



Enhancing the Discovery of Chemistry Theses by Registering Substances and Depositing in PubChem

Vincent F. Scalfani

Science and Engineering Librarian
Rodgers Library for Science and Engineering
The University of Alabama
Tuscaloosa, AL
vfscalfani@ua.edu

Barbara J. Dahlbach

Annex Services Librarian
Libraries' Annex
The University of Alabama
Tuscaloosa, AL
bdahlbac@ua.edu

Jacob Robertson

Institutional Repository Specialist
Gorgas Library
The University of Alabama
Tuscaloosa, AL
jrobertson3@ua.edu

Abstract

Chemical substances from theses are not widely accessible as searchable machine-readable formats. In this article, we describe our workflow for extracting, registering, and sharing chemical substances from the University of Alabama theses to enhance discovery. In total, 73 theses were selected for the project, resulting in about 3,000 substances registered using the IUPAC International Chemical Identifier and deposited in PubChem as either structure-data files or Simplified Molecular-Input Line-Entry System notations. In addition to substances being deposited in PubChem, an archive copy was also deposited in the University of Alabama Institutional Repository. The PubChem records for the substance depositions include the full bibliographic reference and link to the thesis full text or thesis metadata when the full text is not yet available. Excluding mixtures, we found that 40% of the shared substances were new to PubChem at the time of deposition. We conclude this article with a detailed discussion about our experiences, challenges, and recommendations for librarians and curators engaged in sharing chemical substance data from theses and similar documents.

Introduction and Literature Review

Chemical substances are core to the organization and retrieval of chemical information. Modern chemical databases contain millions of substances and associated data such as literature references, names, identifiers, and property data ([Krallinger et al. 2017](#)). Common to all chemical databases is the necessary step of substance data entry as a machine-readable format such as a chemical line notation or connection table ([Weininger 1988](#); [Dalby et al. 1992](#); [Heller et al. 2015](#)). After substance data entry, substances are registered ([Warr 2011](#)). Substance registration is an algorithmic process that uniquely identifies substances so that duplicates are not inadvertently entered into the database ([Buntrock 2001](#); [Gobbi & Lee 2012](#); [Martin et al. 2012](#)). Preventing duplicates ensures that associated known data, literature references, and identifiers can all be linked to one substance record. Computerized chemical substance registration systems have been developed over the past half century and are well documented in the literature. Examples of substance registration systems include the CAS REGISTRY from Chemical Abstracts Service ([Dittmar et al. 1976](#)), the Beilstein Registry File ([Domokos 1991](#)), Gmelin Factual Database ([Roth et al. 1992](#); [Lawson et al. 2014](#)), the PubChem Compound Standardization system ([Kim et al. 2016a](#); [Hähnke et al. 2018](#)) and systems based on the IUPAC International Chemical Identifier (InChI) such as UniChem and ChemSpider ([Chambers et al. 2013](#); [Heller et al. 2015](#); [Richardson 2018](#)).

Major efforts to extract and register substances into secondary chemical literature databases over the past half century have mainly focused on substances from journals, technical reports, and patents. To our knowledge, Reaxys ([Elsevier 2021](#)) and PubChem ([Kim et al. 2016b](#)) do not contain thesis bibliographic data nor substance data that is extracted and shared systematically from university theses. Both SciFinder ([Gabrielson 2018](#)) and ChemSpider ([Pence & Williams 2010](#)) contain data from the Selected Organic Reactions Database (SORD) ([Garritano 2013](#); [Royal Society of Chemistry 2020](#)). The SORD database efforts in the early 2000s partnered with academic institutions to gain access to thesis collections and index reaction and substance data ([Wife 2010](#)). The latest information we could find about SORD data content suggests that substance and reaction data were extracted from 1,300 theses in total, mostly from Europe ([Wife 2010](#)). Within SciFinder, there is SORD reaction data from about 900 theses ([Garritano 2013](#)). And within ChemSpider, a data source search for SORD reveals 57,000 substances ([Royal Society of Chemistry 2020](#)). The SORD thesis substance data extraction efforts appear to no longer be active, as we were unable to locate a live website or other current information about SORD. From the SORD efforts, it was estimated that 80% of content in university chemistry theses is never published and was termed “Lost Chemistry” ([de Laet et al. 2000](#); [Wife 2010](#)).

In addition to SORD content in CAS databases, CAS indexes thesis references (CAPlus) and registers substances from theses into the CAS REGISTRY ([Garritano 2013](#)). As of July 2020, CAS has about 670,000 thesis and dissertation records with substance and concept indexing in CAPlus (Chemical Abstracts Service, personal communication, July 13, 2020). In our experience with the chemistry theses at The University of Alabama (UA), the CAS REGISTRY typically contains two or less registered substances for each UA chemistry thesis. As chemistry theses with a synthetic focus typically contain many more than a handful of synthesized substances, there is an opportunity to increase thesis substance registration, sharing, and ultimately information discovery.

There are few reports in the literature focused specifically on extracting substance and related data from university chemistry theses. In fact, we are only aware of two such reports, one from Downing et al. ([2010](#)) and one from Andrews et al. ([2016](#)). Downing et al. developed automated text mining tools to extract chemical information such as names, experimental procedures, and characterization data directly from PDF/DOCX electronic theses. About 40 theses were used as a

proof of concept and the extracted data was arranged in markup language format to allow for data repository storage and semantic searching. There were many challenges with data cleanup and false hits, but overall, their approach validated the use of machine extraction of chemical data from theses with acceptable precision in some cases ([Downing et al. 2010](#)). It is worth noting that there is a large amount of related ongoing research and available tools in cheminformatics focused on automated data extraction from digital documents, including methods to optically recognize chemical substance diagrams and convert them to machine-readable format ([Filippov & Nicklaus 2009](#); [Valko & Johnson 2009](#); [Swain & Cole 2016](#); [Krallinger et al. 2017](#); [Nguyen et al. 2019](#)). However, the Downing et al. (2010) article is the only article we are aware of that systematically evaluates machine extraction of chemistry thesis data. We suspect many of these automated text and chemical substance recognition methods could certainly be applied to thesis documents, but the focus has been on other documents like patents and journals.

More recently, Andrews et al. (2016) reported on an initiative that extracted substances from United Kingdom chemistry theses and deposited the substances publicly in ChemSpider. After an initial evaluation of substance optical recognition software, Andrews et al. selected a manual extraction workflow largely due to the initial uncertainty of potential copyright restrictions associated with automated extraction and complexities associated with validating the accuracy of machine extracted substances. In total, about 45,000 substances from over 700 chemistry theses were manually extracted, redrawn, encoded in machine format and deposited publicly in ChemSpider ([Andrews et al. 2016](#)). Importantly, the bibliographic information was submitted along with the substances, which greatly enhances the discovery of the data as well as providing a provenance record for users. About 70% of the substances were new to ChemSpider at the time of deposition, which is close to the estimate of 80% “lost chemistry” from Wife ([2010](#)) and de Laet et al. ([2000](#)).

Interestingly, chemistry theses are rarely cited in the chemical literature. For example, an analysis of chemistry theses at Mississippi State University ([Zhang 2013](#)) and the University of Texas at Austin ([Flaxbart 2018](#)) found that citations to theses amounted to less than one percent of the total citations. Similar results were reported in a recent analysis of citations in ten different American Chemical Society journals where the citations for “other” information types, which includes theses, were found to be less than 5% ([Rose-Wiles & Marzabadi 2018](#)). The lack of discoverability and access to theses, including any substance content, could be one factor contributing to low chemistry thesis citation counts. We recognize this is a minor factor as peer-reviewed articles are the main information resource used in chemistry ([Flaxbart 2018](#)). Regardless, it is evident that chemistry theses contain useful data such as substances that are currently not electronically discoverable in chemical databases and, therefore, offer a unique opportunity for research libraries seeking to improve discoverability of chemical research at their institution.

The recent efforts by Andrews et al. (2016) to manually extract thesis substances and deposit them openly in ChemSpider was a workflow we envisioned could be adapted for research libraries; that is, subject chemistry librarians along with data repository librarians could extract and register substances from their local university chemistry theses and share them in public disciplinary repositories such as ChemSpider or PubChem. Such efforts would greatly enhance the discovery and utility of the theses, as users would have the ability to discover not only the text of the thesis, but the actual substances via standard chemical specific searches such as by molecular structure, formula, or identifier. As we began our own efforts to extract and register substances from UA theses, we quickly realized the various complexities of registering substance data, and the general lack of available detailed workflows and guidelines for the research library community. While the Andrews et al. (2016) report was helpful to think about the overall goals

and significance of the project, specific workflow details such as how to redraw the substances so that machines interpret them accurately, how to organize the substance data locally, or how to create substance-to-document links, was not discussed in detail.

In this article, we describe our workflow and results with registering nearly 3,000 substances from 73 UA theses. We manually extracted the substances, encoded them in machine-readable format, and shared the substances in PubChem openly with links to the original thesis document on our Institutional Repository (UA IR) or library record if full text was not available. In addition to sharing the data openly in PubChem, an archival copy of the substance data is available in the UA IR, and all data, programmatic scripts, and notes are openly available in GitHub (Supporting Information). We conclude this article with a discussion of the workflow challenges and our recommendations for librarians and curators related to registering and sharing substance data from university theses.

Methods

The following methods section was adapted from our GitHub repository README file ([Scalfani 2020](#)). The GitHub repository contains all data, scripts, and working notes from this project.

Theses Selected

A total of 73 UA Chemistry Ph.D. or M.S. theses were used. Theses selected were related to organic chemistry and contained synthetic details for small molecule preparations. All theses were not embargoed; theses selected were available for public use, either digitally via the UA IR or in print from the UA Libraries. Nearly all theses selected were from 1984 through 2019. The few exceptions include three theses from the 1960s and one thesis from 1929. The thesis date was not the primary selection criteria; theses were selected based on their organic chemistry content as we discovered them. About 30% of the theses were available as full-text PDFs. The full-text PDFs were mostly theses that were born-digital (post-2009); a few were retroactively digitally scanned.

Software Environment

All software and data analysis were run on Linux Ubuntu 18.04, with the exception of Bio-Rad KnowItAll ChemWindow 2018 ([Wiley Science Solutions 2020](#)), which was run on Windows 10. The open source RDKit cheminformatics software package v2019.09.2 ([Landrum 2020](#)) was installed in an Anaconda 3 Linux conda environment with Python 3.6.

Substance Selection

In general, substances selected from theses included those that could be represented using the machine-readable Simplified Molecular-Input Line-Entry System (SMILES) line notation ([Weininger 1988](#)). These substances included small molecule organic chemistry with some limited organometallic and coordination compounds. In addition, selected substances had associated synthetic preparatory procedures and experimental characterization details such as nuclear magnetic resonance spectroscopy, infrared spectroscopy, melting point, elemental analysis, or mass spectrometry data. In rare cases, substances selected for registration included only a quantitative analytical test. The preparatory procedures were typically specific to the substance, however in certain cases, substances were selected that had only general synthetic procedures associated with them; that is, when the same reaction was run across similar substrates. We avoided selecting substances where the synthetic preparatory method, as noted by the author, directly followed a prior reported literature preparation.

Substance Drawing

The majority of chemical substances were redrawn similarly to the depiction in the theses using ChemAxon MarvinSketch v19.27.0 ([ChemAxon 2019a](#)). Stereochemistry including double bond configuration and chiral centers were reproduced as originally defined. In cases where the substance name included racemic notation, (\pm), both enantiomers were drawn and included within one registry identifier. In rare situations where the author defined the stereochemistry drawn as absolute in the 2D depiction, but named the compound with relative notation symbols, R* or S*, the depiction was considered the correct absolute stereochemistry. When substances were drawn by an author with stereo non-specific wavy bonds, these were reproduced as drawn with the non-specific stereocenters, which is equivalent to plain bonds ([Brecher 2006](#)). However, when additional information was provided such that the final product was not an isolated stereoisomer, and instead an identified mixture of enantiomers or diastereomers, we drew both substance configurations and combined them into one registry identifier with two components. In cases where the diastereomeric mixture was not easily identifiable; that is, when it was not clear which stereocenter or bond to flip, or when the diastereomeric mixture was greater than two substance configurations, we drew those substances as stereo non-specific single component substances. Lastly, atropisomers were encoded as non-specific bonds.

For substances depicted as projections, special care was required to preserve the stereochemistry ([Brecher 2008](#); [Martin et al. 2012](#)). Haworth projections were manually converted to Mills skeletal depictions ([Brecher 2006](#)) and drawn in ChemAxon MarvinSketch. When substances were presented as chair conformations or Fischer projections, Bio-Rad's KnowItAll ChemWindow 2018 software was used to draw the structures and determine the stereochemistry automatically ([Abshear et al. 2018](#)).

Some substances (< 5% estimated) required adjustments to the original representation to maintain the correct hydrogen count and represent the structures within the limitations of chemical valence rules and cheminformatics file formats. These internal adjustments are described in Scalfani ([2020](#)). Our intention was to accurately maintain the author's original chemical structures as drawn. As such, we endeavored to keep these local business rules ([Hersey et al. 2015](#)) to a minimum, and instead rely on the well documented and established PubChem Compound standardization process to standardize the structures ([Hähnke et al. 2018](#)).

Generation of Machine-Readable Substance File Formats

For substances drawn in ChemAxon MarvinSketch, the representations were exported as ChemAxon SMILES (v19.27.0, Daylight variant). ChemAxon extended SMILES (CXSMILES) ([ChemAxon 2021](#)) were used for substances containing radicals or carbenes. Next, the SMILES were compiled in a spreadsheet along with the thesis bibliographic information and processed with RDKit v2019.09.2 using a custom Python script to generate a structure-date file (SDfile) ([Dalby et al. 1992](#)) containing the molecular representation connection table, a registry identifier (UALIB-1 and increasing sequentially), RDKit calculated Kekule SMILES, InChI (v1.05), thesis bibliographic reference, and link to the full-text thesis or library record. Any dative bonds were then added to the RDKit processed SDfile manually using the PubChem nonstandard bond syntax ([National Center for Biotechnology Information \[date unknown-b\]](#)). For substances drawn in KnowItAll ChemWindow 2018, the representations were exported as SMILES and InChI (v1.05) and compiled into a CSV spreadsheet with a substance registry identifier, thesis bibliographic reference, and link to the full-text thesis or library record, without any further local processing.

Registration and Consistency Check Using the Standard InChI

Standard InChIKeys (v1.05) were calculated from ChemAxon MarvinSketch exported SMILES using the ChemAxon command line program, Molconverter ([ChemAxon 2019b](#)) in a terminal window as follows: `$ molconvert -g "inchikey:SAbs,AuxNone" in.smi -o out.inchikey`

The InChI absolute stereochemistry, SAbs, option was used to force the calculation of a Standard InChI ([International Union of Pure and Applied Chemistry 2017](#)). Next, duplicate substances were checked against a main local registry file list of InChIKeys containing a local list of previously registered substances. This step was completed in a Unix terminal: `$ sort InChIKeys_list.inchikey | uniq --count --repeated`

The above command outputs a list of any duplicate InChIKeys. If any duplicates were identified, the duplicate substances were assigned the original registry identifier. The same sort/uniq command was used to check for duplicate substances with the InChIKeys generated from KnowItAll.

InChIKeys were also used as an interoperability check when transferring data between cheminformatics toolkits locally; that is, the InChIKeys generated from ChemAxon Molconverter were compared to RDKit generated InChIKeys for consistency ([Akhondi et al. 2012](#)).

PubChem Deposition

The RDKit generated SDfiles and KnowItAll compiled CSV spreadsheet files were submitted to PubChem for processing into the database through the PubChem Upload web interface. Our local registry file was then updated with the deposited PubChem Substance Identifier (SID) and standardized Compound Identifiers (CID).

Institutional Repository Archiving

After depositing the substance data in PubChem, an archive copy of the substance data in SDfile or CSV format was deposited in UA's DSpace Institutional Repository (UA IR). A new record was created for each collection of thesis substance data. Each UA IR record used the Dublin Core metadata schema with the following elements: dc.contributor, dc.date.issued, dc.description (includes a description of the substance data and CC-BY 4.0 license), dc.publisher, dc.relation.isbasedon (reference to original thesis), dc.title, dc.type, local.GitHub.URL, local.SDFPubChemExternalIDs.URL. The latter two local metadata elements provide cross links to the substance data on GitHub and PubChem.

New Substance Count Data Collection

To find the number of newly deposited substances in PubChem, the total number of substances (SIDs for same, mixtures, and all) linked to each of the UA deposited compound standardized records were retrieved. If there was only one associated SID, the structure was considered new to PubChem, and represents a new deposition. The data was programmatically collected using a script written in MATLAB R2020a. The MATLAB script uses the PubChem Power User Gateway web requests to retrieve the data and is detailed in a separate article ([Scalfani et al. 2020](#)) The substance count data was collected in May 2020 for each of the UA substances deposited in PubChem.

Results

Thesis Content

The variety of chemistry encountered in the selected 73 organic chemistry theses was diverse and reported substances synthesized included, for example, ionic liquids, natural products, carbene complexes, silyl compounds, furanones, ribose derivatives, and boronates. On average each thesis contained 39 synthesized substances with associated characterization data. By evaluating thesis titles, we estimated that ~200 of the theses at UA from 1924-2020 are in the organic chemistry subject domain and have a significant focus on small molecule synthesis, and as a result, our selected sample represents about 40% of suitable theses at UA for organic chemistry substance registration. About 30% of the theses selected are available in digital full-text format.

Substance Drawing and Machine-Readable File Creation

Using the workflow described in the methods section, a total of 2,885 unique substances were manually redrawn. The majority of structures (~94%) were drawn in ChemAxon MarvinSketch with the remaining substances, originally depicted as perspective representations, drawn in KnowItAll ChemWindow. For reference, if no challenges were encountered, we could typically draw 60 substances in about 3 hours, and then complete the remainder of the workflow in minutes.

Substance Registration and Interoperability Check with InChI

In our workflow, we tracked substances by their calculated Standard InChIKey. A compiled tabular list of InChIKeys and a unique identifier (e.g., UALIB-1) served as our internal registry list. If the calculated Standard InChIKey for a substance was unique, the substance was determined as new and added to our internal registry list. If the substance was identified as a duplicate InChIKey within our substance registry list, it was assigned the previously known registry identifier. In all of the substances we selected, there were 76 duplicates identified (2.6%) using the Standard InChI. For these substances, the result is that they have more than one associated thesis reference in our local registry identifier list.

The Standard InChI was also used to check the consistency of the chemical substance data exchange between the cheminformatics toolkits. SMILES and file format reading differences between toolkits are known to exist ([O'Boyle et al. 2018](#)), and since we were transferring ChemAxon generated SMILES to RDKit, comparing the individual toolkit calculated Standard InChIs was a convenient way to check for consistency ([Akhondi et al. 2012](#); [O'Boyle et al. 2018](#)). In total, there were 19 chemical substances (0.7%) where the calculated Standard InChIs did not match between the toolkits. We hypothesize that the differences are a result of a 2D drawing limitation ([Clark et al. 2006](#); [Frączek 2016](#)) leading to a different calculated InChI (there are no coordinates in SMILES). Out of caution, we submitted these 19 substances as ChemAxon molfiles, which includes the original 2D coordinates, directly to PubChem, without any local transfer to other toolkits. No critical differences were observed compared to the original ChemAxon SMILES to RDKit molfile derived submission after the PubChem standardization process. Compounds were standardized in PubChem Compound in the same manner.

PubChem Deposition

The substance data was deposited in PubChem through their PubChem Upload interface as either a SDfile or a CSV text file. After submission, it typically took 3-7 days for PubChem to process

the data and assign public PubChem SIDs from the Substance database along with the linked standardized PubChem CIDs from the Compound database. Each SID record in PubChem deposited by UA Libraries uses the External ID field to link to the full-text thesis in the UA IR or the catalog metadata record if the full text is not available yet (Figure 1).

We also added the full bibliographic citation of each thesis in the Depositor Comment Field. We notified PubChem staff that our depositions contained linked synthetic preparatory procedures in the original thesis reference. As a result, PubChem created a workflow on their end during the standardization process, which created a “Synthesis Reference” annotation from the bibliographic reference in the Depositor Comment field. The thesis reference is then displayed on the associated CID record page in the “Synthesis Reference” Literature section (Figure 2).

PubChem SID: 404338015

Structure: 

Source: [The University of Alabama Libraries](#)

External ID: [UALIB-913](#)

Source Category: Curation Efforts

Version: 1 [Revision History](#)

Status: Live



Related Compounds: PubChem CID: [CID 145865763](#) (3-[[4-(4-methoxy-N-(4-methoxyphenyl)anilino)phenyl]diazonyl]-N,N-bis(4-methoxyphenyl)aniline)



Dates: Deposit: 2020-02-08 Available: 2020-02-08

Please note that the substance record is presented as provided to PubChem by the source (depositor). For standardized chemical structure and/or annotation information, please visit the summary page for [CID 145865763](#).

[PubChem](#)

Figure 1. Example of a PubChem SID record with link to the UA Libraries

5 Literature  

5.1 Synthesis References  

Saint-Louis, C.J. Control of molecular geometries using new photo-electro-switchable azobenzenes. Ph.D. Thesis, The University of Alabama, 2015.

▼ The University of Alabama Libraries

Source: [The University of Alabama Libraries](#)

URL: <https://pubchem.ncbi.nlm.nih.gov/substance/404338015>

Description: Synthesized chemical substances from The University of Alabama Dissertations and Theses.

License URL: https://github.com/uilibweb/UALIB_ChemStructures/blob/master/README.md

Figure 2. Example of PubChem CID record showing the thesis reference in the literature section

Evaluation of Substances Deposited

Using PubChem programmatic web requests, we found that 1,461 (51%) of the UA thesis PubChem standardized compounds had only one associated substance identifier (SID), and were, therefore, new to the PubChem database at the time of deposition. PubChem Compound considers mixtures as unique and since our depositions include mixtures, the unique percentage of 51% may be slightly inflated. We had a total of 298 mixture submissions. If we assume that all of the mixtures had known individual components, this brings the new compound percentage down to 40%.

Discussion

Thesis Selection, Full Text Limitations, and Copyright Considerations

Theses containing organic, and some limited organometallic, substances are great candidates for substance data sharing as these molecules are most easily represented as machine-readable formats with available cheminformatics software ([Clark 2011](#); [Warr 2011](#); [Hähnke et al. 2018](#)). We, therefore, considered organic chemistry theses to be the priority area for substance registration and substance data sharing. The majority of the theses we identified at UA from 1924 through 2020 as having an organic chemistry focus (~200), were only available in print. As such, we considered and experimented with retrospective digital scanning of theses and deposition of the full text in the UA IR as permitted by copyright ([Copyright Advisory Network 2020](#)). However, this manual scanning process of theses was too time consuming and deemed not essential to the goals of the substance registration project. As each substance registered and shared would include the thesis bibliographic information, users discovering the substance data can contact UA Libraries for the full text.

It is our personal understanding as academic researchers, not lawyers, that according to the *Compendium of U.S. Copyright Office Practices* ([2017](#)), chemical substances are excluded from copyright protection. However, it is not clear to us if automated machine extraction of chemical substances would be considered copying the thesis content and a violation of the author's copyright. Andrews et al. ([2016](#)) had similar concerns with machine extraction in their thesis data extraction pilot. Given this uncertainty of machine extraction and copyright law, combined with the fact that most of our theses were only available in print, we had to use a manual substance extraction approach, which created the necessity for us to redraw all substance structures, as opposed to any automated substance machine-extraction techniques.

Experiences and Challenges with Substance Drawing

The majority of substances we encountered could be redrawn in ChemAxon MarvinSketch similarly to how they were depicted in the original thesis; that is, the subsequent export of the machine-encoded SMILES faithfully preserved the input structure atoms, bonds, connectivity, and stereochemistry. These “well-behaved” substances (> 90%) were substances that were drawn with organic chemistry 2D skeletal formulas, which followed, or at least loosely followed, graphical representation standards from IUPAC ([Brecher 2006](#); [Brecher 2008](#)). Some of the key features include using lines for bonds, omitting hydrogen atoms, atomic symbols for heteroatoms, plus or minus symbols for charges, and hashed or solid wedges/bonds for stereochemistry (Figure 3). These types of structures are most easily interpreted by cheminformatics software ([Brecher 2008](#); [Martin et al. 2012](#)).

We found that we were efficient with drawing structures in ChemAxon MarvinSketch; however, different structure editors can certainly be used and there are a variety of other editors available depending upon preferences such as ChemDraw or PubChem Sketcher ([Ihlenfeldt et al. 2009](#)).

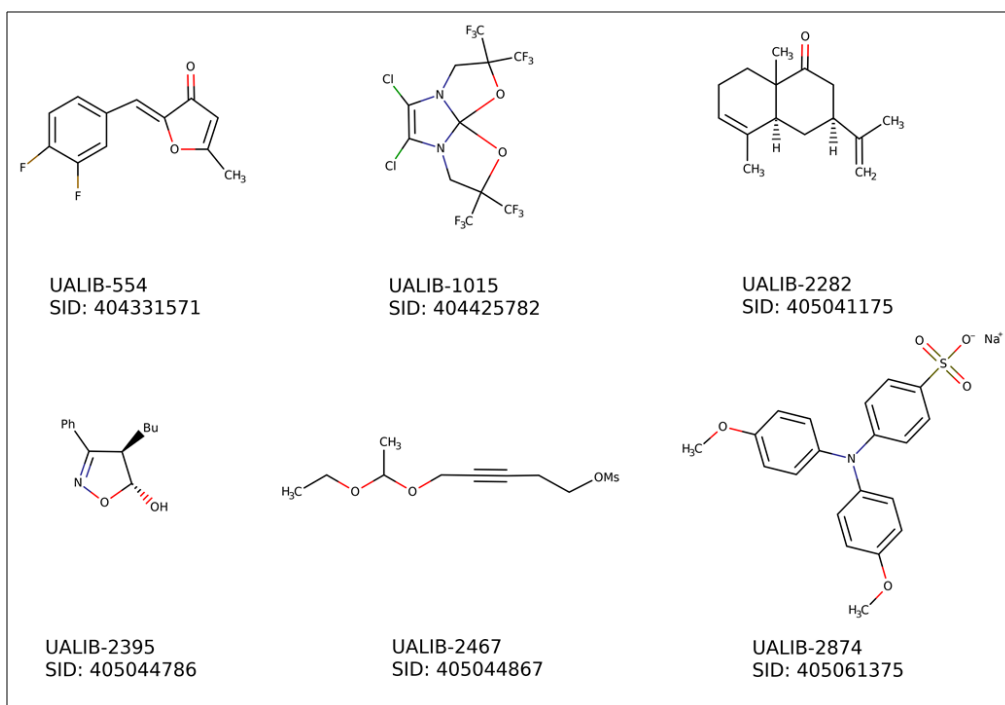


Figure 3. Examples of “well-behaved” substance drawings redrawn as depicted

Preserving stereochemistry from perspective chemical substance drawings (Figure 4) and handling stereochemical mixtures was the most challenging aspect of redrawing chemical substances. Perspective drawings including Haworth, chair, and Fischer projections are designed for humans and are generally not fully interpreted by most chemical drawing software tools, with the major limitation being loss of stereochemical information ([Brecher 2008](#); [Gobbi & Lee 2012](#); [Martin et al. 2012](#)). To our knowledge, the only consumer/academic software that can automatically assign stereochemistry in perspective drawings is the KnowItAll ChemWindow structure editor ([Abshear et al. 2018](#)). Given the software limitations of interpreting perspective drawings, we either had to manually infer the stereochemistry and redraw the structures with standard hash/solid wedges for stereochemistry or use the KnowItAll ChemWindow software to perceive the stereochemistry automatically. We found that it was most efficient for us to manually redraw Haworth projections as non-perspective drawings in MarvinSketch. However, for the chair and Fischer projections, it was faster for us to draw these in ChemWindow than to manually perceive the stereochemistry.

Ideally, for stereochemical mixtures including racemic, enantiomers, and diastereomers, we would use a file format such as molfile V3000 or ChemAxon Extended SMILES that support relative configuration of stereocenters ([Gobbi & Lee 2012](#); [Martin et al. 2012](#); [ChemAxon 2021](#)). However, PubChem does not support relative stereochemistry or defined mixtures of stereoisomers as a single structure. Support for enhanced stereochemistry is technically possible and defined within the PubChem stereochemistry specification, but this feature is not currently supported ([National Center for Biotechnology Information \[date unknown-a\]](#)). Further, we selected the Standard InChI as our local substance uniqueness check, and this process considers the substances as only absolute stereochemistry. As a result of these limitations, using file formats that support enhanced stereochemistry was not an option for us and we instead represented stereochemical mixtures including racemates, enantiomers with any ratio, and diastereomers within one registry identifier as separate disconnected substances. Such

representation limitations within chemical databases are discussed by Hersey et al. (2015), and there is, unfortunately, not currently an accepted standard across public databases for how to represent stereochemical mixtures; some sources choose to draw racemic mixtures as one substance with no stereochemistry, while others draw multiple enantiomers or diastereomers in

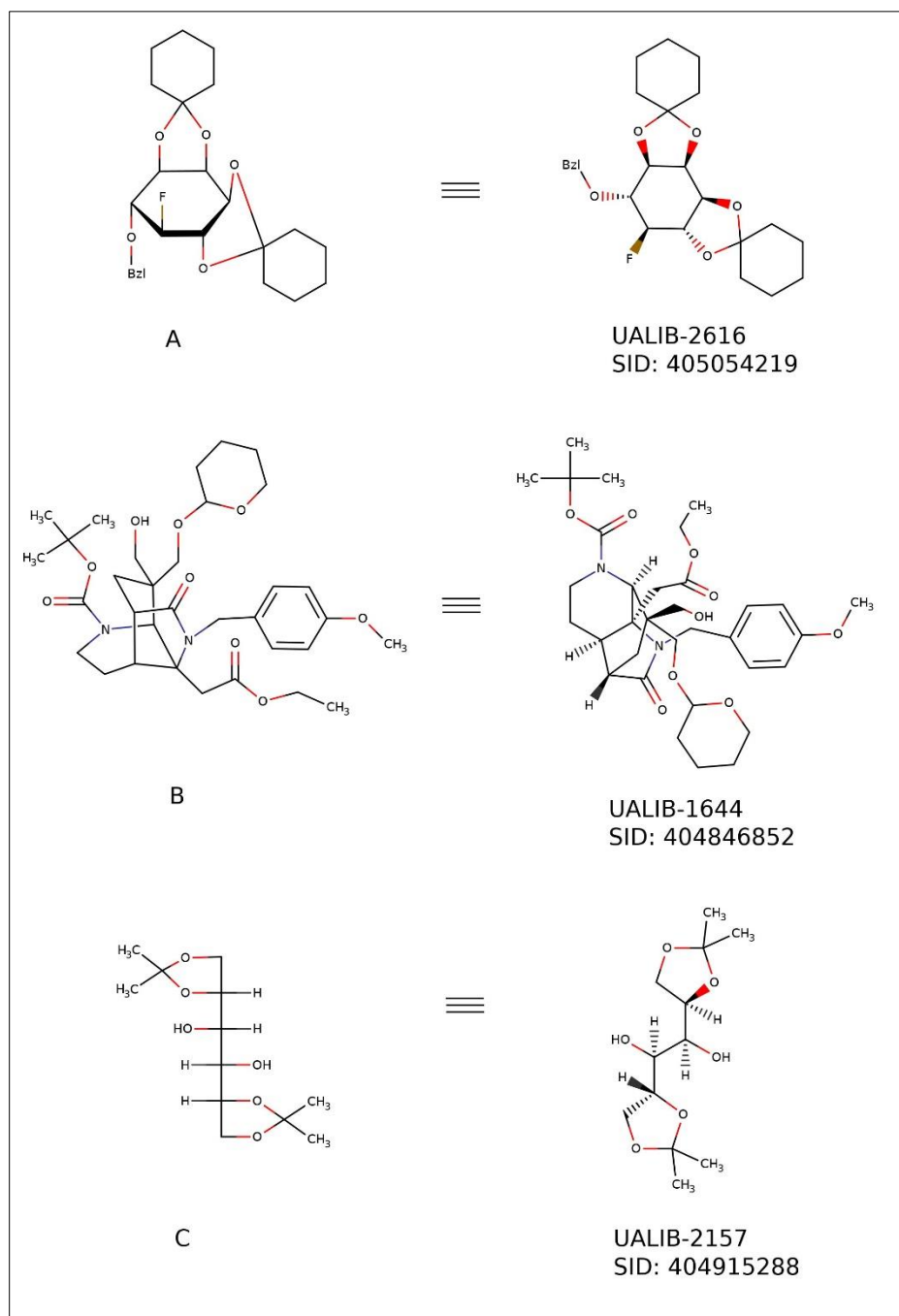


Figure 4. Example Haworth (A) , chair (B), and Fischer projections (C) (adapted from Martin et al. (2012))

one record (Food and Drug Administration 2007; Hersey et al. 2015). Lastly, drawing multiple substances in a record creates a way to describe molecules by the AND operator, for example: (2S)-2-bromobutane AND (2R)-2-bromobutane. It is unclear how to best represent a substance with a defined “OR” scenario within one registry identifier in public databases, without using extended stereochemistry file formats, such as in the case of (2S)-2-bromobutane OR (2R)-2-bromobutane.

Machine-Readable File Creation Experiences and Recommendations

As noted above, most of the substance representations were processed using the RDKit to create SDfiles. We selected RDKit because of its strong integration with the Python programming language and our familiarity with it. RDKit is a cheminformatics toolkit and does not contain a graphical structure editor. As such, this required drawing the structures in a separate program, in our case ChemAxon MarvinSketch, and then transferring the molecular representation data to RDKit. In hindsight, processing the chemical substance SMILES data in a separate cheminformatics toolkit to create an SDfile was unnecessary for data sharing in PubChem and necessitated the incorporation of a local data interoperability check using InChI. A more efficient approach is to compile the molecular representation data as SMILES along with the thesis bibliographic information in a spreadsheet application and then submit this file directly to PubChem, as we did in the case of substances drawn with KnowItAll ChemWindow; that is, the same spreadsheet workflow could have been used for substances drawn in ChemAxon MarvinSketch. The major limitation with submitting a spreadsheet of SMILES chemical representations to PubChem is that, to our knowledge, it is not possible to specify PubChem nonstandard bonds such as dative bonds defined in PubChem substance tags ([National Center for Biotechnology Information \[date unknown-b\]](#)) within the spreadsheet or represent features such as radicals, as these SMILES extensions are not recognized by PubChem. In these specific cases, a molfile/SDfile representation format would need to be used for PubChem submissions. Such a task can still be completed within a single toolkit, as both MarvinSketch and KnowItAll can export molfile/SDfile formats. Finally, it should be noted that there are other differences and limitations with molfile/SDfile encoded molecular representations compared to SMILES and this may be a consideration when submitting data to PubChem ([Daylight Chemical Information Systems 2011](#); [Dassault Systemes 2017](#)). However, as SMILES contain the entire representation on one line, we found SMILES much more convenient to work with compared to molfile/SDfiles.

InChI Algorithm for Substance Registration

The use of InChI was critical to our substance registration process, as well as being useful as a consistency check within our overall workflow. InChI is an open non-proprietary chemical identifier, which is well supported across cheminformatics software. The InChI algorithm is currently used to check for structure uniqueness in several public chemical databases and cross-referencing services such as ChEMBL ([Mendez et al. 2018](#); [Hersey \[date unknown\]](#)), ChEBI ([Chambers et al. 2013](#); [Hastings et al. 2016](#)), ChemSpider ([Richardson 2018](#)), and UniChem ([Chambers et al. 2013](#)). Combined with the ability to compare InChIs across cheminformatics toolkits and the established record of using InChI as a uniqueness check, InChI proved to be a great choice for our structure uniqueness check.

There are different levels of uniqueness that InChI can describe, depending on if a standard or non-standard InChI is calculated. Standard InChIs, for example, are tautomer independent, represent organometallics with disconnected metals, and only support absolute stereochemistry. These limitations can be overcome by calculating a non-standard InChI, which allows for specific options related to tautomers, metal representation, stereochemistry and more ([Heller et al. 2015](#)). Both the Standard InChI and non-standard InChI are suitable choices for checking the uniqueness of chemical substances. ChEMBL and UniChem use the Standard InChI, ([Chambers et al. 2013](#); [Hersey \[date unknown\]](#)) while ChemSpider (Royal Society of Chemistry, personal communication, July 24, 2020) and ChEBI use a non-standard InChI ([Chambers et al. 2013](#); [Hastings et al. 2016](#)). Chambers et al. (2013) argue that the community considers the Standard InChI to be an acceptable measure of substance uniqueness relevant to chemical biology and drug discovery. Ultimately, we selected the Standard InChI because of the primary consideration

of data reuse; that is, since all of our data, including intermediate working files and registry lists are public, we felt it was best to share Standard InChIs for data exchange considerations.

PubChem Data Sharing, Provenance, and Access

At the time that we submitted the UA thesis substances, PubChem had 103 million unique compounds. As such, the fact that 40% or more of our contributions were new to the database and unique from the already present 103 million compounds is highly significant, and we believe supports the claim that contributing substances from university theses is valuable to the community. For duplicate substances submitted, there is still value as the data is merged with other records in PubChem and adds a new bibliographic reference to the record.

There are many steps involved in sharing chemical substance data from theses and with that comes many opportunities for data loss or corruption. No matter how careful data depositors are locally, there is still a possibility that any of the substances shared in PubChem could be interpreted differently after being processed with their selected cheminformatics software and standardization workflow ([Hähnke et al. 2018](#)). To evaluate how our substance representations changed after PubChem deposition, we compared our locally computed Standard InChIs for all substances that passed PubChem standardization to the PubChem Compound Standard InChIs and found that 150 (5.2%) of the substances did not have identical InChIs after PubChem processing, suggesting a possible change in structure representation. Chemical substance interpretation differences such as a stereochemistry loss or hydrogen count disagreement highlights the importance of maintaining provenance to the original data and link to bibliographic record. While we can endeavor to limit errors (~95% precision based on Standard InChI comparison), ultimately the end user should always validate the data with the original source.

One of the biggest advantages of depositing data in PubChem is that users can now search for UA thesis substance data with chemical specific search query options, such as by chemical structure, substructure, molecular formula, and identifier. Notably, there is limited information available about chemists' use of PubChem as citations to databases in the literature are rare ([Tomaszewski 2019](#)). In a recent information seeking behavior study of chemists, however, it was found that about 17 percent of the chemists surveyed use PubMed ([Gordon et al. 2018](#)), which is closely integrated with PubChem. Moreover, throughout 2020, PubChem had between two and four million unique users per month ([Kim et al. 2021](#)).

Full access to the UA thesis substance data is available through PubChem via the web interface or any of their programmatic interfaces such as PUG-REST ([Kim et al. 2015](#)), PUG-VIEW ([Kim et al. 2019](#)) or E-Utilities ([National Center for Biotechnology Information 2021](#)). We recommend accessing UA thesis substance data through PubChem, since PubChem standardizes the data and combines the data with related information. However, to maintain the provenance of the substance data, and allow users to validate the data, there is a link from the Source field in our PubChem deposited data directly to our UALIB_ChemStructures GitHub repository which contains notes about reuse (CC-BY 4.0), the original substance data files, and thesis bibliographic reference.

Another advantage of depositing substance data in PubChem is the ability to update records. If the updated data is submitted with the original registry identifier, PubChem will maintain substance record versioning and reprocess the data into PubChem Compound. This is important, and allows us to update our substance records, for example, as we become aware of errors or need to update a bibliographic reference link. Moreover, we expect to submit updates as

cheminformatics file formats improve and as our workflows and understanding of how to handle chemical representation increases.

Cost versus Benefit Considerations

A reasonable question to consider is what is the cost versus benefit of spending the time to extract, register, and share substances retrospectively from theses? It is a hard question to answer, but we do have a couple of supporting quantitative data points. For example, we found that at least 40% of the substances we shared were new to PubChem. The 40% we found is less than the 70% reported by Andrews et al. (2016) for new substances deposited to ChemSpider from UK theses; however, it is still a large percentage of new substances deposited. There is also a potential to quantify any increased web traffic views of UA theses with substances shared versus theses that do not have their substances shared in machine-readable format. We hope to have some meaningful usage data to analyze after a few years, which should provide a reasonable time frame for discovery of the new data in PubChem.

More broadly, theses represent the history of the research at an institution (Scalfani 2017), and we feel strongly that one of the most important tasks a librarian should engage in is to help promote, share, and preserve their institutions' research for others to discover and build upon.

We acknowledge that a significant time investment will be necessary for the workflow setup and becoming familiar with chemical structures, software, and machine-readable chemical file formats. However, after a workflow is set up similarly to that described in this article, the actual process of redrawing structures and sharing them is reasonable and practical to incorporate within regular liaison workloads. With a bit of practice, we were able to complete an entire thesis with 60 substances in about 3 hours.

Conclusion

We successfully implemented a workflow to manually redraw chemical substances from UA theses and share them in machine-readable format in PubChem. The main workflow used a combination of ChemAxon MarvinSketch and RDKit to create a machine-readable SDfile containing the substance connection tables, SMILES, InChI and bibliographic reference. The greatest challenge was the manual redrawing of the chemical substances, particularly when encountering perspective drawings and stereochemical mixtures. In total, about 3,000 chemical substances from 73 UA theses were shared. At least 40% of the substances were new to PubChem at the time of deposition. Substance depositions in PubChem include the full thesis bibliographic information and link to the thesis full-text PDF or metadata record if the digital full text is not yet available. Users can now discover UA theses in PubChem using specific chemical literature search strategies like molecular formula, structure, and identifier searches.

For librarians and curators seeking to share chemical substance data from theses, it is necessary to first become familiar with chemical file formats and their limitations. It will take time to register and share a significant amount of retrospective thesis substance data from research libraries; however, we are hopeful that this article will help stimulate interest among chemistry librarians and support the idea that enhancing the discovery of theses is of value to the community and profession.

Supporting Information

GitHub Repository: https://github.com/ualibweb/UALIB_ChemStructures – includes all working substance data files, programmatic scripts, and notes.

PubChem Data: <https://pubchem.ncbi.nlm.nih.gov/source/15645> – includes substance data submitted to PubChem (SIDs) and standardized PubChem data (CIDs)

Institutional Repository Archived Data: <https://ir.ua.edu/> - includes a copy of the original machine-readable files submitted to PubChem.

Acknowledgments

VFS thanks Jeweles Moton, Rachel Humphrey, Donald Williams, and Mary Alexander for help with early data collection and workflows. We thank ChemAxon for the MarvinSketch academic research license, Bio-Rad for the KnowItAll academic license (now Wiley Science Solutions), and all contributors to RDKit. VFS acknowledges The University of Alabama Libraries for granting research leave for this work. Finally, we thank NIH/NLM/NCBI PubChem staff for their helpful responses related to chemical file formats and data submission questions.

References

Abshear, T., Banik, G., Dalvi, S., D'Souza, M., Kunitsky, K. & Nedwed, K. 2018. Validation of the KnowItAll stereochemistry toolkit: Tech note 210434. Philadelphia (PA): Bio-Rad Laboratories.

Akhondi, S.A., Kors, J.A. & Muresan, S. 2012. Consistency of systematic chemical identifiers within and between small-molecule databases. *Journal of Cheminformatics* 4:35. DOI: [10.1186/1758-2946-4-35](https://doi.org/10.1186/1758-2946-4-35).

Andrews, D.M., Broad, L.M., Edwards, P.J., Fox, D.N.A., Gallagher, T., Garland, S.L., Kidd, R. & Sweeney, J.B. 2016. The creation and characterisation of a National Compound Collection: The Royal Society of Chemistry pilot. *Chemical Science* 7(6):3869–3878. DOI: [10.1039/C6SC00264A](https://doi.org/10.1039/C6SC00264A).

Brecher, J. 2006. Graphical representation of stereochemical configuration - (IUPAC recommendations 2006). *Pure and Applied Chemistry* 78(10):1897–1970. DOI: [10.1351/pac200678101897](https://doi.org/10.1351/pac200678101897).

Brecher, J. 2008. Graphical representation standards for chemical structure diagrams. *Pure and Applied Chemistry* 80(2):277–410. DOI: [10.1351/pac200880020277](https://doi.org/10.1351/pac200880020277).

Buntrock, R.E. 2001. Chemical registries in the fourth decade of service. *Journal of Chemical Information and Computer Sciences* 41(2):259–263. DOI: [10.1021/ci000109q](https://doi.org/10.1021/ci000109q).

Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S. & Overington, J.P. 2013. UniChem: A unified chemical structure cross-referencing and identifier tracking system. *Journal of Cheminformatics* 5:3. DOI: [10.1186/1758-2946-5-3](https://doi.org/10.1186/1758-2946-5-3).

ChemAxon. 2019a. MarvinSketch v19.27.0 [Internet]. [cited 2021 Jan 13]. Available from <https://chemaxon.com>.

ChemAxon. 2019b. Molconverter v19.27.0 [Internet]. [cited 2021 Jan 13]. Available from <https://chemaxon.com>.

ChemAxon. 2021. Extended SMILES and SMARTS - CXSMILES and CXSMARTS [Internet]. [cited 2021 Apr 15]. <https://docs.chemaxon.com/display/docs/chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts.md>.

Clark, A.M. 2011. Accurate specification of molecular structures: The case for zero-order bonds and explicit hydrogen counting. *Journal of Chemical Information and Modeling* 51(12):3149–3157. DOI: [10.1021/ci200488k](https://doi.org/10.1021/ci200488k).

Clark, A.M., Labute, P. & Santavy, M. 2006. 2D structure depiction. *Journal of Chemical Information and Modeling* 46(3):1107–1123. DOI: [10.1021/ci050550m](https://doi.org/10.1021/ci050550m).

Copyright Advisory Network. 2020. Public Domain Slider [Internet]. Available from <https://librarycopyright.net/>.

Dalby, A., Nourse, J.G, Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A. & Laufer, J. 1992. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of Chemical Information and Modeling* 32(3):244–255. DOI: [10.1021/ci00007a012](https://doi.org/10.1021/ci00007a012).

Dassault Systemes. 2017. BIOVIA CTFE formats: BIOVIA databases [Internet]. [accessed August 2020]. Available from http://help.accelrys.com/ulm/onelab/1.0/content/ulm_pdfs/direct/reference/ctfileformats2016.pdf.

Daylight Chemical Information Systems. 2011. Daylight theory manual v4.9 [Internet]. Available from <https://www.daylight.com/dayhtml/doc/theory/>.

de Laet, A., Hehenkamp, J.J.J. & Wife, R.L. 2000. Finding drug candidates in virtual and lost/emerging chemistry. *Journal of Heterocyclic Chemistry* 37(3):669–674. DOI: [10.1002/jhet.5570370324](https://doi.org/10.1002/jhet.5570370324).

Dittmar, P.G., Stobaugh, R.E. & Watson, C.E. 1976. The Chemical Abstracts Service chemical registry system. I. General design. *Journal of Chemical Information and Computer Sciences* 16(2):111–121. DOI: [10.1021/ci60006a016](https://doi.org/10.1021/ci60006a016).

Domokos, L. 1991. The Beilstein Structure Registry System. 1. General design. *Journal of Chemical Information and Modeling* 31(2):320–326. DOI: [10.1021/ci00002a019](https://doi.org/10.1021/ci00002a019).

Downing, J., Harvey, M.J., Morgan, P.B., Murray-Rust, P., Rzepa, H.S., Stewart, D.C., Tonge, A.P. & Townsend, J.A. 2010. SPECTRa-T: Machine-based data extraction and semantic searching of chemistry e-theses. *Journal of Chemical Information and Modeling* 50(2):251–261. DOI: [10.1021/ci9003688](https://doi.org/10.1021/ci9003688).

Elsevier. 2021. Reaxys content. [Internet]. [cited 2021 Jan 13]. Available from <https://www.elsevier.com/solutions/reaxys/features-and-capabilities/content>.

Filippov, I.V. & Nicklaus, M.C. 2009. Optical Structure Recognition Software to recover chemical information: OSRA, an open source solution. *Journal of Chemical Information and Modeling* 49(3):740–743. DOI: [10.1021/ci800067r](https://doi.org/10.1021/ci800067r).

Flaxbart, D. 2018. Analysis of citations to books in chemistry PhD dissertations in an era of transition. *Issues in Science and Technology Librarianship*. 88. DOI: [10.5062/F4DV1H4T](https://doi.org/10.5062/F4DV1H4T).

Food and Drug Administration. 2007. Substance registration system standard operating procedure [Internet]. Available from <https://www.fda.gov/media/75274/download>.

Frączek, T. 2016. Simulation-based algorithm for two-dimensional chemical structure diagram generation of complex molecules and ligand–protein interactions. *Journal of Chemical Information and Modeling* 56(12):2320–2335. DOI: [10.1021/acs.jcim.6b00391](https://doi.org/10.1021/acs.jcim.6b00391).

Gabrielson, S.W. 2018. SciFinder. *Journal of the Medical Library Association* 106(4):588–590. DOI: [10.5195/JMLA.2018.515](https://doi.org/10.5195/JMLA.2018.515).

Garritano, J.R. 2013. Evolution of SciFinder, 2011–2013: New features, new content. *Science & Technology Libraries* 32(4):346–371. DOI: [10.1080/0194262X.2013.833068](https://doi.org/10.1080/0194262X.2013.833068).

Gobbi, A. & Lee, M-L. 2012. Handling of tautomerism and stereochemistry in compound registration. *Journal of Chemical Information and Modeling* 52(2):285–292. DOI: [10.1021/ci200330x](https://doi.org/10.1021/ci200330x).

Gordon, I.D., Meindl, P., White, M. & Szigeti, K. 2018. Information seeking behaviors, attitudes, and choices of academic chemists. *Science & Technology Libraries* 37(2):130–151. DOI: [10.1080/0194262X.2018.1445063](https://doi.org/10.1080/0194262X.2018.1445063).

Hähnke, V.D., Kim, S. & Bolton, E.E. 2018. PubChem chemical structure standardization. *Journal of Cheminformatics* 10:36. DOI: [10.1186/s13321-018-0293-8](https://doi.org/10.1186/s13321-018-0293-8).

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. & Steinbeck, C. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research* 44(D1):D1214–D1219. DOI: [10.1093/nar/gkv1031](https://doi.org/10.1093/nar/gkv1031).

Heller, S.R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. 2015. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* 7:23. DOI: [10.1186/s13321-015-0068-4](https://doi.org/10.1186/s13321-015-0068-4).

Hersey, A. [date unknown]. ChEMBL database: Meeting chemical and biological information needs of scientists of the future [Internet]. Available from https://www.rsc.org/images/ChEMBL-anne-hersey_tcm18-213324.pdf.

Hersey, A., Chambers, J., Bellis, L., Bento, A.P., Gaulton, A. & Overington, J.P. 2015. Chemical databases: Curation or integration by user-defined equivalence? *Drug Discovery Today Technology* 14:17–24. DOI: [10.1016/j.ddtec.2015.01.005](https://doi.org/10.1016/j.ddtec.2015.01.005).

Ihlenfeldt, W.D., Bolton, E.E. & Bryant, S.H. 2009. The PubChem chemical structure sketcher. *Journal of Cheminformatics* 1:20. DOI: [10.1186/1758-2946-1-20](https://doi.org/10.1186/1758-2946-1-20).

International Union of Pure and Applied Chemistry. 2017. International chemical identifier (InChI) version 1, software version 1.05 API reference [Internet]. Available from <https://www.inchi-trust.org/downloads/>.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. 2021. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research* 49(D1):D1388–D1395. DOI: [10.1093/nar/gkaa971](https://doi.org/10.1093/nar/gkaa971).

Kim, S., Thiessen, P.A., Bolton, E.E. & Bryant, S.H. 2015. PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. *Nucleic Acids Research* 43(W1):W605–W611. DOI: [10.1093/nar/gkv396](https://doi.org/10.1093/nar/gkv396).

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. 2016a. PubChem Substance and Compound databases. *Nucleic Acids Research* 44(D1):D1202–D1213. DOI: [10.1093/nar/gkv951](https://doi.org/10.1093/nar/gkv951).

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., et al. 2016b. Literature information in PubChem: Associations between PubChem records and scientific articles. *Journal of Cheminformatics* 8:32. DOI: [10.1186/s13321-016-0142-6](https://doi.org/10.1186/s13321-016-0142-6).

Kim, S., Thiessen, P.A., Cheng, T., Zhang, J., Gindulyte, A. & Bolton, E.E. 2019. PUG-View: Programmatic access to chemical annotations integrated in PubChem. *Journal of Cheminformatics* 11:56. DOI: [10.1186/s13321-019-0375-2](https://doi.org/10.1186/s13321-019-0375-2).

Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. 2017. Information retrieval and text mining technologies for chemistry. *Chemical Reviews* 117(12):7673–7761. DOI: [10.1021/acs.chemrev.6b00851](https://doi.org/10.1021/acs.chemrev.6b00851).

Landrum, G.A. 2020. RDKit: Open-source cheminformatics software [Internet]. Available from <https://www.rdkit.org/>.

Lawson, A.J., Swienty-Busch, J., Géoui, T. & Evans, D. 2014. The making of Reaxys—Towards unobstructed access to relevant chemistry information. In: McEwen, L.R. & Buntrock, R.E., editors. *The Future of the History of Chemical Information*. Washington (DC): American Chemical Society. p. 127–148.

Martin, E., Monge, A., Duret, J-A., Gualandi, F., Peitsch, M.C. & Pospisil, P. 2012. Building an R&D chemical registration system. *Journal of Cheminformatics* 4:11. DOI: [10.1186/1758-2946-4-11](https://doi.org/10.1186/1758-2946-4-11).

Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. 2018. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research* 47(D1):D930–D940. DOI: [10.1093/nar/gky1075](https://doi.org/10.1093/nar/gky1075).

National Center for Biotechnology Information. 2021. Entrez programming utilities help [Internet]. Available from <https://www.ncbi.nlm.nih.gov/books/NBK25501/>.

National Center for Biotechnology Information. [date unknown-a]. PubChem specification: PC-StereoGroup [Internet]. [accessed 2020 Jul 2]. Available from https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/asn_spec/PC-StereoGroup.html.

National Center for Biotechnology Information. [date unknown-b]. PubChem substance tags [Internet]. [accessed 2020 May 18]. Available from https://pubchem.ncbi.nlm.nih.gov/upload/html/tags_substance.html.

Nguyen, A., Huang, Y-C., Tremouilhac, P., Jung, N. & Bräse, S. 2019. ChemScanner: Extraction and re-use(ability) of chemical information from common scientific documents containing ChemDraw files. *Journal of Cheminformatics* 11:77. DOI: [10.1186/s13321-019-0400-5](https://doi.org/10.1186/s13321-019-0400-5).

- O'Boyle, N.M., Mayfield, J.W. & Sayle, R.A. 2018. Can we agree on the structure represented by a SMILES string? A benchmark dataset [Internet]. Available from https://www.nextmovesoftware.com/products/SMILESBenchmark_ICCS_May2018.pdf.
- Pence, H.E. & Williams, A. 2010. ChemSpider: An online chemical information resource. *Journal of Chemical Education* 87(11):1123–1124. DOI: [10.1021/ed100697w](https://doi.org/10.1021/ed100697w).
- Richardson, S. 2018. ChemSpider pre-deposition filters [Internet]. Available from <https://blogs.rsc.org/chemspider/2018/09/18/chemspider-pre-deposition-filters/>.
- Rose-Wiles, L.M. & Marzabadi, C. 2018. What do chemists cite? A 5-year analysis of references cited in American Chemical Society journal articles. *Science & Technology Libraries* 37(3):246–273. DOI: [10.1080/0194262X.2018.1481488](https://doi.org/10.1080/0194262X.2018.1481488).
- Roth, B., Böhmer, H-U. & Deplanque, R. 1992. Registration of substances in the Gmelin Factual Database. *Analytica Chimica Acta* 265(2):301–304. DOI: [10.1016/0003-2670\(92\)85036-6](https://doi.org/10.1016/0003-2670(92)85036-6).
- Royal Society of Chemistry. 2020. ChemSpider data source search: SORD [Internet]. [cited 2020 May 12]. Available from <https://www.chemspider.com/Search.aspx?dsn=SORD>.
- Scalfani, V.F. 2017. Text analysis of chemistry thesis and dissertation titles. *Issues in Science and Technology Librarianship* 86. DOI: [10.5062/F4TD9VBX](https://doi.org/10.5062/F4TD9VBX).
- Scalfani, V.F. 2020. UALIB_ChemStructures GitHub repository [Internet]. Available from https://github.com/ualibweb/UALIB_ChemStructures/blob/master/README.md.
- Scalfani, V.F., Ralph, S.C., Alshaikh, A.A. & Bara, J.E. 2020. Class and home problems: Programmatic compilation of chemical data and literature from PubChem using MATLAB. *Chemical Engineering Education* 54(4):230-241. DOI: [10.18260/2-1-370.660-115508](https://doi.org/10.18260/2-1-370.660-115508).
- Swain, M.C. & Cole, J.M. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling* 56(10):1894–1904. DOI: [10.1021/acs.jcim.6b00207](https://doi.org/10.1021/acs.jcim.6b00207).
- Tomaszewski, R. 2019. Citations to chemical databases in scholarly articles: To cite or not to cite? *Journal of Documentation* 75(6):1317–1332. DOI: [10.1108/JD-12-2018-0214](https://doi.org/10.1108/JD-12-2018-0214).
- U.S. Copyright Office. 2017. Compendium of U.S. Copyright Office practices [Internet]. 3rd ed. Available from <https://www.copyright.gov/comp3/>.
- Valko, A.T. & Johnson, A.P. 2009. CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition. *Journal of Chemical Information and Modeling* 49(4):780–787. DOI: [10.1021/ci800449t](https://doi.org/10.1021/ci800449t).
- Warr, W.A. 2011. Representation of chemical structures. *WIREs Computational Molecular Science* 1(4):557–579. DOI: [10.1002/wcms.36](https://doi.org/10.1002/wcms.36).
- Weininger, D. 1988. SMILES, a chemical language and information system 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28(1):31–36. DOI: [10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005).

Wife, D. 2010. Selected organic reactions database [Internet]. Available from https://www.acdlabs.com/download/publ/2010/eum10_wife.pdf.

Wiley Science Solutions. 2020. ChemWindow chemical structure drawing software [Internet]. Available from <https://sciencesolutions.wiley.com/chemwindow-chemical-structure-drawing-software/>.

Zhang, L. 2013. A comparison of the citation patterns of doctoral students in chemistry versus chemical engineering at Mississippi State University, 2002–2011. *Science & Technology Libraries* 32(3):299–313. DOI: [10.1080/0194262X.2013.791169](https://doi.org/10.1080/0194262X.2013.791169).



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Issues in Science and Technology Librarianship No. 97, Winter 2021. DOI: [10.29173/istl2566](https://doi.org/10.29173/istl2566)