Summary

The research project focuses on two aspects: the development of *ChemScanner* - a software library that can be used for the extraction of chemical information from scientific documents that contain ChemDraw sketches - and the integration of the *ChemScanner* library into the electronic lab notebook (ELN).

ChemDraw is one of the most well-known chemical drawing software that is used by chemists and researchers in recent years[1]. ChemDraw binary (CDX) or ChemDraw XML-based (CDXML) files are the most common file formats for molecular structure drawings. Their contents can also be found embedded within DOC, DOCX, or XML documents. *ChemScanner* was developed to retrieve graphics and schemes directly from these file formats by extracting and interpreting information created by chemical professionals. The obtained data are processed together with the additional text and values to form chemical reactions and molecules.

The outputs from the *ChemScanner* library can facilitate the reuse of chemical information embedded into various documents used as standard storage and communication instrument in chemical sciences (e.g., theses, publications, or patents). The software aims to support the chemists in their efforts to re-use chemistry research data by providing them missing tools for an automated assembly of reaction data.

ChemScanner processing results can be visualized via the *ChemScanner User Interface*, as the central part of the integration of *ChemScanner* into Open Source ELN Chemotion[2]. Via the *ChemScanner User Interface*, users would be able to manage the uploaded ChemDraw files and their outputs from the *ChemScanner*.

The *ChemScanner User Interface* supports the export to Excel and CML, the direct import of the extracted data to the Chemotion ELN, or the "copy and paste" feature for selected information. Imported data could be searched using substructure searching or similarity search. Computational properties of molecules can also be calculated and visualized within the integration UI. The integration system is an essential process, not only to improve the ease of use and feasibility of *ChemScanner* but also to keep track of the project development process with further development extensions.

Zusammenfassung

Das Forschungsprojekt konzentriert sich auf zwei Aspekte: die Entwicklung von *ChemScanner* - einer Software zur Extraktion chemischer Informationen aus wissenschaftlichen Dokumenten mit integrierten ChemDraw Dateien - und die Integration von *ChemScanner* in das elektronische Laborjournal (ELN).

ChemDraw ist eine der bekanntesten und am weitesten verbreitetsten Programme zum Zeichnen chemischer Strukturen[1]. ChemDraw Binärdateien (CDX) oder ChemDraw XML-basierte (CDXML) Dateien sind die gängigsten Dateiformate für Molekülzeichnungen, deren Inhalt auch in DOC-, DOCX- oder XML-Dokumente eingebettet werden kann. *ChemScanner* wurde entwickelt, um chemische Informationen direkt aus den Grafiken und Schemata direkt dieser Dateiformate zu interpretieren und zu extrahieren. Die gewonnenen Daten werden zusammen mit zusätzlichen Texten und Werten zu Molekülen und Reaktionen verarbeitet.

Die Ergebnisse der *ChemScanner*-Bibliothek können die Wiederverwendung von chemischen Informationen erleichtern, die in verschiedenen Dokumenten eingebettet sind, die als Standard-Speicher- und Kommunikationsinstrument in den Chemiewissenschaften verwendet werden (z.B. Dissertationen, Publikationen oder Patente). Die Software soll Chemiker bei der Wiederverwendung von Forschungsdaten unterstützen, indem sie ihnen fehlende Werkzeuge für eine automatisierte Zusammenstellung von Reaktionsdaten zur Verfügung stellt.

Die Ergebnisse der von *ChemScanner* prozessierten Dateien können über die *ChemScanner* Benutzeroberfläche als zentraler Bestandteil der Integration von *ChemScanner* in das Open Source Chemotion ELN[2] visualisiert werden. Durch die Verwendung der *ChemScanner* Benutzeroberfläche können Benutzer die hochgeladenen ChemDraw-Dateien und deren Ausgaben von ChemScanner verwalten.

Die *ChemScanner-Benutzeroberfläche* unterstützt den Export in Excel und CML Dateiformate, den direkten Import der extrahierten Daten in das Chemotion ELN und für ausgewählte Informationen die "Copy and Paste" Funktion. Importierte Daten können über die Substruktursuche oder die Ähnlichkeitssuche ausfindig gemacht werden. Die rechnerischen Eigenschaften von Molekülen können auch innerhalb der Integrations Benutzeroberfläche berechnet und visualisiert werden. Das Integrationssystem ist ein wesentlicher Prozess, nicht

nur um die Benutzerfreundlichkeit von *ChemScanner* zu verbessern, sondern auch den Projektentwicklungsprozess mit weiteren Erweiterungen zu unterstützten.

Chapter 1. Introduction

Chemical databases have been playing an essential role for the researcher in organic chemistry. Chemists use chemical information from knowledge databases, experiences, lab notebooks, and literature for synthesis route planning, drug discovery-development, and prediction of new compounds and properties. In the recent years, with the revolution of artificial intelligence (AI) and machine learning (ML), chemical databases are used as the training data sets for the development of machine learning applications in retrosynthesis and reaction prediction progress. Although there are more and more results and improvements, the most crucial obstacle for every machine learning application is the size of the chemical data sets in organic chemistry.

The chemical information retrieval system has been used to automate the extraction progress of chemical information out of scientific literature, comprises patents, theses, and publications. Legacy printed, non-digital resources can be digitalized by scanning then converted into machine-encoded text using Optical Character Recognition (OCR) techniques. The chemical information retrieval systems then curate the digitalized materials for typical information demands in organic chemistry. In practice, chemical entities (e.g., chemical compounds, chemical families, reactions) and their associated information (e.g., preparation steps, safety information) in documents are recognized.

Krallinger et al. [3] described a chemical information retrieval system as a combination of two major components: textual contents and graphical contents, illustrated in figure 1-1.

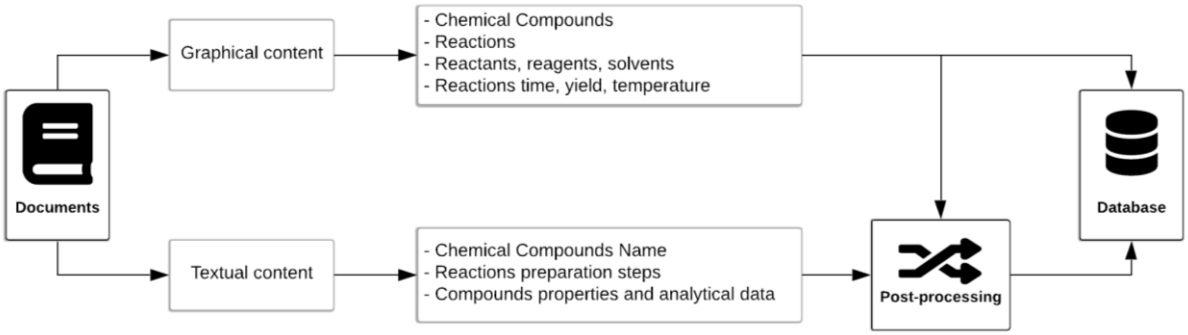

Figure 1-1 The Chemical Information Retrieval System

The *textual contents retrieval component* detects the appearances of chemical entities in the document. Textual chemical entities could be expressed in many methods[4] and do not have a standardized naming convention. Chemical compounds could be described using systematic nomenclature such as IUPAC nomenclature (e.g., dihydrogen monoxide), trivial names (e.g., water), acronyms or abbreviations (e.g., THF, DMF), sum formulas (e.g., C_6H_6), name of groups (e.g., ketones, aldehydes), registered trademarks or brand names (e.g., aspirin, paracetamol). The *textual contents retrieval component* responsibility is to detect and retrieve these entities by employing text mining and natural language processing (NLP) approaches.

Chemical entities that are detected from the *textual contents retrieval component* only have practical meaning if one knows which molecular structures they are referring to. The *post-processing component* is responsible for the linking of chemical entities with the structural information. This can be achieved by employing various approaches. *Martin et al.*[5] employed the search results from chemical databases such as PubChem[6], ChemSpider[7], or SciFinder[8]. *Grego et al.*[9] used a dictionary-based approach for the ChEBI database. The ChemSpot system[10] integrated the name-to-structure software OPSIN[11] to convert IUPAC names to chemical structure. In addition to the name-to-structure conversion, another primary approach for the linking of chemical entities to molecular structures is to use the structural information within documents from the *graphical contents retrieval component*.

Typically, the *graphical contents retrieval component* is an application of OCR[12–14], or specifically in chemistry, is Optical Chemical Structure Recognition (OCSR) to convert image to chemical structure. Although many OCSR applications are developed, reconstructing chemical molecules is an error-prone process[15]. Furthermore, even if the reconstruction is successful, only molecules are rebuilt, the chemical reactions information from images are still missing. Since most chemical professionals are using chemical drawing software (e.g., ChemDraw, ChemSketch, ISIS/Draw …) for their research, the outputs from this software are computer-readable formats so that they would be processed more precisely and reliable.

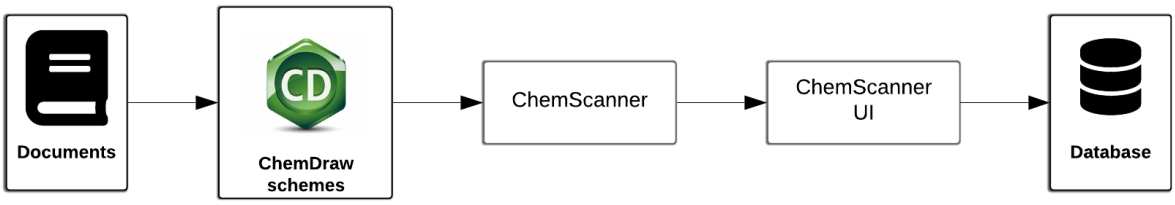

Figure 1-2 Proposed Architecture

To make use of the output from chemical sketcher software, the research project proposes and develops *ChemScanner*, to extract and interpret chemical structural information out from ChemDraw, one of the most prevalent chemical drawing software. The research project also includes the integration of *ChemScanner* into the Open-source Chemotion Electronic Lab Notebook, as shown in figure 1-2. The integration process improves the usability of the software by employing the web application user interface, the *ChemScanner User Interface*. Also, a better management mechanism is provided with the storage management UI together with the searching feature.

1.1 Motivation

The research project is motivated to develop *ChemScanner*, a novel software library that retrieves chemical structural information from output files of the ChemDraw sketcher software.

ChemDraw is known as one of the most popular chemical drawing software. Many chemical researchers and professionals are using ChemDraw for their daily research. However, when their works are shared with others, publicly by publications or internally within their organizations, drawn schemes are treated as digitized images (e.g., BMP, TIFF, PNG, JPG/JPEG or GIF). Its consequences that useful structure-data generated by drawing software is lost, and would not be used properly.

Fortunately, figures drawn by ChemDraw are not entirely disappeared, and there are plenty of resources for mining. Copy and paste figures into Word documents (DOC or DOCX) embed them into the documents while maintaining the original contents of ChemDraw files. Word files (e.g., theses, manuscript) are saved afterward and shared internally or ready to submit for publications. The United States Patent and Trademark Office (USPTO) accepts and stores over 24 million patents ChemDraw files[16].

Current approaches with images to structures techniques (e.g., OSRA[14], CLiDE[17], ChemoCR[18], Imago OCR[19]) are struggling with handle complicated OCR challenges appropriately. For example, OCR software is usually confused and makes many mistakes while dealing with the recognition of wavy bonds, or crossing bonds. Comparing with these approaches, sketches produced by ChemDraw or other drawing software are more computer-readable. By reading the chemical structural contents from the sketches directly, one can overcome many OCR challenges.

For all the above reasons, we believe that our research is a combination of many advantages for the chemical information retrieval system. Our system has the potential to allow more chemical databases to be created. The research project benefits chemical researchers and organizations, as well as the publishers, to mine their data using *ChemScanner* and its integration with the ELN.

1.2 Problem Statement

Although mining sketches seems easy because we can read what users want to draw, without losing any information during translation as OCR approaches, interpreting chemical schemes as what they mean is challenging. Some symbols and graphics are widely used, but their real meaning is entirely different from their original purposes. For example, "*Ar*" purpose is to indicate the chemical element "*Argon*". Nevertheless, in most cases, it is interpreted as "*aryl*". Alternatively, "*Ac*" often means "acetyl" instead of the "*Actinium*" element. More similar ambiguous symbols like "*B*", "*V*", "*W*", "*Y*" intend to represent an ordinary generic atom/group label instead of "*Boron*", "*Vanadium*", "*Tungsten*", "*Yttrium*" accordingly.

Besides, non-chemical elements, including text and graphics, also need to be used in combination with structural information to interpret schemes into proper molecules and reactions. ChemScanner is designed to cover all of these scenarios to correctly derives what they are meant to be. The implementation details are described in chapter 3.

Practically, researchers want to describe the information within the scheme as much as possible. Since there is flexibility for people to use their creativity to create their perfect pattern that fits with their documents, it is almost impossible to reach a perfect conversion without human interaction. In order to solve this problem, the research project also develops the

ChemScanner UI, which is part of the integration of ChemScanner into the Chemotion ELN, to support human monitoring and to improve the conversion time.

1.3 Related Studies

Many commercial solutions introduce sketches mining approaches in the past. Wiley[20] employed the templating approach by defining sets of problems in order to extract molecules with their R-groups label and information. The templates are predefined and are used to extract only molecules. ICSCHEMEPROCESSOR[21] from InfoChem addressed molecular drawing challenges and came up with a hybrid with a templating approach and an algorithmic approach. They are aiming more on molecule extraction than reaction extraction by skipping many drawing patterns of reactions.

Recently, *May et al.*[22] from NextMove Software report a mining approach on the USPTO ChemDraw files, by combining extracted R-group labels and repeated group with R-group table to assemble combinations of molecules from core structures and R-groups. However, they are struggling with undefined chemical entities that cannot be converted.

In summary, these above approaches are only targeting on molecular extraction, without associated reactions information while reactions are more vital to organic chemistry, especially on retrosynthesis and reaction predictions, since reactions are defining they synthetic pathway to desired molecules. Besides, they are all commercial solutions that are closed-source and hard to access by researchers. Also, they do not provide an interface for human interaction with the extraction progress.

1.4 Overview of the research project

In the rest of this thesis, we present our approach to design and implement *ChemScanner* and the integration of *ChemScanner* into Chemotion ELN. The progress is organized as follows.

Chapter 2 will deeply explain the background information about *ChemScanner* development. Basic cheminformatics formats are covered, explaining how molecules and reactions are stored in computer-encoded formats. This chapter also describes how ChemDraw organized the internal sketcher data into CDX and CDXML formats. A review of the current state-of-the-art chemical information retrieval is included in the end.

Chapter 3 presents the overall design and implementation of *ChemScanner*. Challenging issues and particular problems scenario handling are explained in this chapter, together with the use of the library.

Chapter 4 introduces the interface of *ChemScanner*, created as a web application. This chapter cover in detail each feature of the *ChemScanner User Interface* as an intermediate component between *ChemScanner* and the electronic lab notebook.

Chapter 5 describes the integration into the *Chemotion Electronic Lab Notebook* process. This chapter shows every detail of the molecule searching, deployment of the web service used for computational properties, and propose the Green Chemistry attributes.

This thesis is enclosed with chapter 6, which summarizes the development as well as future enhancements.

Chapter 2. Background

This chapter introduces the fundamental concepts that are used for *ChemScanner* and *ChemScanner* development.

First of all, it is explained how molecular information is stored and retrieved with two popular file formats in chemical information (cheminformatics): SMILES and MDL Molfile. Secondly, the details of the ChemDraw file-formats family are described. The formats CDX, CDXML are covered, and it is described how they are embedded inside Word files. Finally, a quick review of current state-of-the-art approaches in chemical information retrieval is given.

2.1 Molecular representing

2.1.1. SMILES

In cheminformatics, a line notation is a single-line string (a sequence of characters), nowadays the most well-known line notation formats are the IUPAC International Chemical Identifier (InChI)[23] and the Simplified Molecular-Input Line-Entry System (SMILES)[24]. InChI is the latest line notation and is more modern than SMILES. However, SMILES is still the best-known because it is more human-readable and is supported by most molecule editors.

The original SMILES specification was introduced by David Weininger[25], then being adapted and modified by many following information systems, especially Daylight Chemical Information System[26]. The latest version of SMILES is an open standard OpenSMILES[27], which was introduced by the Blue Obelisk[28] community.

In general, the SMILES notation is a sequence of characters that end with a whitespace terminator character (space, tab, newline, carriage-return) or the end of the string, while hydrogen atoms could be included or omitted. The SMILES string is obtained by picking one first atom, then printing atomic symbols of atoms while traversing the chemical graph of the molecule structure in any order with five basic rules following.

- **Atoms**: All periodic table elements are supported, asterisk symbol (“*”) is accepted as a wildcard or unknown atom.
    - An atom is represented with its respective atomic symbol.

    - Upper-case letters indicate non-aromatic atoms, and lower-case letters refer to aromatic atoms. With atomic symbols that have more than one letter, the second letter must be lower-case.
    - Atoms do not belong to the organic subset (B, C, N, O, P, S, F, Cl, Br, I), and atoms with abnormal valences must be enclosed in brackets. Figure 2-1 describes typical valences of the organic subset atoms.

- **Bonds**: Single, double, triple, aromatic bonds and disconnected structures are expressed by "-", "=", "#", ":", and "." respectively. Single bonds can be omitted.
- **Branches**: Branched atoms in the chemical graph must be enclosed in parentheses. Branches could be nested or stacked to any depth.
- **Ring**: a number is placed right after the opening and closing ring atoms to identify ring structures with SMILES.
- **Disconnections**: Structures that are disconnected between others are separated in SMILES string by a period symbol (".").

B	C	N	O	P	S	Halogens
3	4	3,5	2	3,5	2,4,6	1

Figure 2-1 Organic subset atoms valences

The same molecule can be represented by different SMILES strings using the basic rules. Figure 2-2 describes four different ways to represent Ethanol. The first form represents Ethanol with implicit hydrogens, and the other three forms represent without implicit hydrogens with a different order of the starting atom. The second and third forms are used more in general since they look simpler and shorter. Therefore, a normalization process is proposed by OpenSMILES, this normalization process is not mandatory, but it is used in many guidelines of chemical information systems to generate SMILES strings.