline for display in the recommended way[60] but such a

depiction is difficult for most procedures used to create "clean" structures. These algorithms remain an area of development for most drawing software development teams.

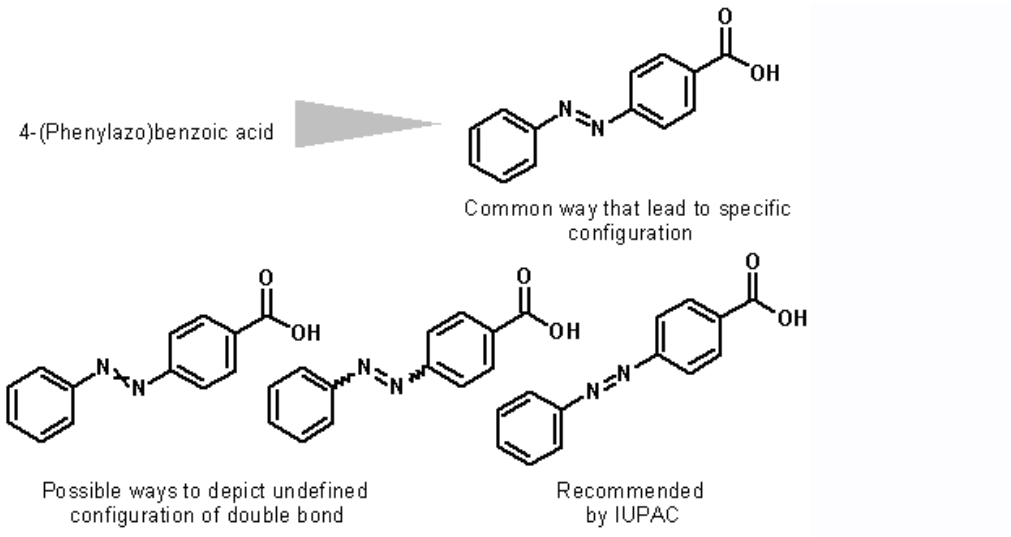


Figure 11. Graphical representations of undefined vs. defined double bond configurations.

Conclusions

In the near future we can be hopeful that the need to convert chemical names to chemical structures will be less important than we find at present. The ability to encapsulate the majority of organic molecules into an internationally accepted string representing a chemical structure already exists (see Chapter 5 regarding the InChI identifier) and publishers are starting to embed the InChI string directly into their articles to facilitate structure-based communication[61]. Software tools from a number of the commercial vendors can already search across chemical structures embedded in electronic documents and generate either PDF files[62] or image files with chemical structure information embedded directly into those files[63]. As the InChI identifier is extended to include other chemical structures of interest to the community (for example polymers, organometallics, inorganics and Markush) then the opportunity to further structure enable all electronic documents for searching is facilitated. As publishers initiate the inclusion of

structure-based tags associated with either chemical names or chemical structure depictions, the future of data-mining will require the coordinated extraction of information from documents containing chemical entities in both textual and graphical formats.

Until that time there remains a real need to continue the efforts to convert chemical identifiers, be they names or registry numbers, to their source chemical structures. As the optical recognition performance improves, and supporting technologies such as RECAPTCHA[64] contribute to the challenge of text digitization, then the conversion of chemical names will be limited only by the quality of the conversion algorithms and the appropriateness of the chemical names. The available name to structure conversion algorithms have already demonstrated value and are maturing in capability. The choice of accuracy versus throughput is one for the user. What these algorithms cannot resolve, however, is the potentially errors and ambiguities inherent to chemical names present in various documents and it is the authors opinion that moving forward future issues of this nature can only be resolved by adoption of structure identifier embedding inside the document (the suggested format being the InChI identifier), the unlikely development of improved nomenclature skills in all publishing chemists or preferably the adoption of electronic tools for the generation of high quality systematic names.

While NTS algorithms and other structure mining tools continue to improve there will likely be many opportunities for errors. Trusting the conversion of chemical names to a computer program without prior knowledge of the nature and quality of the input could be a recipe for disaster when handling publications and, based on our experience, especially when dealing with patents. NTS software is a very useful support aid at best but quality and curation remain the responsibility of the users of the software who are responsible for the generation of chemical information via application of the software.

References

1. IUPAC names are systematic names generated according to guidelines issued by the International Union of Pure and Applied Chemistry - http://www.iupac.org. An overview of their nomenclature efforts is provided online at http://en.wikipedia.org/wiki/IUPAC_nomenclature_of_organic_chemistry (accessed December 10, 2007).
2. CA index names are chemicals names issued according to the nomenclature standards of the Chemical Abstracts Service – http://www.cas.org.
3. CAS Registry Numbers are unique numerical identifiers for chemical compounds, polymers, biological sequences, mixtures and alloys. http://www.cas.org/newsevents/10digitrn.html (accessed December 11, 2007).
4. An EC-Number refers to a seven-digit code allocated by the Commission of the European Communities for commercially available chemical substances within the European Union. This EC-number replaces the previous EINECS and ELINCS numbers issued by the same organization. http://ecb.jrc.it/data-collection/ (accessed December 11, 2007).
5. The International Union of Pure and Applied Chemistry - http://www.iupac.org/dhtml_home.html (accessed December 11, 2007).
6. Recommendations on Organic & Biochemical Nomenclature, Symbols & Terminology http://www.chem.qmul.ac.uk/iupac/ ; The IUPAC Nomenclature Books Series http://www.iupac.org/publications/books/seriestitles/nomenclature.html.
7. An overview of the history of Optical Character Recognition - http://en.wikipedia.org/wiki/Optical_character_recognition (accessed December 8, 2007).

8. Proceedings of the First International Workshop on Text Mining in Bioinformatics (TMBio), Arlington, VA, USA. *BMC Bioinformatics.* 2007, 8**.** http://www.biomedcentral.com/1471-2105/8?issue=S9 (accessed December 25, 2007)

9. *ACD/Name;* Advanced Chemistry Development, Toronto, Ontario, Canada.

10. The CAS Registry Number and Substance Counts, Updated Daily: http://www.cas.org/cgi-bin/cas/regreport.pl (accessed December 24, 2007).

11. *The Beilstein structure database*; Beilstein GmbH, Germany.

12. *DiscoveryGate*®; Symyx, California, USA.

13. *The Spresi Database*; Infochem GmbH, Germany.

14. *VINITI*, (*Vserossiiskii institut nauchno-tekhnicheskoi informatsii*) (All-Russia Scientific Research Institute of Information), Moscow, Russia.

15. *Elsevier-MDL Patent Chemistry Database,* Elsevier, Reed-Elsevier, Amsterdam, The Netherlands. http://www.mdli.com/products/knowledge/patent_db/index.jsp (accessed December 26th 2007)

16. *Derwent World Patents Index*, Thomson Scientific, The Thomson Corporation. http://scientific.thomson.com/products/dwpi/ (accessed December 23 2007)

17. *The SureChem Portal*, SureChem Inc., San Francisco, CA, USA. http://www.surechem.org (accessed December 10 2007)

18. *IBM Chemical Search Alpha*, IBM, Almaden Services Research, San Jose, CA 95120, USA, https://chemsearch.almaden.ibm.com/chemsearch/SearchServlet.

19. Goncharoff, N.2007. SureChem – Free Access to Current, Comprehensive Chemical Patent Searching, Presentation given at the ICIC meeting, Infonortics, Barcelona, Spain..

20. Boyer, S. 2007. IBM, Almaden Services Research, San Jose, CA 95120, USA, Presentation given at the ICIC meeting, Infonortics, Barcelona, Spain. http://www.infonortics.com/chemical/ch04/slides/boyer.pdf (accessed December 3, 2007)
21. Advanced Chemistry Development, Toronto, Ontario, Canada http://www.acdlabs.com
22. Cambridgesoft, Boston, Massachusetts, USA http://www.cambridgesoft.com
23. Openeye, Santa Fe, New Mexico, USA http://www.eyesopen.com
24. Rhodes, J., Boyer, S., Kreulen, J., Chen, Y. and Ordonez, P. 2007. Mining Patents Using Molecular Similarity Search. In *Proceedings of the 12th Pacific Symposium on Biocomputing*. 12: 304-315.
25. An overview page regarding the IBM Chemical Search Alpha https://chemsearch.almaden.ibm.com/chemsearch/about.jsp
26. World Intellectual Property Organization, (WIPO). http://www.wipo.int/about-wipo/en/what_is_wipo.html (accessed December 29 2007).
27. Medline/Pubmed - http://www.ncbi.nlm.nih.gov/sites/entrez
28. SureChem, Nicko Goncharoff, Private Communication, (December 2007)
29. Accelrys, San Diego, CA, USA. http://www.accelrys.com (accessed December 16 2007)
30. Accelrys' Scitegic Pipeline Pilot ChemMining Collection: http://www.scitegic.com/products/chemmining/ChemMining.pdf (accessed December 15 2007).
31. Murray-Rust, P., Mitchell, J.B., and Rzepa, H.S. 2005. Communication and re-use of chemical information in bioscience. In *BMC Bioinformatics.* 6:180.

32. Corbett, P. and Murray-Rust, P. 2006. High-Throughput Identification of Chemistry in Life Science Texts. In *Lecture Notes in Computer Science: Computational Life Sciences II*. 107-118

33. PubChem: Information on biological activities of small molecules: http://pubchem.ncbi.nlm.nih.gov/ (accessed December 12 2007).

34. OPSIN, an Open Parser for Systematic Identification of Nomenclature: http://depth-first.com/articles/2006/10/17/from-iupac-nomenclature-to-2-d-structures-with-opsin (accessed December 10 2007).

35. MDL was acquired by Symyx in 2007. http://www.symyx.com/press_release.php?id=4&p=255 (accessed December 12 2007)

36. TEMIS S.A. Skill Cartridge Chemical Entity Relationships: http://www.temis.com/fichiers/t_downloads/file_109_Fact_sheet_TEMIS_Skill_Cartridge_Chemical_Entity_Relationships_En.pdf (accessed December 16 2007)

37. The National Center for Text Mining, United Kingdom: http://www.nactem.ac.uk/software.php?software=namedentity (accessed December 16 2007)

38. Optical Character Recognition, http://en.wikipedia.org/wiki/Optical_character_recognition (accessed December 10 2007)

39. Gifford-Fenton, E. and Duggan, H. N. Electronic Textual Editing: Effective Methods of Producing Machine-Readable Text from Manuscript and Print Sources http://www.tei-c.org/About/Archive_new/ETE/Preview/duggan.xml

40. OmniPage SE Version 2.0 by ScanSoft Inc. (ScanSoft Inc. has since merged with Nuance and assumed their name, http://www.nuance.com/company/) (accessed December 29 2007)

41. Hodge, G., Nelson, T. and Vleduts-Stokolov, N. 1989. Automatic Recognition of Chemical Names in Natural Language Text. Paper presented at the *198th American Chemical Society National Meeting*, Dallas, TX, April 7-9.

42. Kemp, N. and Lynch, M. F. 1994. The extraction of information from the text of chemical patents. 1. Identification of specific chemical names. In *Journal of Chemical Information and Computer Sciences*, 38(4), 544-551.

43. Corbett, P., Batchelor, C. and Teufel, S. 2007. Annotation of Named Chemical Entities, *BioNLP 2007: Biological, translational, and clinical language processing*, 57-64, Prague, Czech Republic

44. Copestake, A., Corbett, P., Murray-Rust, P., Rupp, C. J., Siddharthan, A., Teufel, S. and Waldron, B. 2006. An Architecture for Language Processing for Scientific Texts. *Proceedings of the UK e-Science Programme All Hands Meeting*, Nottingham, UK.

45. Garfield, E. 1961. Chemico-Linguistics: Computer Translation of Chemical Nomenclature. In Nature, 192-196

46. Vander Stouw**,** G. G., Naznitsky, I., and Rush, J. E., 1967. Procedures for converting systematic names of organic compounds into atom-bond connection tables. In *Journal of Chemical Documentation, 7*(3), 165–169

47. Vander Stouw, G. G., Elliott, P. M., and Isenberg, A. C. 1974.Automated conversion of chemical substance names into atom-bond connection tables. In *Journal of Chemical Documentation, 14*(3), 185–193

48. Structure=Name Pro version 11.0, Cambridgesoft, Boston, Massachusetts, USA. http://www.cambridgesoft.com/software/details/?ds=0&dsv=122 (accessed December 12 2007)

49. Brecher J. S., Method, system, and software for deriving chemical structural information, United States Patent 7054754, Issued on May 30, 2006

50. ACD/Name: Generate Structure From Name, Advanced Chemistry Development, Toronto, Ontario. http://www.acdlabs.com/products/name_lab/rename/tech.html (accessed December 16 2007)

51. NamExpert, ChemInnovation Software, San Diego, CA, USA. http://www.cheminnovation.com/products/nameexpert.asp (accessed December 16 2007)

52. Bonniot, D., IUPAC Naming. Presented at the ChemAxon UGM Meeting, Budapest, Hungary. http://www.chemaxon.com/forum/viewpost12244.html#12244 (accessed Dec 16 2007)

53. IC_{N2S}, InfoChem GmbH, Munchen, Germany. http://infochem.de/en/mining/icn2s.shtml (accessed December 14 2007)

54. Brecher, J.S. 1999. Name⇔Struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature. In *Journal of Chemical Information and Computer Sciences (JCICS)*, 39, 943-950.

55. Eller, G.A. 2006. Improving the Quality of Published Chemical Names with Nomenclature Software. In *Molecules*, 11, 915-928.

56. Thielemann, W. 2007. Information Extraction from Full-Text. *The International Conference in Trends for Scientific Information Professionals*, Barcelona, Spain.

http://www.infonortics.com/chemical/ch07/slides/thielemann.pdf (accessed December 6 2007)

57. ChemIDplus, National Library of Medicine, Maryland, USA, http://chem.sis.nlm.nih.gov/chemidplus/ (access December 16 2007)
58. Clarck A.M, Labute P. and Sabtavy M. 2006. 2D Structure Depiction. In *J. Chem. Inf. Comput. Sci.* 46, 1107-1123.
59. Fricker P.C, Gastreich M. and Rarey M. 2004. Automated Drawing of Structural Molecular Formulas under Constraints. In *J. Chem. Inf. Comput. Sci*. 44, 1065-1078
60. Brecher, J. S. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2007), Provisional recommendations. Available at: http://www.iupac.org/reports/provisional/abstract07/brecher_300607.html (accessed December 25 2007)
61. Project Prospect, Royal Society of Chemistry, Cambridge, UK. http://www.rsc.org/Publishing/Journals/ProjectProspect/Features.asp (accessed December 10 2007)
62. ACD/ChemSketch, Advanced Chemistry Development, Toronto, Ontario, Canada. http://www.acdlabs.com/products/chem_dsn_lab/chemsketch/features.html#Reporting (accessed December 10 2007)
63. Wikipedia Discussions Forum. 2007. Embedded InChIs in images (http://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemistry/Structure_drawing_workgroup/Archive_Jun_2007#ACD_ChemSketch_-_the_company_is_willing_to_make_a_Wikipedia_Template_on_their_Freeware)

64. RECAPTCHA™ - Digitizing Books One Word at a Time - http://recaptcha.net/learnmore.html (accessed December 16 2007)