INFORMATION FUSION IN TAXONOMIC DESCRIPTIONS

BY

QIN WEI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor P. Bryan Heidorn, Chair and Director of Research
Professor Linda Smith
Associate Professor Catherine Blake
Research Scientist James Macklin, Agriculture and Agri-Food Canada

Abstract

Providing a single access point to an information system from multiple sources is helpful in many fields. As a case study, this research investigates the potential of applying information fusion techniques in biodiversity area since researchers in this domain desperately need information from different sources to support decision making on tasks like biological identification. Furthermore, there are massive collections in this area and the descriptive materials on the same species (object) are scattered in different places. It is not easy to manually collect information to form a broader and integrated one.

As one of the most important descriptive materials in this field, floras are selected as the target of this research. This research tests a hypothesis concerning the organization of text and the constancy of fact-based information in text. It is observed that individual descriptions may not contain sufficient information to differentiate the target species from others, and different information sources might contain not only overlap information but also complementary information that is helpful. We also observe non-trivial complementary information could also be from different-level descriptions [family, genus, or species level] from the same source. By using the sample dataset from Flora of North America (FNA) and Flora of China (FOC), we found that about 50% information could only be found in single source and another 25% complementary information could be identified by fusion. And the most importantly, confliction information could only be detected by direct comparison.

The question is how could we fuse the records in an automatic or semi-automatic manner, so that each resulting record provides a broader while non-redundant description of each species? The proposed system demonstrates the feasibility with currently available techniques. The prototype system contains 4 modules: Text segmentation and Taxonomic Name Identification, Organ-level and Sub-organ level Information Extraction, Relationship Identification, and Information fusion. By using the sample descriptions from Flora of North America and Flora of China, we demonstrate that the method gain promising fusion result based on Cross-Description Relationships. With the evaluation results, we identified the key factors contribute to the performance of fusion. Some methods that might lead to further improvement on fusion performances are discussed.

This study also demonstrates that to a certain extent, this fusion approach is generalizable. The generalizability of this fusion approach is a challenging problem due to the typical domain- and task- oriented nature of the fusion methods. We identified the challenges while applying the approach to different data set.

To Leo, Zhi, Mom and Dad

Acknowledgements

This dissertation would not have been possible without the help from my dissertation committee: Dr. P. Bryan Heidorn, Dr. Linda Smith, Dr. Catherine Blake, and Dr. James Macklin. I can not say thank you enough to my advisor, dissertation committee chair and director of research, P. Bryan Heidorn, who offers me the support and guidance throughout my doctoral study here. Thanks to Dr. Linda Smith for her consistent support during my study in the program. Also thanks to Dr. Catherine Blake and Dr. James Macklin who answered my numerous questions during the course of the project. I also want to thanks Dr. Stephen Downie and Dr. David Boufford who answered many biology-related questions in this project.

Many thanks go to Dr. Leigh Estabrook who gives supports and encouragement during my graduate study. I also want to thanks Dr. Allen Renear, Dr. Stephen Downie, Dr. David Dubin, Dr. John MacMullen, Dr. Jerome McDonough, Dr. Vetle Torvik, who served in my various committes and discussed with me many interesting research or non-research questions. I am thankful to my peer doctoral students who shared with me the joy and frunstrations. I also want to thank all the participants in the *Second Fine-Grained Semantic Markup of Descriptive Data for Knowledge Applications in Biodiversity Domains* for their help in evaluating the results.

Finally, thanks to my husband, son, mom and dad who endured this long process with me with their great patience, support and love. This dissertation is delicated to them.

Table of Contents

Chapter 1: Introduction

1.1 Background

A routine but important task for biologists is biological identification. Biological identification refers to the assignment of individual specimens to previously classified and named groups (Jardine, 1969). For example, if one brings a specimen in from the field and attempts to find a scientific name for it, that constitutes identifying the specimen. In biological identification, the morphology (the external form) of an organism has been, and still is, the type of data that is the most-commonly used in identification.

Current methods of identification rely heavily upon a system known as keys, which was initially proposed by Linnaeus back in the eighteenth century. In simplest form, keys are dichotomous, quite similar to a decision tree. The process of constructing a decision tree is similar to a decision-making procedure. The underlying idea of a decision tree is to break up a complex decision into several simpler decisions. Each decision made leads to a different choice. In our case, each decision made about morphological features leads to categorization at different levels of a hierarchy: family, genus, or species. An example of a modern key is presented in Table 1.1.

The following is an explanation of how a key works. In the first state (marked as 1 in the first column of Table 1.1), if one's specimen has 3-foliolate leaves, follow to 2. If the leaflets are "abaxially densely white woolly, chalazal knot rounded or broadly elliptic",

the specimen should be classified as *Ampelocissus artemisiifolia*. Otherwise, it is *Ampelocissus butoensis*.

Table 1.1: Genus *Ampelocissus* keys[1]

1	Leaves 3-foliolate	(2)
+	Leaves simple	(3)
2 (1)	Leaflets abaxially densely white woolly, chalazal knot rounded or broadly elliptic.	4 A. artemisiifolia
+	Leaflets abaxially pilose, chalazal knot narrowly elliptic.	5 A. butoensis
3 (1)	Branchlets and petioles tomentose and with purple bristles.	3 A. hoabinhensis
+	Branchlets and petioles glabrous or tomentose, but not hispid	(4)
4 (3)	Leaves cordate-oval, glabrous.	1 A. sikkimensis
+	Leaves broadly ovate, adaxially densely pubescent, abaxially pilose, with woolly hairs on veins	2 A. xizangensis

Using keys is straightforward, and keys are used widely and heavily for identification. However, there are weaknesses. First of all, a researcher must know the exact answer for all branches of the decision tree in order to reach the leaf nodes (the scientific name). However, in real cases, a user may have insufficient knowledge to be able to navigate the

[1] *Ampelocissus* key retrieved from http://efloras.org/florataxon.aspx?flora_id=2&taxon_id=101412 on April 4, 2010

[2] Biodiversity Heritage Library (BHL) Accessible at

key and thus make a bad decision leading to a wrong identification. The error may be based on the specimen being incomplete or the users inability to make good decisions.

Given the limits of the keys method, taxonomists have attempted to use different techniques to make the key generation process less time-consuming and to improve their accuracy by making use of computer-aided methods. The best example is is DELTA (DEscription Language for TAxonomy) (Dallwitz, 1993). DELTA is a method for encoding taxonomic descriptions into digital format, and the DELTA system was developed based on the DELTA format that contains a set of programs. One of the most used programs within the DELTA system is known as INTKEY, which is an interactive identification component. Unlike paper-based keys, INTKEY allows users to express characteristics in an arbitrary order, and can be used to retrieve identification characteristics more efficiently. In other words, the DELTA system is quite similar to a digital-based keys system in which access to the characters is searchable. The DELTA system can be more effective than paper-based keys, but several important weaknesses have limited DELTA's success and usefulness. First of all, the system requires manual entry of all attributes of every character into the system. This in turn requires expert knowledge on how to make decisions regarding information to be entered. Also, coding from textual descriptions to the DELTA format is very time-consuming. In addition, just as is the case with traditional paper-based keys, it is impossible to express any degree of uncertainty that might exist within the identification process.

On the other hand, there are major collections of morphological descriptions from

journals, monographs, books, etc., which are written in sub-language that will be discussed in detail in Chapter 2 and deposited in digital libraries such as the Biodiversity Heritage Library (BHL)[2] that can help researchers. These variably detailed descriptions were generated during the course of centuries of research, and contain valuable comparative information. Sub-language texts need to be transformed into semi-structured or structured formats in order to become potentially useful in information retrieval systems. This dissertation proposes a method to improve on systems like DELTA to access taxonomic descriptions. The goal is to develop a knowledge base comprised of the collections of morphological descriptions in sub-language.

Table 1.2: Descriptions of *Alternanthera sessilis*[3]

Level	Flora of North America	Flora of China
Genus	Leaves opposite, sessile or petiolate; blade lanceolate to ovate, ovate-rhombic, or obovate-rhombic, margins entire.	Leaves opposite, margin entire.
Species	Leaves sessile; blade elliptic to oblong or oblanceolate, 1.2-5 × 0.5-2.2 cm, apex obtuse to acute, glabrous.	Petiole 1-4 mm, glabrous or pilose; leaf blade linear-lanceolate, oblong-obovate, or ovate-oblong, 1-8 × 0.2-2 cm, glabrous or pilose, base attenuate, margin entire or slightly serrate, apex acute or obtuse.

For example, Table 1.2 shows four typical descriptions that can be found in the literature. They describe species *Alternanthera sessilis* on different levels and from different

[2] Biodiversity Heritage Library (BHL) Accessible at http://www.biodiversitylibrary.org/

[3] Descriptions from efloras.org

sources. Note that the generic descriptions include characteristics for all the species found in the area covered. So, you need to be a little more careful how you describe this relationship.

This example shows that individual descriptions contain some unique information that could not be found in other descriptions. Different information sources might contain both overlapping information as well as complementary information that can be relevant and useful. Aside from complementary information taken from the same species description that appears in different sources, non-trivial complementary information can also be extracted from different taxonomic-level descriptions (family, genus, or species level) from the same source.

The example shows that combining multiple descriptions of the same species from different sources and different levels can provide the researchers more complete information than any single description. The question is: how could we fuse matching records in an automatic manner so that each resulting record provides a broader and non-redundant description of each species?

This process is known as information fusion (IF). Information fusion is an area of research that studies approaches to combining data or information that can originate from different sources. Llinas, et al. (2004) describes information fusion as "an information process that associates, correlates, and combines data and information from single or multiple sensors or sources to achieve refined estimates of parameters, characteristics,

events and behaviors".

There are two major reasons why information fusion has recently become important. On one hand, we are experiencing information overload because digitally-formatted information increases in volume every minute. Meanwhile, information is distributed by many different sources and in different formats, e.g., text, audio, and video news from different publishers. Complementary and/or conflicting information is scattered around in many places and only when such information is combined can it provide us with a more complete and unbiased outlook.

This dissertation project is an exploratory study of the application of information fusion techniques to a specific biological information source, taxonomic descriptions. We hypothesize that:

There is non-trivial complementary information existing in different floras on the same species. It is also true that genus-level (higher-level) descriptions contain non-trivial complementary information on the species-level (lower-level). Automatic/semi-automatic information fusion is useful both because it provides a single access point to the user and provides better data quality. It also provides us with opportunities to detect conflicting information.

It is argued that data quality can be measured by several dimensions: completeness, validity, accuracy, consistency, integrity, timeliness (Ravn & Høedholt, 2009). In this

dissertation project, we tried to improve the data quality by focusing on its completeness and consistency.

1.2 Objectives

This dissertation work seeks to study the problem of access to information from biodiversity sources by providing a single access point to multiple sources of information. The same object or the same attributes of an object may be described in different sources. We would thus expect that we could obtain better data quality and a broader scope of data by using multiple sources. The problem is that current information systems are primarily based on a single source, therefore, researchers must consult multiple sources when they are doing their comparative studies and want to find the information they require.

This is a proof of concept project that will focus on taxonomic descriptions in the biodiversity area because users desperately need information from multiple sources in order to make sound decisions. In another sense, there are large numbers of collections in this academic area. Due to that there are so many sources it is difficult (or time consuming) for a researcher to manually access and compile the information from different sources.

The primary objectives of this work are:

1. Obtain in-depth insight into the informational properties of morphological

descriptions and the constancy of fact-based information in those texts.

2. Develop a cross-description relationship taxonomy based on leaf descriptions using multiple sources.
3. Study the implications of cross-description relationships with respect to text mining in general and information fusion in particular.
4. Develop an informational fusion system that provides a single point of access to multiple floras.
5. Identify the key factors that contribute to the performance of information fusion.
6. Identify the key challenges when applying the method to different texts.

This study provides a quantitative approach for estimating the quality and quantity of the information contained in various taxonomic treatments and investigates the underlying cross-description relationships between them. Developing a functional application demonstrates the feasibility of such a system using currently available techniques. This approach is evaluated by expert and non-expert users in terms of its accuracy.

1.3 Research Questions

Before information fusion can be conducted, we must investigate the information patterns inside the text we will process. Lydon, et al. (2003) conducted similar research on this topic. Their findings are useful and indicative (additional details appear in the Chapter 2 Literature Review). Their findings are based exclusively on a case study of five descriptions of species, so it is far from comprehensive. For this specific research area,

more detailed and systematic investigation regarding the information patterns in taxonomic treatments is needed. Implementation will not be possible until such time as the following questions answered:

Q1: Does non-trivial complementary information exist in different sources and at different taxonomic levels? If yes, what are their types and how frequently do they occur?
Q1.1 What are the major semantic relationships existing in taxonomic descriptions based on cross-sentence relationships?
Q1.2 What types of complementary information can be derived from different sources? And from different levels?
Q1.3 Does conflicting information exist?

The results from these questions allow us to move on to the implementation stage. The answers to Question 1.2 will guide us on how to design the system, and the answers to Question 1.3 will help us evaluate the system. The answers to Question 1 will be used to help address the following questions:

Q2 Is a semi-automatic or automatic information fusion feasible and how can it be evaluated?
Q2.1 What are the challenges involved in implementing such a system? What techniques can be used in the proposed system?
Q2.2 How should such a fusion system be evaluated?
Q2.3 What are the key factors that contribute to the performance of the system?

In order to evaluate the proposed information fusion method, an actual fusion system will be built and an evaluation experiment will be conducted to test the performance. The answers to Question 2.2 and Question 2.3 will inform us on the current performance of the system. The answers to question 2 will help us to answer the following question:

Q3 What are the challenges when applying the method to different texts?

By answering this question, we want to demonstrate the generalizability of this method and provide suggestions for future research in information fusion.

The three research questions are closely related and each is built upon the previous one. Together, they answer the overarching question of why information fusion is necessary and how it could be successfully accomplished.

1.4 Summary

This chapter introduces the notion of information fusion and explains briefly why information fusion is important in text mining. The three research questions that were proposed in this chapter are closely related and their answer will shed light on our decisions about designing the fusion system.

Chapter 2: Literature Review

2.1 Cross-document Sentence Theory (CST)

CST (Cross-document Sentence Theory) is based on RST (Rhetorical Structure Theory). RST was proposed in Mann & Thompson (1988) and presents the rhetorical relationships between sentences. The theory is based on the assumption of text coherence that is defined by Taboada and Mann (2006) as "coherence by postulating a hierarchical, connected structure of texts, in which every part of a text has a role, a function to play, with respect to other parts in the text."

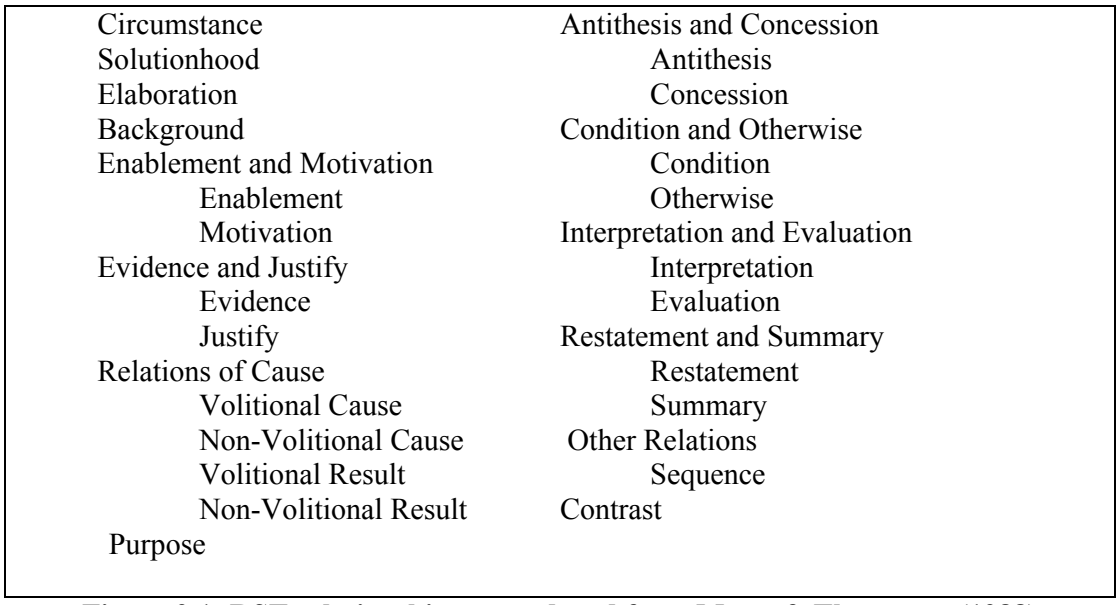


Figure 2.1: RST relationships reproduced from Mann & Thompson (1988)

In RST, a sentence (which can sometimes be a noun phrase or a verb phrase) is called a text span. The text span could either be a "nucleus" or "satellite." The nucleus is the most

important part of a text while a satellite contributes to the nucleus. Multiple spans can be combined into a single larger span. Four fields must be defined in order for each relationship to exist: constraints on the nucleus, constraints on the satellite, constraints on the combinations of nucleus and satellite and the results of their combinations.
The relationships proposed include 12 groups and groups may have subcategories. Figure 2.1 presents the relationships (Mann & Thompson, 1988).

The importance of RST is that it provides a general way to describe relations among clauses within a text regardless of whether they are grammatically or lexically related or not. The relationship can exist in multiple sentences and it consists of one or more nuclei and zero or more satellites. They argue that RST has wide areas of application including computational linguistics, cross-linguistic studies, dialogue and multimedia, discourse analysis, argumentation and writing, etc. An example of document structure analysis using RST appears in Figure 2.2.

CST is largely based on RST and similar assumptions are applied. The difference between CST and RST is that RST is used to analyze the sentence relationships within a document whereas CST focuses on cross-document sentence relationships.

Radev (2000) proposed using CST to analyze document structure and considered it to be the first step toward advanced text mining, e.g. automatic information fusion. CST is a taxonomy of cross-document rhetorical relationships for generic texts.

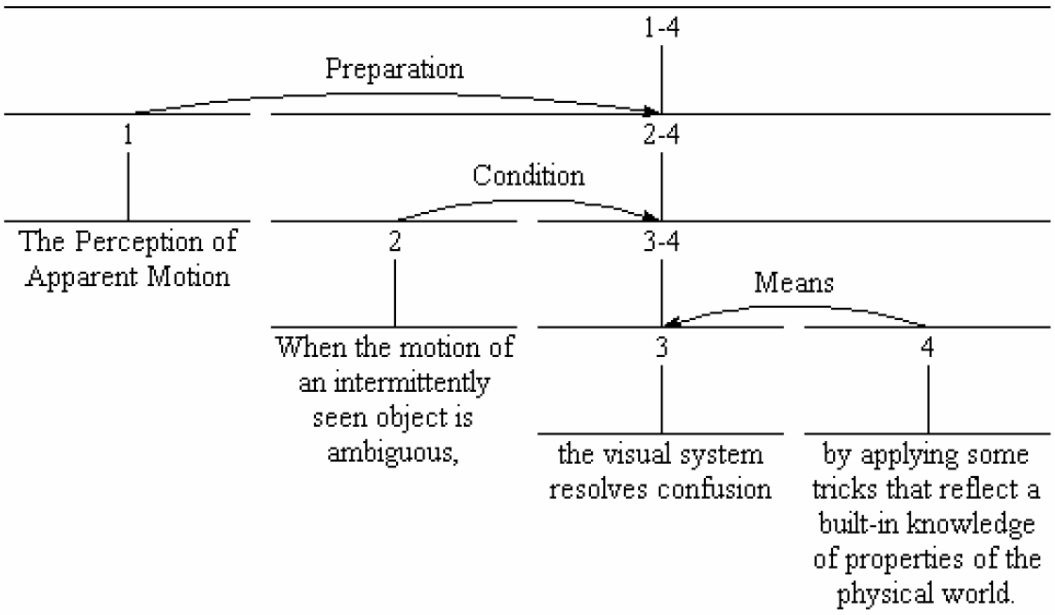


Figure 2.2: Diagram of an RST analysis reproduced from Taboada & Mann (2006)

It has been argued that it is primarily useful in literature where different sources describe the same event or object at the same time, or when a single source publishes a series of publications on the same event or object over a period of time. CST proposes 24 relationships and the relationships that can be on 4 different levels: word (W), phrase (P), sentence or paragraph (S) or entire document (D) level (Radev, 2000). See Table 2.1 for more detailed information about these relationships. It has been argued that CST "is essential for the analysis of contradiction, redundancy and complementarity in related documents" (Radev, 2000).

Several studies have been conducted that have used RST or CST to analyze document structure, e.g. Fox, 1987; Radev, 2000. There are two major benefits to be obtained from applying CST to taxonomic documents: analyzing the relationships in the documents is

both helpful for revealing the underlying information organization patterns in those documents and as the most important input information for information fusion.

Table 2.1: CST relationships and levels

Relationships	Level	Description
Identity	Any	The same text appears in more than one location
Equivalence	S, D	Two text spans have the same information content
Translation	P, S	Same information content in different languages
Subsumption	S, D	One sentence contains more information than another
Contradiction	S, D	Conflicting information
Historical	S	Information that puts current information in context
Cross-reference	P	The same entity is mentioned
Citation	S, D	One sentence cites another document
Modality	S	Qualified version of a sentence
Attribution	S	One sentence repeats the information of another while adding an attribution
Summary	S, D	One textual unit summarizes another
Follow-up	S	Additional information which reflects facts that have happened since the last account
Elaboration	S	Additional information that was not included in the last account
Indirect Speech	S	Shift from direct to indirect speech or vice-versa
Refinement	S	Additional information that is more specific than the one previously included
Agreement	S	One source expresses agreement with another
Judgment	S	A qualified account of a fact
Fulfillment	S	A prediction turned true
Description	S	Insertion of a description
Reader profile	S	Style and background-specific change
Contrast	S	Contrasting two account of facts
Parallel	S	Comparing two accounts of facts
Generalization	S	Generalization
Change of Perspective	S, D	The same source presents a fact in different light

In order to support automatic information fusion, automatic relationship identification should be conducted first since the fusion is done based on the relationships identified between the descriptions. The identification of relationships can be viewed as a

classification problem: given the two sentences and the relationships available, which relationship can be applied to the two sentences? Machine learning has been widely used in this context with supervised methods such as Support Vector Machine (SVM), Conditional Random Field (CRF), Hidden Markov Model (HMM) and Naïve Bayes (NB) having received the most attention and achieved the most promising results, e.g. Wang, et al., 2005; Bellegarda, 2004; Peng, et al., 2004; Langseth and Nielsen, 2006.

Zhang, Otterbacher and Radev (2003) proposed using Boosting that is one of the machine learning algorithms to automatically classify CST relationships. The identification problem could be one of two complementary scenarios: binary classification (whether two sentences are correlated) and multi-class classification (identify the specific relationship(s) that could be applied between the two sentences). The features used in this research are: lexical feature (number of tokens in sentence 1, number of tokens in sentence 2, number of tokens in common), syntactic feature (number of tokens having Part-of-Speech POS x in sentence 1, number of tokens having POS x in sentence 2, number of common tokens having POS x, and semantic features). Semantic features here are the distances between the most important concepts discussed in each sentence with WordNet being used to compute them. For binary classification, the typical evaluation measures used were precision, recall and F-measure. The evaluation measures used for multi-class classification are more interesting. They include one-accuracy (whether the top ranked class is among the correct ones), coverage (how far do we have to go down the ranked list to find all of the correct ones), and average precision. The results reported in Zhang, Otterbacher and Radev (2003) are promising. Using 41 articles about different

topics led to the finding that the highest F-measure they had for binary classification is about 73%. The multi-class classification performances are worse and different for different classes. The performance of each class depends on the training examples they had in that class.

Aside from supervised learning, unsupervised learning has also been proposed to be useful in automatic CST relationship classification. Marcu and Echihabi (2002) presented an unsupervised approach to automatically identifying four relationships of RST (contrast, explanation-evidence, condition and elaboration) found in two arbitrary sentences. They constructed the training examples using several hand built patterns: contrast (but, although), explanation-evidence (because, thus), condition (if, then), and elaboration (for example, which). Their method begins with the hypothesis that the relationship between the sentences is determined by the word pairs in the sentences. They compare the conditional probability of the relationships for different word pairs. The identified word pairs are then used to train a Naïve Bayes classifier to classify new examples. The classifiers performed surprisingly well on their test data. The primary reason for their success was that the four relationships have more lexical clues than the others in RST.

Real applications demonstrate that utilizing the automatically identified CST or RST relationships in the text mining system shows promise.

Zhang, Blair-Goldensohn and Radev (2002) proposed a CST-enhanced document

summarization system. The connectivity of sentences is ranked by automatically identified CST relationships. Their results show that the effects of adding CST are positive. They also found that different CST relationships produce different effects on the final output.

Wan (2008) conducted research on whether the cross-document relationship or the within-document relationship is more important in multi-document summarization. The results showed that for both generic and topic-focused types of summarization, using only cross-document relationships could outperform at least as well as approaches that were based on within-document relationships.

2.2 Text Mining in Taxonomic Literature

Taxonomic descriptions are different from other types of documents in the sense of how they are constructed. Developing tools for parsing multiple taxonomic descriptions must be based on our understanding of the characteristics of the descriptions.

Taxonomic descriptions are usually written in a sub-language. A sub-language can differ from natural language in several ways: vocabulary, grammar, and most important of all, how it carries knowledge. In a sub-language, both the words and the grammar can carry semantic meanings. The primary characteristics of taxonomic descriptions include:

1. Taxonomic descriptions are more compact than natural language (Lehrberger,

1982). Strictly speaking, the sentences in the descriptions are not real "sentences". The verbs are typically omitted and the sentences are shorter. The descriptions are mostly in a form such as: Noun-adjective.

2. Taxonomic descriptions have three levels of restrictions on word class and the combinations of words from different classes.
    - **Lexical**: taxonomic descriptions have a relatively smaller vocabulary in each character. For example, the word appearing in descriptions of leaf arrangement or leaf shape are less broad than the words that can appear in news reports. The vocabularies are thus less broad and less confusing.
    - **Syntactic**: The combinations of the words in taxonomic descriptions are not random but are instead pre-defined. For example, we can only say: "leaves alternate" but not "alternate leaves".
    - **Semantic**: The meaning of a word in a taxonomic description is related to where the word is found or the context of the words associated with it. For example, in the context of a leaf arrangement, alternate only means leaves borne singly at each node on a stem butin the context of a stamen arrangement in a flower, alternate also means stamens borne between the petals.
3. Information Organization patterns. Lydon, et al. (2003) reported an analysis of the information in different paper-based botanical descriptions. Five species descriptions of the genus *Ranuculus L*. from six different English language floras were compared. Several important findings were reported:
    - The descriptions varied in the characters that were described. They found

that more than half of the information (55%) came from a single source. The ratios for absolute agreement, some degree of agreement and disagreement were 9%, 36% and 1%, respectively. Similar patterns were observed in family-level descriptions.

- When the same properties were mentioned, in most cases, they provided overlapping information. However, different vocabularies were used. This non-uniformity is common in the literature.
- Conflicting data does exist. Although the percentage was not high (1% reported), conflicting data has a special importance. Contradiction might be due to: differences in methodology (museum specimen vs. field specimen), mistakes (misidentification, misuse of words, typographic errors), or genuine differences between specimens or populations. Interesting findings can be obtained when genuine differences are identified. They might imply an unrecognized variation or a wider range of character states. If the disagreements are identified because the specimens were studied at different times, they might reflect historical changes.

The characteristics mentioned above made text mining in this field different than generic text mining. Lydon et al. (2003) argues, "techniques by merging the results from several descriptions which is more complete than that in any one source, are likely to provide a way forward in dealing with the vast amounts of overlapping botanical legacy data." Our research focuses on merging the information together to improve accessibility. Related

research includes information extraction and fusion.

Some studies have been conducted on using natural language processing techniques to automatically extract knowledge from literatures and transform sub-language into a structured xml format. The techniques include rule-based parsing and statistical-based parsing which are discussed below.

Rule based. Rule-based approaches "perform deep analysis of linguistic phenomena and are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms" (Liddy, 2001). The simplest type of rule NLP is the use of Regular Expressions (RE) that define a regular grammar in a text string. The NLP processes are then based on a series of REs (Hahn and Wermter, 2006). Rule-based methods were the first proposed for automatic extraction of information from taxonomic descriptions.

Taylor (1995) proposed an information extraction system for building a knowledge base by parsing taxonomic descriptions. Their parser was built using manually constructed grammars that contain 120,000 character/state pairs. Some induction rules were used to handle unknown pairs and words. Neither qualitative nor quantitative evaluation data was provided. Similar to the system that Taylor (1995) proposed, X-tract (Abascal and Sánchez, 1999) was proposed for the extraction of structural information from taxonomic descriptions. Their system contains a pre-parser, a parser, a filter, a semantic interpreter and a structure resolution. The parsing is based on a set of syntactic rules that were

collected based on the corpora and was heavily reliant on presentational tags, e.g. html tags.

Positive results have been reported regarding the use of rule-based algorithms in their application. Tang (2007) reported better information retrieval performance when using the information extraction results than a keyword based retrieval system.

However, some limitations exist. A large amount of data and labor (highly skilled linguists and/or domain experts) are required in order to create, enhance, and maintain the rules (Richardson, 1994). Rules developed by domain experts have usually been deduced from well-formulated examples. It is difficult to encode every exception to the rules because exceptions are difficult to predict prior to application.

Such manually constructed rules have very strict restrictions for application, and lack the flexibility needed for dynamic adaptation to data/corpus. Once the rules have been developed, they can not be easily adapted for use with another domain. The rules need to be analyzed carefully in order to be adapted for application. Given such limitations, researchers began to focus on using statistical methods in this field.

Statistical NLP. Statistical methods "use observable data as the primary source of evidence" and "employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant

linguistic or world knowledge" (Liddy, 2001). Statistical NLP does not take advantage of language structure because it views a sequence of text the same as a sequence of non-meaningful arbitrary symbols (Rosenfeld, 2000). The statistical method can be either supervised or unsupervised.

Cui and Heidorn (2007) developed a MARTT (MARkuper of Taxonomic Treatments) system that automatically converts taxonomic descriptions into XML format. They adopted the supervised machine learning approach to the extraction of domain knowledge from the training data and then applied the induced knowledge to new text. They found that the knowledge that had been deduced from the more structured corpora is useful when such knowledge is applied to less structured corpora.

In comparison with rule-based NLP, supervised learning is useful in the sense that it can allow for the processing of unexpected input with a certain degree of confidence, provided that the training data is sufficient or through the use of smoothing techniques. For supervised learning, explicit rules and knowledge are not required, however, they do require training data for the supervised learning methods. In most cases, this training data needs to be edited by hand. The knowledge is obtained from the data and the learning is a process of estimating a set of parameters for a specific model. The statistical methods were unable to obtain any knowledge outside of the training data that the model processed (Liddy 2001; Chowdhury, 2003). Therefore, unsupervised learning has recently been proposed as a method of overcoming these limitations. Cui (2008) proposed the use of the unsupervised learning method for the automatic markup of biodiversity

literature. It has been argued that unsupervised learning might overcome the limitations of syntactic parsing and supervised learning while retaining their strengths. However, to date no system has been built which uses this method and thus no evaluation data is available.

2.3 Information Fusion

Kludas, et al. (2007) argued that Information Fusion (IF) "has established itself as an independent research area over the last decades." Many areas of research have found IF applications to be useful. The classical IF applications are in multimedia, multi-sensors, and meta-search information systems. Bloch, et al. (2001) stated that the application of information fusion is of particular importance in certain fields:

- Sensor fusion, particularly where multi-sensors are used. In such cases, information is collected from different sensors;
- Multiple-source information systems where several sources can provide precise, imprecise, or uncertain information about the object of interest;
- Expert opinion pooling where different individual statements are about the same event or object of interest are used.

It is appealing that information fusion has the potential to offer significant benefits by synthesizing different types of information across documents and thereby providing a global view. Information fusion will eventually allow us to analyze incomplete/uncertain information in a systematic manner in the hope that it will help resolve the differences.

The development of an information fusion system should include the following steps as showed in Table 2.2 (Torra & Narukawa, 2007). We will discuss how those steps were implemented in our system in later chapters.

Table 2.2: Steps in implementing fusion systems (Torra & Narukawa, 2007)

Stage	What	Method	Output
Acquisition	The process of gathering information from the information sources, determining the representation of the data	Passive (already exists) or active (when the fusion process starts, input starts)	The original data for fusion
Preprocessing	Preparation of the data for fusion	Noise reduction, sensor recalibration, filtering, etc.	Computationally appropriate data, same feature representation
Fusion Method	Define the fusion method to be used in execution	Function definition, selection (studying existing methods) and parameter determination	Fusion methods
Execution	Appropriate procedures are applied to data	A set of actions to be acted upon the data	Final result

Information fusion and domain knowledge. Domain knowledge is the knowledge obtained from the context or the domain of the input data/corpus. The central question that needs to be answered is: how do we know what constitutes domain knowledge and how will we encode the knowledge so that the system could take advantages of it?

Ontologies are explicit formal specifications of the terms in the domain and relationships among them (Gruber, 1993). The ontologies range from controlled vocabularies to highly expressive models (McGuinness, 2003). Ontologies are constructed by using at least two

essential components: a term vocabulary and the relationships between the terms. Ontologies can either be generic or domain-specific. It makes sense that domain ontologies should be used to guide the information fusion process given that domain knowledge is encoded in formal logical principles where induction is possible.
The following example from Bloch, et al. (2001) illustrates the concept of using domain knowledge in fusion.

S1: *The color of the object is blue.*

S2: *The color of the object is green.*

Domain knowledge: *Green and blue are different colors.*

In this case, we detect a conflict in our data.

It will be beneficial if at some point in the near future ontologies can be adopted in information fusion where the differences in structural, syntactic or semantic levels can be solved. However, there are requirements that must be met before the ontologies can be applied:

- A domain ontology must be available.
- The ontology must be encoded on the same level as the data abstraction level.
- An added layer that provides a mapping mechanism between the data and the ontology.

Cholvy & Hunter (1997) argued that the issue of how to handle domain knowledge is

closed related to handling conflicting information. Implicit knowledge can be handled by preference. In a variety of real-world projects in information fusion, domain knowledge is usually encoded in the form of constraints on the fusion process. Hunter published a series of publications regarding the construction of fusion methods in structured text (Hunter, 2000; Hunter, 2002; Hunter & Summerton, 2004; Hunter & Liu, 2006). The target was how to merge weather reports and a logic-based framework was adopted. As argued by Hunter (2002), before fusion procedures can be applied, background knowledge must be exploited and the fusion process will then become a monotonic process that develops a set of actions that define how the reports should be merged. Hunter's later papers examined the issue of how uncertainty information could be considered during the fusion process.

Wache, et al. (2001) reported that current information fusion systems with ontologies suffered from the lack of sophisticated methodologies that can support the development and advanced utilization of ontologies. The areas of evaluation and verification have attracted insufficient attention today.

The similarity between those research projects and this dissertation's research is that we are applying fusion rules to different kinds of text in different domains. The weather reports in Hunter's research were in the form of structured text, the vocabularies were much smaller and the required domain knowledge was less than taxonomic descriptions. In weather reports, the researcher can define the semantic relationship between each term in the vocabulary in order to define the fusion method, which is not possible in our

situation. More details are presented in chapter 3.

Information fusion in taxonomic descriptions. In information science research, information fusion has been intensively investigated on meta-search: combine/rerank the results from several different search systems/search engines (Fox, et al., 1993; Montague & Aslam, 2002; Wu & McClean, 2006; Efron, 2009).

Multi-document information fusion has not been the focal point of research. Some similar studies can be found in bibliographic data (Cowie, et al., 2000; Mann & Yarowsky, 2005). However, their studies focused on information extraction rather than true fusion. It is our hope that information fusion may prove useful for the purpose of improving the accessibility of information from multiple taxonomic sources.

Different descriptions are often used to describe the same object at various levels of generality (species, genus or family). Integrating information from different sources for the purpose of providing more complete information about the same object becomes more important. Due to the challenges of the non-uniform nature of taxonomic literature (Lydon, et al., 2003), information fusion has only recently become the focus of research with a few pilot projects dealing with multiple documents/sources having been conducted, e.g.,Wood, et al., 2004; Wang & Pan, 2006.

Wood et al. (2004) developed an ontology-based Information Extraction system known as MultiFlora. Their results show that retrieval performance significantly improved by

summing over six botanical descriptions. One of the results tripled recall, however no information was provided about precision. They also showed that failures in extraction from one text could be corrected by results from other texts. The correction rate was about 50 percent and no evaluation statistics were reported. Although Wood, et al. (2004) attempted to extract information from different floras, no attempt was made to merge the information together. The first attempt was conducted by Wang & Pan (2006).

Wang & Pan (2006) used an approach similar to Wood's (Wood, et al., 2004) involving construction of a generic shape function to determine how it matched with leaf shape. They selected 21 common shapes from *Botanical Latin*[4] and modeled them into a four-feature shape model. They simultaneously explored how to integrate and query leaf shape descriptions by taking advantage of ontologies. They extracted leaf shape information from the descriptions and built an ontology. Querying was conducted based on the ontology that had been developed. They argued that this method outperforms the keyword method by taking advantage of the semantics of shape. They found that their method could produce better results than those found using the keyword search method. While some statistical evaluations were provided, they were based on a limited sample of only 10 examples. They also found that their method failed to detect some the good results due to the strictness level of the ontology and that those "good results" were exactly what the searchers were looking for.

This dissertation's research is similar to that of Wang & Pan (2006) in the sense that we

[4] Stearn, W.T. (1973). Botanical Latin: history, grammar, syntax, terminology and vocabulary. David and Charles, Newton Abbot, England.

attempt to integrate information together from different sources. However, they focused exclusively on modeling a discrete variable into a continuous variable while this dissertation will focus on how to integrate different types of information together.

2.4 Summary

This chapter offers a brief overview of relevant works regarding CST and its identification, text mining in taxonomic literature and information fusion in related fields. A review of related research provides the context for the system design and evaluation in this dissertation. The following chapters will detail the data we will use, the system design, and how to evaluate the performance of the system.

Chapter 3: Data Acquisition and Preprocessing

The first steps to take when building an information fusion system are data acquisition and preprocessing. This chapter will detail the data acquisition and the preprocessing procedures. We will explore the information properties of the data that might impact our decisions regarding the implementation of the system. The sub-language features in this data set, and the consistency of the information embedded in different sources and different levels, will be exploited. We will examine the redundancy in parallel descriptions in order to produce an accurate, non-redundant, structured record of each species that will guide us through the preprocessing process.

3.1 Data Acquisition

Data sets that meet the following requirements are candidates for this research:

- taxonomic descriptions;
- available in digital format;
- have common descriptions of the same objects;
- hierarchical information organization structure.

Floras are selected because:

- They offer taxonomic descriptions of plants.
- Some prior information extraction studies have been conducted on floras, e.g.,

Cui, 2005; Cui & Heidorn, 2007.

- Floras have been and are freely available online at efloras[5]. Examples include Flora of China, Flora of North America, Flora of Texas and so on.
- Different floras have descriptions on the same species and genus.

Due to the exploratory nature of this study, FNA and FOC were selected as the data set to be used. Several important features led to this decision:

- Both floras are of great importance. "Flora of North America builds upon the cumulative wealth of information acquired since botanical studies began in the United States and Canada more than two centuries ago. The FNA Project will treat more than 20,000 species of plants [that are] native [to] or naturalized in North America north of Mexico, about 7% of the world's total. Both vascular plants and bryophytes are included". [6] Flora of China is "a collaborative international project to publish the first modern English-language account of the vascular plants of China (nearly 12% of the world's plants)" (Brach and Song, 2006).

- FNA and FOC share many species and genera. "Mainland China and the continental United States share a common latitude and similar-sized land areas. The climates in much of the two regions are also similar, especially in the eastern halves." [7] The similar climates also mean that species from one area that have

5 Accessible at http://www.efloras.org
6 FNA Introduction http://www.fna.org/introduction Retrieved on April 4, 2011
7 FOC online. http://hua.huh.harvard.edu/china/mss/plants.htm Retrieved on April

somehow migrated to the other can survive. There are few species that are native to both continents. In North America there are hundreds of introduced species from China and that is why there is overlap. Therefore, there are genera and species described in both FNA and FOC.

- Both floras have very clear written guidelines about information organization at different levels. For example, "constructions of keys, descriptions, and sequence of characters must be parallel for all taxa within a rank; e.g. within genus, subgenus or section, if leaf shape is mentioned in one species, it must be mentioned in all species."[8] This strict guideline for the construction of the descriptions ensures that the information flow between different levels is consistent.
- The text is OCR-error free (we had access to the digital source files). This will allow us to focus on the information extraction and fusion instead of mixing the problem together with OCR errors.

Each taxonomic description in FNA or FOC includes several parts: nomenclature, common names, naming history, morphological description, habitat and geographic distribution, etc. [9] Leaf descriptions inside morphological descriptions were selected as the focus of this study. The rationale for this is explained below. Morphological features

---

4, 2011

[8] Flora of North America Contributors Guide http://fna.huh.harvard.edu/files/FNA%20ContribGuide%202008.pdf Retrieved on April 4, 2011

[9] For example description, please see http://efloras.org/florataxon.aspx?flora_id=2&taxon_id=116064 Retrieved on April 4, 2011

contain the most important information that is useful for identification purposes. Most morphological features have the advantage of being easily seen. Hence, their variability has been more widely appreciated than other types of features. This is particularly true with herbarium materials on which most taxonomic work must be based (Stuessy, 1990). The early plant taxonomists relied almost exclusively upon morphology when classifying plants that were sent to them from many parts of the world. This remains true for a great deal of identification currently being conducted.

Morphological features, meaning the external features of a plant, can be categorized into two types: floral features and vegetative features. Floral features include flowers and other reproductive organs. Vegetative features include leaves, stems, roots and other organs. The difference between them is that the vegetative parts of plants are modular constructions in which there exist repeating units of structure that do not have a fixed numbers of parts. This contrasts with floral features that are more definite in number (Stuessy, 1990).

We chose morphological description of leaves as the target for our research for the following reasons:

(1) The vegetative parts of plants have multiple varied functions such as support, food production, water transportation, etc., which contrasts with the narrower role of floral features in reproduction. The result of these numerous functions is that vegetative features tend to be more plastic and/or variable and hence more difficult to use for taxonomic purposes (Stuessy, 1990).

(2) In comparison with stems and roots, leaves have been given the most attention and the variation in leaves are more significant than stems or roots. (Stuessy, 1990).

(3) There are few plants that lack leaves. There are plants that may appear to not have leaves, while actually having leaves. Possibly the leaves fell off when the plant matured or the leaves were too small to be readily noticed. In most cases, leaves have longer durations than flowers or fruits.

(4) Leaf blades are conspicuous and offer ease of observation along with obvious differences in size and shape. Leaf blades have been examined extensively in taxonomic studies. Different types of data exist that include leaf descriptions that include numeric, numeric-like and nominal variables.

(5) These findings are expected to be applicable to other organs such as flowers, stems, etc.

The species and genus level descriptions were downloaded from the website eFloras[10] which is a collection of online floras, including the Flora of China (FOC), Flora of North America (FNA), Flora of Missouri, Flora of Pakistan, and others. Figure 3.1 is a sample page from the eFloras website. The descriptions were downloaded using wget [11] from the eFloras.org website, which contains Volume 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 22, 24 of FOC and Volume 1, 2, 3, 4, 5, 19, 20, 21, 22, 23, 26, 27 of FNA. Some statistics about FNA and FOC can be found in Table 3.1.

[10] eFloras Accessible at http://www.efloras.org

[11] Wget Accessible at http://www.gnu.org/software/wget

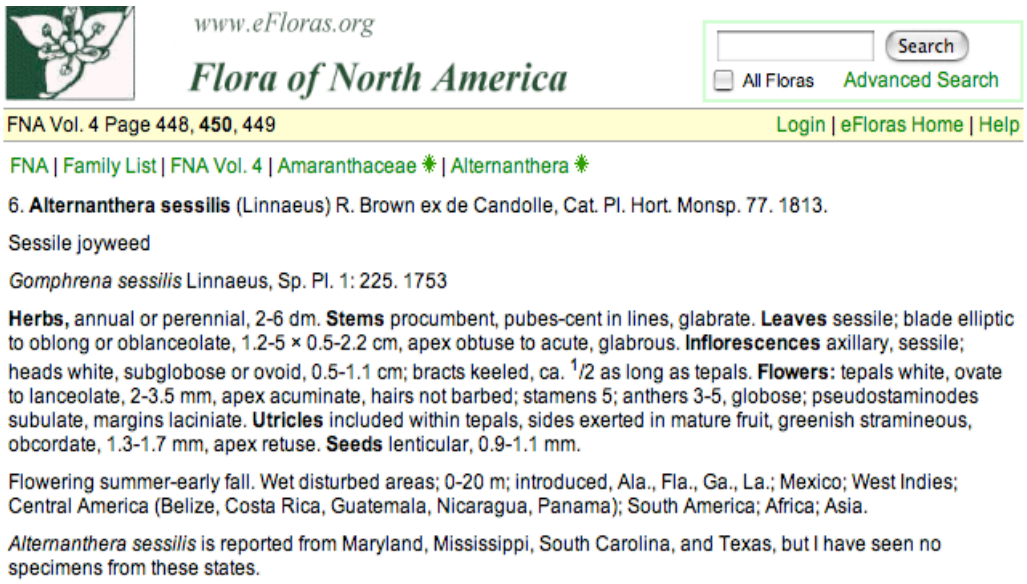

www.eFloras.org
Flora of North America
Search
All Floras Advanced Search
FNA Vol. 4 Page 448, **450**, 449 Login | eFloras Home | Help
FNA | Family List | FNA Vol. 4 | Amaranthaceae | Alternanthera
6. **Alternanthera sessilis** (Linnaeus) R. Brown ex de Candolle, Cat. Pl. Hort. Monsp. 77. 1813.
Sessile joyweed
Gomphrena sessilis Linnaeus, Sp. Pl. 1: 225. 1753
Herbs, annual or perennial, 2-6 dm. **Stems** procumbent, pubes-cent in lines, glabrate. **Leaves** sessile; blade elliptic to oblong or oblanceolate, 1.2-5 × 0.5-2.2 cm, apex obtuse to acute, glabrous. **Inflorescences** axillary, sessile; heads white, subglobose or ovoid, 0.5-1.1 cm; bracts keeled, ca. 1/2 as long as tepals. **Flowers:** tepals white, ovate to lanceolate, 2-3.5 mm, apex acuminate, hairs not barbed; stamens 5; anthers 3-5, globose; pseudostaminodes subulate, margins laciniate. **Utricles** included within tepals, sides exerted in mature fruit, greenish stramineous, obcordate, 1.3-1.7 mm, apex retuse. **Seeds** lenticular, 0.9-1.1 mm.
Flowering summer-early fall. Wet disturbed areas; 0-20 m; introduced, Ala., Fla., Ga., La.; Mexico; West Indies; Central America (Belize, Costa Rica, Guatemala, Nicaragua, Panama); South America; Africa; Asia.
Alternanthera sessilis is reported from Maryland, Mississippi, South Carolina, and Texas, but I have seen no specimens from these states.

Figure 3.1 Sample single treatment webpage from efloras.org

Table 3.1 Data set statistics

Flora statistics	FNA	FOC	In Both
Families	123	195	39
Genera	1157	1599	156
Species	3317	10252	153

The overlapping 153 species constitute the data set we will examine. They will then be processed through the proposed fusion system as described in Chapter 4.

We have 85 unique genera within our group of 153 samples in this data set. Among them, 3 do not have FNA genus level descriptions, and 3 do not have FOC genus level descriptions. So, at the genus level, we have a total of 164 different descriptions. For the species level, one does not have an FNA level description, and 12 do not have FOC species level descriptions. Therefore, we have a total of 293 species level descriptions. This information is presented in Table 3.2.

Table 3.2 Number of descriptions in our data set

Descriptions	FNA	FOC	Total
Genus	82	82	164
Species	152	141	293
Total	234	223	457

3.2 Data Transformation

Data transformation is a process where data is transformed or consolidated into forms that are appropriate for further processing. This is a standard procedure in data mining. The transformation can involve cleaning (removing noise), aggregation (applying aggregations to data), generalization (moving to higher-level features), and normalization (scaling the data within a specified range). Similar transformations are needed in text mining.

3.2.1 Data Cleaning

The importance of data cleaning cannot be over-emphasized. Data cleaning deals with "detecting and removing errors and inconsistencies from data in order to improve the quality of data" (Rahm and Do, 2000). This problem arises when data comes from different sources, or even a single source, where no input controls were imposed. This problem is also related to several data mining challenges: missing data, erroneous data, etc.

An online survey[12] has shown that 64% of researchers in data mining projects reported over 60% of project time was spent on data cleaning. Improved performance during this stage can contribute significantly to the success of the entire project.

Similar to data mining projects, most text mining projects involve data cleaning difficulties because it is difficult to impose any controls over the input of the data. There are usually few, or no, restrictions on what data can be entered and stored. It is sometimes the case that those issues in text mining are known as instance-specific problems because the errors and inconsistencies cannot be controlled on the schema level. Another challenge in text mining is the sparse nature of the data.

Although some techniques used in data mining are not available for use in text mining, the steps taken during the implementation of cleaning are the same. Standard data cleaning steps were proposed in Rahm & Do (2000). These steps are presented in Table 3.3.

Among the five steps involved in data cleaning, the analysis of the data is the most important. The knowledge obtained from analysis will determine the decisions made on the definition of the transformation rules and the outcome of the entire process. Several error patterns in our data are mentioned in Cui (2005). The errors include: missing values (e.g., the unit is missed), extra words or misspelled words, misuse of punctuation marks, misuse of the keyword "to" in character state transitions, and misplaced organ descriptions. It is true that such errors exist in our data set. However,

[12] Data Preparation http://www.kdnuggets.com/polls/2003/data_preparation.htm Retrieved on March 2, 2011

those error patterns are on the instance level, which is not very useful in our context since the correction of such errors would require manual verification. Our target is to find systematic patterns to guide us in the automatic cleaning process.

Table 3.3 Data cleaning steps in Rahm & Do (2000)

Steps	What	How to
Data analysis	A manual inspection of the data sample to obtain data properties and quality information	Data profiling (instance analysis: data type, length, value range, discrete values and frequency, variance, uniqueness, occurrence of null values, typical string patterns, etc.) and data mining (discover data patterns in data sets, e.g., relationships between attributes)
Definition of transformation workflow and mapping rules	Define how data will be transformed on the abstract level	Define schema- and instance- related mappings: can be single-step or multiple steps depending on different data sources. Steps can be specified using declarative queries and mapping languages
Verification	Verify the correctness and effectiveness of the transformation	Test on sample or copy of the data. Verify the correctness in some cases of conflict. Multiple iterations may be needed.
Transformation	Execute the transformation	Apply the transformation steps in the data, including standardization.
Backflow of cleaned data	Store corrected data	Replace “dirty” data

We used the method of data profiling that is mentioned in Rahm and Do (2000). A statistical analysis on our data set was conducted. The study focused on the data type and value range, together with their frequency, variance, and typical string pattern. We tokenize our descriptions by a single word that is the string between two spaces. We also removed the colons and periods. Table 3.4 is the table we obtained.

Generally speaking, the difficulties the cleaning process involves depend on the levels of

variance of the vocabulary that appear in the data. If the vocabulary for one element is quite small, e.g. only three categorical strings, then the cleaning process would differ from the element which has a large vocabulary size, e.g. three thousand categorical strings.

Table 3.4 Vocabulary size for each element

Element	Vocabulary Size	Element	Vocabulary Size
leafletlength	3	bladeshape	66
leafletwidth	3	apex	77
bladewidth	15	base	80
bladelength	19	lobe	104
solidshape	22	vein	111
attachment	28	stipule	130
tooth	28	petiole	134
texture	29	arrangement	155
duration	33	margin	191
complexity	47	surface	284
color	51	other	349

In many text mining projects, the input data is quite sparse, which means that when we have a term frequency matrix over the input data, the value we obtained in most cells is zero. Table 3.5 shows that some elements have larger variance than other elements. Elements such as shape, attachment, and texture have much smaller vocabulary than other, surface and margin.

We plotted the distribution of the term frequency and obtained the power law distribution pattern in our dataset. This means that when we plotted the term frequency and the frequency of the term frequency in a graph, it looks like Figure 3.2 and has a long tail. Figure 3.2 shows the term frequency distribution of element margin.

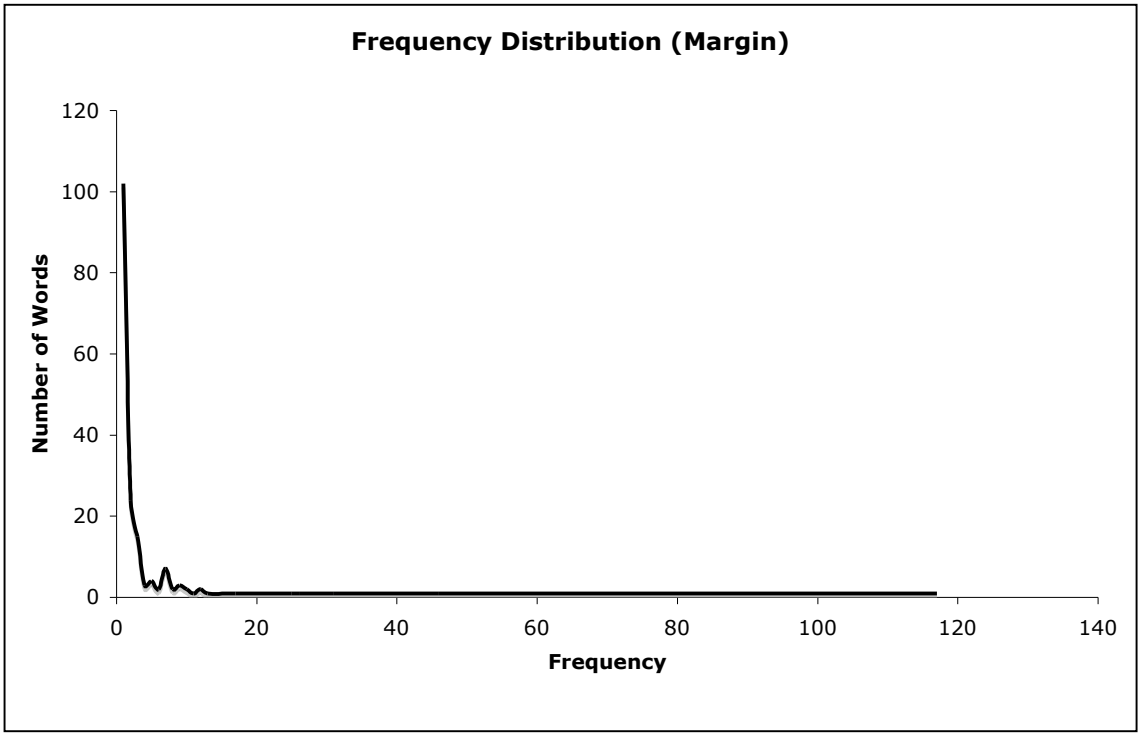


Figure 3.2 Term frequency distribution of Margin

Examining the terms in the vocabularies allows us to separate the terms we obtained into several categories:

(1) Frequently appeared but un-informative or the information contained is not considered in the fusion process. These terms include stop words and modifiers.

Stop words are frequently occurring words that do not carry specific meanings but instead provide support for the structure of the sentence. We used the list of stop terms listed on this site[13].

[13] Stop List http://www.lextek.com/manuals/onix/stopwords1.html Retrieved on February 2, 2011

Modifiers appeared frequently in our dataset. As summarized in SDD (Structure of Descriptive Data)[14] which is adopted as a standard by Biodiversity Information Standards (TDWG)[15], 6 different types of modifiers exist in the treatments. They are frequency, certainty, seasonal, diurnal, treateAsMisinterpretation and otherModifer. Three out of 6 types of modifiers appear frequently in our dataset.

We found another type of modifier in our data set that is not included in SDD. It can be called the "length" or "location" modifier. It appears in scenarios such as "shortly (or long) petiolate." Those modifiers are presented in Table 3.5.

Table 3.5 Types of modifiers

Modifiers	Range	Example
Frequency	Between 0-1	Rarely, usually
Certainty	Between 0-1	Probably, perhaps
Seasonal	Dates	When mature
Length	Categories	Shortly, long, broadly, narrowly

Generally speaking, the modifiers are not the only type of the adverbs used to provide supporting information. Those adverbs usually end with "-ly", e.g., commonly, distinctly, generally and frequently.

Information fusion with modifiers presents different challenges that are discussed in

[14] Structured of Descriptive Data http://wiki.tdwg.org/SDD Retrieved on February 2, 2011

[15] Biodiversity Information Standards (TDWG), also known as the Taxonomic Databases Working Group, more information could be found at http://www.tdwg.org/ Retrieved on February 2, 2011

detail in Chapter 8. For this dissertation project, the information contained in the modifiers is not considered in the entire fusion process. Thus, the modifiers are treated similar as stop words in the entire process.

(2) Frequently appeared and informative. The terms in this category are mainly adjectives, e.g., alternate, acute, green. Those terms contain the most important information we want to capture in our system.

(3) Less frequently appearing terms. The appearance of these terms can be categorized into two types:

- Systematic. The reason is that the expressiveness of the schema we are using is not as expressive as the original description. For example, element Arrangement, includes both leaf-leaf arrangement and leaf-other organ arrangement information.
- Non-systematic. Various authors chose to use different vocabulary words when describing the same character. The production of these treatments requires the personal interpretation of the authors, which leads to variations in the descriptions. During the later processing stage, some terms in this category will be removed.

Excluding the stop words and modifiers, we are left with smaller vocabularies for the elements. Table 3.6 presents the vocabulary size before and after the data cleaning for each element.

Table 3.6 Vocabulary sizes after data cleaning

Element	Vocabulary Size (Before)	Vocabulary Size (After)	Element	Vocabulary Size (Before)	Vocabulary Size (After)
leafletlength	3	3	bladeshape	66	45
leafletwidth	3	3	apex	77	48
bladewidth	15	6	base	80	50
bladelength	19	9	lobe	104	71
solidshape	22	12	vein	111	72
attachment	28	15	stipule	130	95
tooth	28	17	petiole	134	96
texture	29	18	arrangement	155	100
duration	33	20	margin	191	133
complexity	47	34	surface	284	220
color	51	37	other	349	244

3.2.2 Data Normalization

After data cleaning, we then proceed to data normalization. Note that the normalization process might change the original data into a new form.

Numerical variables. Numerical variables include elements blade length, blade width, leaflet length and leaflet width. The normalization process includes:

- Changing the reference value to an actual value. In the descriptions, some property information is stated relative to other properties. For example, the string, "the width is narrower than or as wide as long", does not explicitly state the value of width. However, the information can be parsed out once we know the length. Similar patterns such as "more than," "less than," "as wide as," "equal" or "as long as" represent the reference relationship between elements. The value of the target element is based on the information about another element. We replaced the reference information with the actual value when preprocessing the descriptions

of those numerical variables.

- Normalize the unit used in the elements. In most descriptions of blade length or leaflet length, the unit is missing. In the contributors' guide for Flora of North America, it documents that the unit only appears at the end of the string[16]. Thus, when we see the description of leaf "1-5.5(-12) × 0.5-3.8(-8) cm", it also means that the unit for both length and width is "cm." Therefore, we manually attached the proper unit to the length elements.

We simultaneously saw that several different units were used: cm, mm, dm, etc. "cm" and "mm" are the most popular units in use. Therefore, we transformed all of the units to "mm" and changed the numbers accordingly because "mm" appears most frequently in our dataset.

Numerical-like variables. We found that blade shape was described in different ways in the descriptions. Table 3.7 shows some examples and the patterns they revealed.

We observed that they have similar patterns: "shape1 to shape2", "shape1 or shape2", "shape1-shape2", "shape1 and shape2", and "shape1, shape2". David Boufford, a plant taxonomist explained (personal communication with Dr. David Boufford, Research Taxonomist, Harvard University Herbarium)

> "I believe they all mean a range of shapes between the ones mentioned. It would be unlikely that there would be gaps between the various shapes", "The

[16] Flora of North America Contributors Guide http://fna.huh.harvard.edu/files/FNA%20ContribGuide%202008.pdf Retrieved on April 4, 2011

difficulties arise when different authors try to describe the same features. Each person is somewhat different in the way he or she sees and describes something. Although they all try to explain the same feature, the format and wording of how they do it may vary somewhat."

Table 3.7 Leaf shape descriptions, pattern and meaning

Original Descriptions	Patterns
"obovate, elliptic, or spatulate,"	Shape1, shape2 or shape3
"rhombic-ovate, ovate, or elliptic to broadly lanceolate,"	Shape1-shape2, shape3 or shape4 to shape5
"ovate, rhombic-ovate, or lanceolate,"	Shape1, shape2-shape3 or shape4
"elliptic to oblong or oblanceolate,	Shape1 to shape2 or shape3
"obovate to narrowly spatulate"	Shape1 to shape2

Therefore, all of the patterns we found represent the same meaning. The strings "shape1 to shape2", "shape1 or shape2", "shape1-shape2", "shape1 and shape2", and "shape1, shape2" have the same meaning which is "shape1, shape2 and all of the shapes in between." During the pre-processing, we transformed all of the strings to "shape1 shape2".

3.3 Summary

This chapter details the source data we used and the characteristics of the data. The data cleaning and preprocessing were also discussed in detail. The following chapters will present our system design step by step.

Chapter 4: Methods and System Design

Providing a single access point to an information system that has been constructed from multiple documents is helpful for biodiversity researchers and this is true in other fields as well. It saves the time that would normally be taken going back and forth between different sources and provides the opportunity to semi-automatically? detect conflicts in the information found among different sources. The complementary information in different sources and levels of descriptions can be combined to improve data quality. This is the goal of information fusion. The purpose of the proposed fusion system is to demonstrate that an information fusion system is feasible using currently available text mining techniques. We will use this system to attempt to identify the factors that impact the performance of the system.

4.1 Overall System Design

The proposed fusion system contains 4 major modules: Text Segmentation and Taxonomic Name Identification, Organ-level and Sub-organ level Information Extraction, CST- Relationship Identification, and Information Fusion. The overall system design is shown in figure 4.1. The last two modules, CST-relationship identification and information fusion, will be discussed in detail in Chapter 5 and Chapter 6.

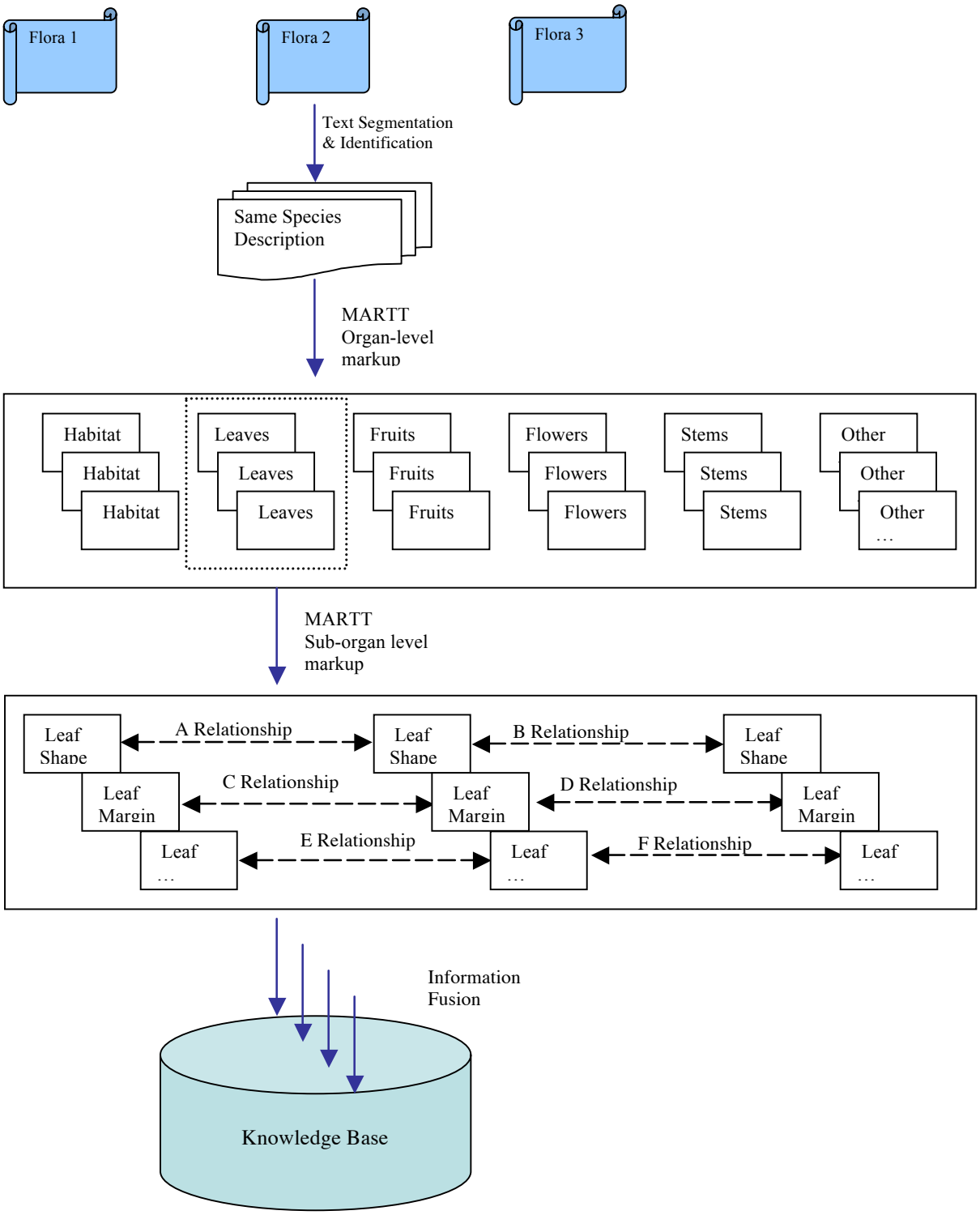


Figure 4.1 Information fusion system architecture

4.2 Text Segmentation and Taxonomic Name Identification

In the previous chapter, Figure 3.1 showed that the downloaded flora files consist of individual web pages that contain multiple kinds of information about particular families, genera or species, including volume information, taxonomic descriptions, keys, lower taxa, related objects, etc. In our system, each description is saved as a separate file. Text segmentation is accomplished by removing all of the other information from the web page source. Each description is then stored in a separate text file with the name of the flora and the taxonid (the taxonid is used by eFloras as a unique identifier for each description in each flora). Table 4.1 shows a sample text file for species *Alternanthera sessilis* (Linnaeus) R. Brown ex de Candolle after removing all unrelated information from the webpage but preserved flora, volume, family name, genus name, species name, species author name, first publication and description information.

Table 4.1 Sample text file of a single species level description

|Flora | Flora volume | family name | genus name | species name| Species Author Name |First publication | Description

|Flora of China | 4 | Amaranthaceae | Alternanthera | sessillis | Linnaeus R. Brown ex de Candolle | Cat. PI. Hort. Monsp 77. 1813 | **Herbs,** annual or perennial, 2-6 dm. **Stems** procumbent, pubes-cent in lines, glabrate. **Leaves** sessile; blade elliptic to oblong or oblanceolate, 1.2-5 × 0.5-2.2 cm, apex obtuse to acute, glabrous. **Inflorescences** axillary, sessile; heads white, subglobose or ovoid, 0.5-1.1 cm; bracts keeled, ca. $^{1}/2$ as long as tepals. **Flowers:** tepals white, ovate to lanceolate, 2-3.5 mm, apex acuminate, hairs not barbed; stamens 5; anthers 3-5, globose; pseudostaminodes subulate, margins laciniate. **Utricles** included within tepals, sides exerted in mature fruit, greenish stramineous, obcordate, 1.3-1.7 mm, apex retuse. **Seeds** lenticular, 0.9-1.1 mm.

Taxonomic name identification is done at the same time as the segmentation. As we can see from table 4.1, we preserved family name, genus name and species name information in the text file.

Table 4.2 Single species leaf descriptions from both FNA and FOC genus and species level

<fnagenus>
Leaves alternate, opposite, or whorled; stipules caducous, free. Leaf blade ovate, lobed or entire, margins dentate; venation appearing palmate or weakly 3-veined from base.
</fnagenus>
<focgenus>
Leaves alternate, spirally arranged or distichous; leaf blade simple to palmately lobed, margin toothed; primary veins 3-5 and plinerved, secondary veins pinnate.
</focgenus>
<fnaspecies>
Leaves: stipules ovate to ovate-oblong, apex attenuate; petiole shorter than or equal to blade. Leaf blade entire or 3-5-lobed, 6-20 x 5-15 cm, base shallowly cordate, often oblique, truncate, or broadly rounded, margins serrate, apex acuminate; surfaces abaxially densely gray-pubescent, adaxially scabrous.
</fnaspecies>
<focspecies>
Leaves spirally arranged; petiole 2.3-8 cm; leaf blade broadly ovate to narrowly elliptic-ovate, simple or 3-5-lobed on young trees, 6-18 x 5-9 cm, abaxially densely pubescent but veins with coarser hairs, adaxially scabridulous and sparsely pubescent, base cordate and asymmetric, margin coarsely serrate, apex acuminate; secondary veins 6 or 7 on each side of midvein.</focspecies>

The relationships between the descriptions are identified by comparing their scientific names. Since we want to merge the information from different levels, we need to identify the genus description for each species or whether two species are the same species. For example, we want to know if two species are the same species or under the same genus or family. We can obtain the information by comparing whether the two species have the same family name and genus name. By comparing the scientific names they have, we are

able to identify the genus description for each species. Table 4.2 shows the text for a single species while containing the genus description after we extracted leaf descriptions (organ-level information extraction is discussed in the next section) from the whole plant descriptions.

4.3 Information Extraction via Natural Language Processing

The organ-level and sub-organ level information extraction involves using natural language techniques to extract the descriptions of the leaves from other descriptions. At the same time, we want markup the extracted leaf descriptions into a designed schema. The schema in Appendix A is based on the schema used in Cui and Heidorn (2007). The major change is the addition of label names. For example, in a description such as "*Petiole 3-6 mm, pubescent to glabrous,*[17]" the word "*Petiole*" is a label name. This means that the text that follows represents the information about it, although the word "*Petiole*" itself does not contain the information.

Natural language processing (NLP) techniques were chosen as the method to use to conduct information extraction. Generally speaking, natural language processing can be either rule-based or statistical- based. (More information about NLP parsing appears in the Chapter 2 Literature Review)

[17] Microdesmis casearifolia Planchon ex J. D. Hooker description in Flora of China (FOC) http://efloras.org/florataxon.aspx?flora_id=2&taxon_id=242332415 Retrieved March 2, 2011

Rule-based NLP can occur on three levels: lexical, syntactic and semantic. Taylor, (1995) and Abascal and Sánchez (1999) proposed using the rule-based method to automatically extract information from descriptions. The major problem with rule-based systems is that most of the rules that have been used have been handcrafted and cannot easily be transformed for use in another scenario or are not useful when used within the context of a broader field (Hahn and Wermter, 2006).

In comparison with the rule-based method, statistical methods have obtained greater popularity in the text mining community. Statistical methods take no advantage of language itself. Statistical methods have been argued to offer both better performance and portability (Rosenfeld, 2000). The features and models used in one corpus can be readily transformed to other corpora without major modifications. The MARTT (Markuper of Taxonomic Treatments) system proposed in Cui and Heidorn (2007) is a system that uses statistical models. The system was tested on descriptions drawn from Flora of North Texas by using training examples drawn from Flora of China and Flora of North America. The results were promising as the system obtained a F-measure on the leaves element of 99.2% for FNA and 98.5% for FOC.

MARTT will be used to conduct the organ level information extraction in our system (permission was obtained from the developer). The main goal for organ level information extraction is to separate the leaf description from all other organs, e.g., flowers or stems. Since the focus of this dissertation project is on the information fusion step, we want to eliminate errors from the earlier steps. Thus, once the leaf descriptions are extracted from

the whole plant descriptions, we then conducted a quick check on the errors and corrected any errors that had been produced. The corrected output from MARTT organ level information extraction became the starting point for sub-organ level information extraction.

The sub-organ level or character-level (a character describes the property of the object, e.g., leaf shape, leaf arrangement) information extraction will also be conducted in a manner similar to the organ-level information extraction. MARTT also contains a module for marking up sub-organ level descriptions that we will use. The author hand-corrected any errors produced during this stage.

The output of this stage will be an xml file that contains both genus and species level descriptions from FNA and FOC genus and species levels. Below is a sample xml file after the sub-organ level information extraction.

<?xml version="1.0" encoding="UTF-8"?>

<?oxygen RNGSchema="http://www3.isrl.illinois.edu/~qinwei2/leaves.rng" type="xml"?>

<Leaves>

<FNA_genus>

<leavesla>Leaves </leavesla><arrangement>opposite, </arrangement><attachment>sessile or petiolate; </attachment><bladela>blade </bladela><bladeshape>lanceolate to ovate, ovate-rhombic, or obovate-rhombic,

</bladeshape><marginla>margins </marginla><margin>entire.</margin>

</FNA_genus>

<FOC_genus>

<leavesla>Leaves </leavesla><arrangement>opposite, </arrangement><marginla>margin

</marginla><margin>entire.</margin>

</FOC_genus>

<FNA_species>

<leavesla>Leaves </leavesla><attachment>sessile;</attachment><bladela> blade

</bladela><bladeshape>elliptic to oblong or oblanceolate,

</bladeshape><bladelength>1.2-5 </bladelength><ns>x</ns><bladewidth> 0.5-2.2 cm,

</bladewidth><apexla>apex </apexla><apex>obtuse to acute,

</apex><surface>glabrous.</surface>

</FNA_species>

<FOC_species>

<petiolela>Petiole </petiolela><petiole>1-4 mm,</petiole><petiolesurface> glabrous or

pilose; </petiolesurface><bladela>leaf blade </bladela><bladeshape>linear-lanceolate,

oblong-obovate, or ovate-oblong, </bladeshape><bladelength>1-8

</bladelength><ns>x</ns><bladewidth> 0.2-2 cm, </bladewidth><surface>glabrous or

pilose, </surface><basela>base </basela><base>attenuate, </base><marginla>margin

</marginla><margin>entire or slightly serrate,</margin><apexla> apex

</apexla><apex>acute or obtuse.</apex>

</FOC_species>

</Leaves>

4.4 Summary

This chapter introduced our overall system design and the implementation of the first two modules. The output of these two steps that is in the form of an xml file that contains marked-up information on the sub-organ/character level. We will detail another two modules in the following two chapters.

Chapter 5: Cross-Description Relationships Identification

This chapter will define the Cross-Description Relationships (CDR) within our data and how we can automatically classify them using machine learning. The ultimate goal for relationship identification is to facilitate the information fusion process.

We have 3 different types of variables in our dataset. We will apply the fusion operators directly to numeric and numeric-like variables. For the categorical variables, fusion will be based on cross-description relationships. The details of the CDR identification process are described in this chapter. The implementation will address the problems concerning feature selection and model selection. An evaluation of the relationship identification is reported at the end of this chapter. The classification programs are written in java and developed in NetBeans[18] 5.5.1 and JDK[19] 5. The learning module comes from WEKA[20].

5.1 CDR Relationships

Cross-document Sentence Theory (CST) is introduced in Chapter 2. It is a method of analyzing the semantic relationships between sentences from different documents. There are 24 different relationships that have been identified in CST.

---

[18] NetBeans. Accessible at http://netbeans.org/

[19] Java Development Kit (JDK). Accessible at http://www.oracle.com/technetwork/java/javase/downloads/index-jdk5-jsp-142662.html

[20] Weka is a collection of machine learning algorithms in JAVA. Accessible at http://www.cs.waikato.ac.nz/ml/weka/

5.1.1 CDR Relationships in the Data

The text we use in this research consists of taxonomic descriptions and they appear in sub-language. Strictly speaking, "complete sentences" do not exist in the descriptions but rather phrases.

It is not the case that all 24 relationships exist in our data. We made manual examinations of each variable in the descriptions, and successfully identified 7 relationships out of the 24 relationships that frequently appear.

Those 7 relationships that exist between the description pairs constitute Cross-Description Relationships (CDR). The only change we made is that the "agreement" relationship is changed to "overlap" relationship. In our cases, the "agreement" relationship would include "identity", "equivalence", and "overlap." The CDR relationships are listed in Table 5.1, including both inter-level and intra-level relationships. Table 5.2 shows some instances of these relationships and Table 5.3 shows some characteristics of the relationships.

Table 5.1 CDR Relationships and their information property

(A and B mean the string in Description 1 and Description 2, respectively. X and Y mean the information included in A and B, respectively. They include sets of information items. Small x and small y represents the information items in X and Y)

Relationships	**Explanation**	**Information properties**
Identity	Contain the same text.	A=B and X=Y
Equivalence	Same meaning with different text	For any x in X, you can find a y in Y where x=y and for any y in Y, you can find a x in X where y=x while A!=B
Subsumption	One description contains more information than the other description.	For any x in X, you can find a y in Y where y=x and for some y in Y, you can not find a x in X where x=y, and those y's together are non-trivial while A!=B
Complementary	Totally different information/ properties	For any x in X, you can not find a y in Y where x=y and for any y in y, you can not find a x in X where y=x while A!=B
Overlap (Partial Equivalence)	Two descriptions contain some information that is the same.	For some x in X, you can find a y where y=x, those xs are non-trivial. For some y in Y, you can find some x in X where x=y, those ys are non-trivial. For some x in X, you can not find a y in Y where y=x and For some y, you can not find a x in X where x=y, those xs + ys together are non-trivial. (It could be the case that x's or y's alone is trivial) while A!=B
Conflict	Conflict information	For some x in X, there is a y in Y where y=not x while A!=B
Refinement	More specific and detailed information	For any x in X, you can find a y in Y where y=x. For some y in Y, you can not find a x in X where x=y, and those y's together are non-trivial while A!=B

Table 5.2 Instances of CDR Relationships

Relationships	Instances
Identity	S1: Leaves [arrangement] alternate. S2: Leaves [arrangement] alternate.
Equivalence	S2: [arrangement] alternate, spirally arranged or distichous; S2: [arrangement] alternate, opposite, or whorled;
Subsumption	S1: [margin] entire, dentate, sinuate, or serrate, occasionally lobed S2: [margin] entire or irregularly serrate or lobed
Complementary	S1: [surface] flattened, not jointed, not spinose S2: [surface] light green on both surfaces
Overlap	S1: [margin] entire, dentate, sinuate, or serrate S2: [margin] serrate or subentire
Conflict	S1: [color] not glaucous S1: [color] glaucous
Refinement	S1: [base] truncate, cordate, hastate, or cuneate(genus level) S2: [base] cuneate(species level)

Table 5.3 Characteristics of CDR relationships

Relationships	Co-exist	Directional
Identity	No.	No.
Equivalence	No.	No.
Subsumption	No.	Yes.
Complementary	No.	No.
Overlap	Yes.	No.
Contradiction	Yes.	No.
Refinement	No.	Yes.

5.1.2 CDR Relationships Distribution

Lydon, et al. (2003) conducted a pilot study regarding information organization patterns in taxonomic descriptions. This study is inspiring but far from systematic and comprehensive but attempts to reveal the underlying structure using cross-description relationships.

In order to examine the distribution of CDR relationships in our dataset, we randomly

selected 10 species from the 153 species that are both included in FNA and FOC. Detailed information about the 10 species is listed in Table 5.4. They are scattered around in 6 families and 8 genera. Processing the files through the first two steps of the designed fusion system and correcting the errors obtained from information extraction resulted in 10 xml files containing the descriptions from the genus level and the species level from FNA and FOC.

Table 5.4 Sample information

Taxon_id	Family	Genus	Species
200006341	Moraceae	Broussonetia	papyrifera
200006809	Chenopodiaceae	Chenopodium	album
200006814	Chenopodiaceae	Chenopodium	chenopodioides
200006975	Amaranthaceae	Alternanthera	philoxeroides
200006977	Amaranthaceae	Alternanthera	sessilis
200006986	Amaranthaceae	Amaranthus	retroflexus
200008245	Ranunculaceae	Thalictrum	sparsiflorum
200008288	Lardizabalaceae	Akebia	quinata
242000840	Caryophyllaceae	Spergularia	diandra
242412161	Amaranthaceae	Alternanthera	paronychioides

One of the purposes of fusion is to generate new knowledge by merging information in different sources and levels. We will explore the potential of inter-level and intra-level information fusion by combining complementary information.

Two types of complementarity are possible when the information comes from different sources or levels. They are:

(1) Different properties (characters) are mentioned in different sources or levels. Lydon et al. (2003) found that there are significant amounts of information (around 55%) that can only be found in different sources or levels.

(2) The same properties are described in different sources or levels, but contain overlapping, contradictory or complementary information. No new information will be generated in relationship Identity, Equivalence, Subsumption or Refinement.

We are interested in both inter-level and intra-level relationships which includes 4 pairs of cross-description relationships (FNA_genus, FOC_genus), (FNA_genus, FNA_species), (FOC_genus, FOC_species), (FNA_species, FOC_species). Breaking down the relationship distributions among different levels and sources, we were able to identify the sources and quantity of complementary information.

Different properties (characters) are mentioned in different sources or levels. The overall information on how many different properties described in different sources or levels is presented in Table 5.5. We observe that there is a significant amount of information that comes from only one source.

Table 5.5 Overall information on different characters

Number of Characters	**(FNA_Genus, FOC_Genus)**	**(FNA_Species, FOC_Species)**	**(FNA_Genus, FNA_Species)**	**(FOC_Genus, FOC_Species)**
Total Characters	61	95	96	97
Characters in Different Sources or Levels	30	38	63	79
Characters from Single Source(%)	49.18%	40.00%	65.63%	81.44%

There are a total of 61 characters that are mentioned at Genus level; 31 of these characters are mentioned in both floras; 30 of these characters are mentioned in only one of the floras. Therefore, about 50% of the information can only be found by combining the information in different sources. The distribution of the 30 characters is presented in Table 5.6.

Table 5.6 Characters at the genus level in different floras

Element	FNA_Genus	FOC_Genus
Apex	4	0
Arrangement	1	1
Attachment	4	0
Base	3	0
Blade shape	7	0
Lobe	2	0
Margin	0	1
Other	3	0
Petiole	0	1
Stipule	1	0
Surface	1	0
Vein	1	0
Total	27	3

A total of 95 characters are mentioned as being at the species level; 57 characters are mentioned as being in both floras; 38 characters are mentioned in only one of the floras.

Table 5.7 below shows the distribution of the 38 characters on the species level from different floras. Among them, 14 characters are from FNA species description and 24 of them are from FOC species description.

Table 5.7 Characters on the species level from different floras

Element	FNA_Species	FOC_Species
Apex	0	1
Arrangement	1	1
Attachment	4	0
Base	0	3
Blade shape	1	1
Blade width	0	2
Duration	0	1
Lobe	0	2
Margin	1	3
Other	3	0
Petiole	2	3
Petiole surface	0	3
Stipule	1	0
Surface	1	3
Vein	0	1
Total	14	24

A total of 96 characters are mentioned as being at the FNA genus level and species level; 33 characters are mentioned as being on both levels; 63 characters are mentioned as being in only one of them; 25 characters are only on the genus level. The 25 characters are presented in Table 5.8.

Table 5.8 The characters only on FNA genus level

Element	FNA_Genus
Arrangement	7
Attachment	4
Blade shape	1
Lobe	4
Margin	4
Other	2
Surface	1
Vein	2

A total of 97 characters are mentioned as being on the FOC genus level and species level;

18 characters are mentioned as being on both levels; 79 characters are only mentioned as being in one of them; 16 characters are only from the genus level.

Table 5.9 The characters only on the FOC genus level

Element	FOC_Genus
Arrangement	8
Attachment	4
Blade shape	1
Margin	2
Other	1

Tables 5.6, 5.7, 5.8 and 5.9 show that a significant amount of information could only be found in different sources or levels. This is true for both the genus level (about 1/2) and the species (about 1/3) level. The same phenomenon can be found at different levels.

We observe that the same properties are mentioned as being present in different floras and levels. Table 5.10 contains more detailed information. We will apply the 7 CDR relationships in Table 5.1 to the pairs of descriptions and determine the details of their distributions.

Table 5.10 Same properties in different sources or levels

Number of Characters	(FNA_Genus, FOC_Genus)	(FNA_Species, FOC_Species)	(FNA_Genus, FNA_Species)	(FOC_Genus, FOC_Species)
Total Characters	61	95	96	97
Same Characters	31	57	33	18

Table 5.11 CDR Relationships distribution on the genus level in different floras

Relationship	Count	Percentage
Identity	14	43.75%
Equivalence	1	3.13%
Subsumption	10	31.25%
Complementary	1	3.13%
Overlap	5	15.63%
Contradiction	1	3.13%
Total	32	100.00%

Table 5.12 CDR Relationships distribution on the species level in different floras

Relationship	Count	Percentage
Identity	2	3.64%
Equivalence	4	7.27%
Subsumption	24	43.64%
Complementary	2	3.64%
Overlap	22	40.00%
Contradiction	1	1.82%
Total	55	100.00%

Table 5.13 CDR Relationship distribution on different levels in FNA

Relationship	Count	Percentage
Identity	0	0.00%
Equivalence	3	8.11%
Complementary	3	8.11%
Overlap	10	27.03%
Contradiction	2	5.41%
Refinement	19	51.35%
Total	37	100.00%

For the characters mentioned at the genus level, the distributions of the 7 relationships (Relationship Refinement only appears in different levels, Subsumption only appears in the same level) are shown in Tables 5.11 through 5.14. Complementary information can be due to Relationship Overlap, Contradiction, and Complementary. They consist of nearly 40% - 50% of the total number of relationships. Note that multiple relationships

may exist in a single pair of descriptions.

Table 5.14 CDR Relationships distribution on different levels in FOC

Relationship	Count	Percentage
Identity	1	5.00%
Equivalence	0	0.00%
Complementary	3	15.00%
Overlap	3	15.00%
Contradiction	3	15.00%
Refinement	10	50.00%
Total	20	100.00%

In summary, by combining information in different sources and levels we are able to add about 60-75% new information to the original descriptions, of which 50% could be added when the character is only described in another source and level and 10-25% can be gained while complementary, overlapping and contradictory relationships hold.

5.2 Automatic CDR Relationships Classification

CDR Relationships identification can be viewed as a standard classification problem. Suppose we are given two descriptions S1 and S2 and S1 and S2 describe the same property of the same object. We are interested in determining which CDR relationship could be applied to the two descriptions. The classification can be one of the following two cases:

- Binary classification. In this scenario, suppose that we have a relationship *x*. *x* could be any one of the 7 relationships in CDR. The question is whether the

relationship between S1 and S2 is *x*. If there does exist a relationship *x* between S1 and S2, then *x*(S1, S2)=1, otherwise *x*(S1, S2)=0.

- Multi-class classification. In this scenario, suppose that we have multiple relationships, *x, y, z* ..... The question here is that given S1, S2 and relationships *x, y, z……,* which relationship exists between S1 and S2. If the relationship between S1 and S2 is *x,* then *R(S1, S2) = x.*

We must note that in the binary classification, the relationships are not mutually exclusive of each other. This means that the result of the classification might be that there exist several relationships that can be identified with the same description pairs. In the multi-class scenario, each instance can only be classified to one class. Except for this difference, the two scenarios are fundamentally the same.

Note that in our data, only overlapping and contradictory relationships can co-exist with each other. However, Relationship contradiction is not included in the automatic classification process, and all other relationships are mutually exclusive of each other. Therefore, multi-class classifications fit better to our task and are conducted in our system.

As discussed above, relationship identification is a typical classification problem and machine learning can be argued to be an effective method of addressing it (Peng, et al., 2004; Langseth & Nielsen, 2006). Satisfactory performances were obtained in various classification problems (Wang, et al., 2005; Bellegarda, 2004; Peng, et al., 2004;

Langseth & Nielsen, 2006). The machine learning method will be used in our system. The problem we have here is to find the appropriate feature set and appropriate machine learning model.

5.2.1 Feature Selection

It has been argued that feature selection is the most important step in machine learning applications (Zweigenbaum, et al., 2007). Feature selection is a process in which we remove irrelevant or redundant information from the original data set in order to improve the performance. The selection process can include several processes, e.g., normalization, transformation, association and filtering.

Feature selection usually begins with the process of finding all of the candidate features that might potentially be useful for the purpose of separating the instances into different classes. It then searches for the optimal subset out of all of the candidate features. The reasons for feature selection include: improving the performance of the classifier, reducing the dimension complexity of the data, and simplifying the classifier in order to avoid overfitting. There are two common approaches to conducting reduction: the wrapper approach and the filter approach.

The wrapper approach uses the intended learning algorithm itself to evaluate the usefulness of the features. The most common wrapper approach technique is known as cross validation that involves using a statistical re-sampling technique where the tests are conducted using a different subset of the data several times. Five-fold and ten-fold cross

validations are popular. The performances are determined by the average performance over the iterations. The process ends with subsets of features for which the best performance has been identified.

The filter approach is independent of the learning algorithm, and involves using heuristics based on the characteristics of the data. In this case, prior knowledge is quite important. Any known relationship that exists between the data or the feature can be used to test the features. One of the methods involves removing all of the features whose information has been subsumed by any combination of other features (Koller and Sahami, 1996).

There is some related research that has been conducted using machine learning to automatically classify cross-document sentence (CST) relationships.

Zhang, et al. (2003) and Maziero, et al. (2010) similarly attempted to classify news into a subset of CST relationships.

Miyabe, et al. (2008) attempted to classify sentences in newspaper articles concerning the same topic into two relationships: “equivalence” and “transition.”

Hatzivassiloglou, et al. (1999) attempted to detect similarities between small textual units using machine-learning methods. The features used in those works are presented in following table. Not all of the works reported the performance of their systems.

Table 5.15 Features used in automatic CST classification

Projects	Lexical	Syntactic	Semantic
(Zhang, et al., 2003)	1. number of tokens in sentence 1 2. number of tokens in sentence 2 3. number of tokens in common	1. number of tokens having POS x in sentence 1 2. number of tokens having POS x in sentence 2 3. number of common tokens having POS x	1. find the most prominent concepts discussed in each sentence pair and compute the distance using WordNet
(Maziero, et al., 2010)	1. difference in length of sentences (in number of words) 2. percentage of common words in sentences 3. position of each sentence in the text that it belongs to 4. whether a sentence is shorter than the other 5. whether the sentence is identical	1. number of nouns 2. number of proper nouns 3. number of adverbs 4. number of adjectives 5. number of verbs 6. number of numerals	
(Hatzivassiloglou, et al.,1999)	1. word co-occurrence 2. shared proper nouns (entity) 3. matching noun-phrases		1. WordNet synonyms 2. common semantic classes for verbs
(Miyabe et al., 2008)	1. cosine similarity measure 2. normalized lengths of sentences 3. position of the sentence in the document 4. the number of words and phrases 5. head verb	1. named entity 2. named entity type 3. numeric expression and units	1. semantic similarities by the frequency vectors of semantic classes of nouns, verbs, and adjectives. 2. salient words

Zhang, et al. (2003) obtained different performances in different relationships where the F-measure ranges from 39% to 5%. Maziero, et al. (2010) reported the F-measure as 94% on the "identity" relationship and the performance on other relationships ranged from 63% to 33%. Miyabe, et al. (2008) reported the highest F-measure performance on "equivalence" as being 75.5%. We found that in those related works, different performances in different relationships were obtained because the behavior of one CST relationship is quite different than another. Examining the examples in each relationship allows us to identify the features that might be useful for the purpose of automatic classification.

In the following text, we will use D1 and D2 to represent description 1 and description 2. We will use X and Y to represent the information set contained in description 1 and description 2. The information items in X are the union of $x_1 \ldots x_n$. The information items in Y are the union of $y_1 \ldots y_n$.

Relationship: Identity

Constraints: D1 and D2 contain the same string, any level, non-directional, and does not co-exist with other relationships.

Effects: X = Y

Examples can be found in Table 5.16.

The relationship can exist inter-level or intra-level and is not directional where R(X,Y) = Identity indicates R(Y, X) = Identity. If the relationship between two descriptions is

identity, this means that two descriptions contain the exact same text.

Table 5.16 Instances of Identity relationship

Element	Description 1	Description 2
arrangement	Alternate	Alternate
margin	entire	entire

For the purpose of identifying this relationship, lexical clues can be used. If two descriptions contain the same string, this relationship can be identified.

Relationship: Equivalence

Constraints: D1 and D2 contain the different strings, any level, non-directional, and does not co-exist with other relationships.

Effects: X = Y

Examples can be found in Table 5.17.

Table 5.17 Instances of Equivalence relationship

Element	Description 1	Description 2
arrangement	alternate, spirally arranged or distichous	alternate, opposite, or whorled
margin	coarsely and sharply doubly serrate,	coarsely or incised doubly serrate
surface	abaxially with veins glabrous or with soft pubescence, without straight, erect hairs, glands yellow, adaxially margins of younger leaf blades with few or no cystolithic hairs	abaxially glabrous or with scattered soft pubescence but without rigid spinulose hairs on veins, adaxially with few or no cystolith hairs marginally when young

In the first example, we see that “alternate” exists in both descriptions. “Spirally

arranged" has the same meaning as "whorled" and "distichous" has the same meaning as "opposite". Therefore, the two descriptions thus contain the same information.

In the second example, "sharply" and "incised" have the same meaning. The word "and" and "or" are stop words.

In the third example, two descriptions present the same meaning that "abaxially glabrous or with soft pubescence without rigid on veins, adaxially margins of younger leaf blades with few or no cystolithic hairs."

In these examples, we found that there is a significant proportion of identical words that appear in both descriptions. We also found that, except for those identical words, the other words in the descriptions are synonyms. The other words that do not have synonyms are usually modifiers. Same as Identity relationship, two descriptions that have the relationship Equivalence contain the same meaning, so we group the two relationships together into one class.

Relationship: Subsumption

Constraints: S1 and S2 contain different strings, inter-level, directional, does not co-exist with other relationships.

Effects: $X \supset Y$

Examples can be found in Table 5.18.

Table 5.18 Instances of Subsumption relationship

Element	Description 1	Description 2
margin	entire, dentate, or crenate.	dentate or entire;
lobe	sometimes 3-lobed or more,	sometimes lobed,
margin	sinuous dentate to serrate or entire	serrate
apex	acute or obtuse to emarginate, mucronulate	obtuse, with a mucro

These examples show that in this relationship, the description that subsumes another description is usually longer. One special case is that if all of the words appear in the second description that also appears in the first description, we assume that it represents a subsumption relationship.

The proposed features can include: number of tokens in description 1, number of tokens in description 2, percentage of common words in description 1, percentage of common words in description 2, differences in length of descriptions, and whether description 1 is longer than description 2.

Relationship: Complementary

Constraints: D1 and D2 contain different string, any level, non-directional, does not co-exist with other relationships.

Effects: $X \cap Y = \varnothing$

Examples could be found in Table 5.19.

We see that in this relationship, the descriptions contain quite different information. There are few words that are identical in both descriptions.

Table 5.19 Instances of Complementary relationship

Element	Description 1	Description 2
stipule	2 per node, inconspicuous, white, ovate to deltate, margins ciliate, apex acute;	small, membranous, caducous.
stipule	4 per node, white, ovate to triangular, margins entire but splitting variously with age, apex obtuse to acuminate;	free, scarious.
vein	forking to 3 times per side.	9-16 on each side of midvein.
petiole	1-6 cm, often as long as leaf blade.	puberulous;

The features to be considered include: number of tokens in description 1, number of tokens in description 2, percentage of common words in description 1, percentage of common words in description 2.

Relationship: Overlap

Constraints: D1 and D2 contain different strings, any level, non-directional, can co-exist with Contradiction relationship

Effect: $X \cap Y \neq \varnothing$ and $X \neq Y$ and $X \not\subset Y$ and $Y \not\subset X$

Examples can be found in Table 5.20.

We see that in the overlap relationship, the descriptions contain some words that are identical in both descriptions. However, they also contain some information that is different from the other. Therefore, there does exist some words or phrases that are common to both descriptions.

The features to consider include: number of tokens in description 1, number of tokens in

description 2, percentage of common words in description 1, percentage of common words in description 2.

Table 5.20 Instances of Overlap relationship

Element	Description 1	Description 2
apex	acute to slightly acuminate;	acute or obtuse;
vein	with 5-18 pairs of lateral veins,	lateral veins 6-8 on each side of midvein.
surface	abaxially glabrous to sparsely pubescent, covered with minute, resinous glands.	abaxially densely resinous punctate, adaxially glabrous
surface	abaxially pale, glabrate, adaxially dark green, lustrous, glabrous;	abaxially pea green and pubescent when young, adaxially dark green, lustrous, and pubescent only on midvein,
vein	forking 5 or more times per side.	depressed; 10-15 on each side of midvein.
base	shallowly cordate, often oblique, truncate, or broadly rounded	cordate and asymmetric,

To sum up the findings from the examples, we identified the following features used in our system: the number of tokens in description 1, the percentage of common words in description 1, the percentage of common words in description 2, the differences in length of descriptions, and whether description 1 is longer than description 2. Among these features, the most important feature is the percentage of common words that appear in both descriptions.

We must first be clear regarding the word “synonymous.” “The English language rarely has absolutely true synonyms-that is, situations where two words mean the same thing and have no variations in nuance” (Taylor, 1999). However, there are words and phrases that have meanings that are so close to each other that they are interchangeable in certain

contexts. Therefore, when producing the descriptive materials, the author has a great deal of freedom when choosing words from the variations that are available. Therefore, we must "teach" the machine to understand in the same sense and process the data properly. Controlled vocabularies (e.g., dictionaries, ontologies) are the tools that make automatic processing possible. As can be expected, the quality of the dictionary will impact the performance of the system that is discussed in detail in Chapter 7.

WordNet is not the best dictionary for this project given the context. It is a large lexical database that is not designed for any specific domain. It would be better to use a biological dictionary or, even better, a botanical dictionary. However, WordNet currently has two advantages that are not found in other dictionaries:

- It is machine-readable.
- It has a semantic network. "WordNet is organized by the concept of synonym sets (synsets), groups of words that are roughly synonymous in a given context".[21]

Therefore, WordNet 3.0[22] is used to identify the synonyms and antonyms in our fusion system.

5.2.2 Model Selection

Different learning models are available including supervised and unsupervised models.

---

[21] WordNet Frequently Asked Questions http://wordnet.princeton.edu/wordnet/faq/ Retrieved on April 16, 2011

[22] WordNet 3.0 Accessible at http://wordnet.princeton.edu/wordnet/download/

Supervised models such as Decision Tree (C4.5), Support Vector Machine (SVM), Conditional Random Field (CRF), Hidden Markov Model (HMM) and Naïve Bayes (NB) are still the learning models that attract the most attention and usually yield positive results, e.g., Wang et al., 2005; Bellegarda, 2004; Peng et al., 2004; Langseth and Nielsen, 2006.

When selecting classification methods, several criteria should be taken into consideration: predictive accuracy (the ability of the method to classify new data into proper classes), speed (the computation costs involved in generating and using the model), robustness (the ability of the model to make correct predictions given noisy data or data with missing values), scalability (the ability to construct the model efficiently given large amounts of data) and interpretability (the level of understanding and insight provided by the model) (Han & Kamber, 2001). Research models used in recent related research studies are presented in Table 5.21.

Table 5.21 Models used in related research

Projects	Model Used
(Zhang, et al., 2003)	Boosting
(Maziero, et al., 2010)	J48 Decision Tree
(Hatzivassiloglou, et al.,1999)	RIPPER (rule learning, similar to decision tree)
(Miyabe, et al., 2008)	Clustering

Among all the learning models, the decision tree model is one of the most widely used and practical methods for classification. It was used in Maziero, et al (2010) and Hatzivassiloglou, et al. (1999). It is a method for approximating discrete-valued functions

that is robust when working with noisy data and is capable of learning disjunctive expressions (Mitchell, 1997). Therefore, the J48 Decision Tree model was used in our system. Finding the best performance machine-learning algorithm for classification purposes is not the focus of this research. The focus of this research is fusion, as will be discussed in the following section. However, it is generally accepted that sophisticated algorithms such as SVM, CRF or HMM can obtain better performance by using the same feature sets.

5.2.3 Performance

We implemented the classification by using the features identified in section 5.2.1 and J48 Decision Tree model in java. The J48 decision tree model package is from WEKA. We separated the classification into two sets of experiments: inter-level and intra-level classifications. We did this because we have different CDR relationships (classes) for classification. The data were collected based on LEAVE ONE OUT cross validation. We did this because we are going to use LEAVE ONE OUT in the real application stage. As the name suggests, LEAVE ONE OUT involves using a single instance from the original dataset as validation data, and the remaining observations are used as training data. This is repeated so that each instance in the sample is used once as the validation data. This is the same as an n-fold cross-validation where n is equal to the number of instances in the sample.

Intra-level Classification Performance includes both (genus, genus) and (species, species) description pairs. We have a total of 501 description pairs. We correctly classified 363

(72.5%) of the pairs using the classifier. Table 5.22 presents the detailed accuracy information by class. Table 5.23 shows the confusion matrix.

Table 5.22 Accuracy by class

(TP rate: True Positive rate, FP Rate: False Positive rate)

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.742	0.05	0.783	0.742	0.762	Subsumed
0.48	0.049	0.522	0.48	0.5	Overlap
0.744	0.11	0.701	0.744	0.722	subsumption
0.765	0.038	0.878	0.765	0.818	Identity
0.753	0.1	0.631	0.753	0.686	complementary

Table 5.23 Confusion matrix

a	b	c	d	e	<-- classified as Class
72	5	5	6	9	a = subsumed
5	24	10	4	7	b = overlap
5	7	96	4	17	c = subsumption
5	7	11	101	8	d = identity
5	3	15	0	70	e = complementary

Table 5.22 shows that the relationship overlap has the lowest performance. Table 5.23 shows that the overlap was primarily incorrectly classified as subsumption and complementary.

Inter-level Classification Performance includes both (FNA genus, FNA species) and (FOC genus, FOC species) description pairs. We have a total of 426 description pairs, 361 description pairs (84.7%) are classified correctly. Detailed information is presented in Tables 5.24 and 5.25.

Table 5.24 Accuracy by class

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.905	0.341	0.911	0.905	0.908	refinement
0.25	0	1	0.25	0.4	overlap
0.704	0.015	0.76	0.704	0.731	identity
0.642	0.078	0.54	0.642	0.586	complementary

Table 5.25 Confusion matrix

a	b	c	d	<-- classified as class
306	0	6	26	| a = refinement
5	2	0	1	| b = overlap
6	0	19	2	| c = identity
19	0	0	34	| d = complementary

We observe that refinement is the major relationship among the inter-level description pairs (77% = 331 out of 427). This means that for most of the cases, the descriptions on the species level are more detailed and it is true that the species level descriptions contain more specific information about the particular species than the genus level description. The confusion matrix shows that the major confusion occurs between refinement and complementary. Similar performance occurs in inter-level classification and in intra-level classification where overlap exhibits the lowest performance.

We can see from both inter-level and intra-level performance that given we only used 5 features for the classification, the results are promising, particularly for intra-level cases. The performances can be enhanced by using a better dictionary (for example, a leaf or plant ontology dictionary is likely to improve performance) and a better feature set. However, we will leave such improvements for future researchers.

5.3 Conflict Detection

There are different degrees of conflict within our data. We can argue that any pair of two descriptions when describing the same property of the same species that are not identical or equivalent representing some degree of conflict with each other. Therefore, CDR relationships (subsumption, complementary, overlap, contradiction) are different forms of conflict For example, in the cases of the subsumption relationship in the same level, conflict occurs when one of the descriptions contains more information than the other. Example one has two descriptions describing the apex of the same species.

Example one:

<apex>acute<apex>

<apex>acute or obtuse<apex>

In this case, we have partial agreement and partial disagreement. Both descriptions agree that the species has an acute apex. However, the second description also indicates that some instances of the species can have an obtuse apex. Here we detect a conflict between the two descriptions because acute and obtuse have quite different meanings. Let us examine some other examples:

Example two:

<texture>papery to leathery</texture>

<texture>leathery</texture>

Example three:

<color> green, red, scarlet, maroon, purple, and yellow</color>

<color> green, red, purple or yellow </color>

In example two, papery is different than leathery, but we are unable to define how different they are. In example three, “scarlet, maroon” is not mentioned in the second description, and we cannot judge how different are they from “red” and “purple.” Can we say that we detect a conflict in these two cases? Probably not in both cases.

This is fundamentally a problem that concerns modifiers, particularly frequency and certainty modifier. The information in the second “apex” description of example one means that there exists some possibility that the species can have an acute apex, and some possibilities that the species can have an obtuse apex. Information fusion with possibilities is discussed in chapter 8 in detail.

In our context, we are interested in direct conflict. Here direct conflict is defined as: one description says character c is “x” but the other description says that character c is “not x”. Then we say that there exists a direct conflict between the descriptions. In looking through our entire dataset, we were only able to identify 10 instances of direct conflict. Therefore, we did not include conflict in the automatic classification for the reason that we do not have a sufficient number of training examples for machine learning. In this case, the rule-based method is used. Table 5.26 shows some examples of conflict.

Table 5.26 Instances of Conflict relationship

Element	Description 1	Description 2
base	generally not oblique,	obliquely to symmetrically obtuse to rounded,
lobe	unlobed,	palmately lobed,
base	not swollen	Expanded
color	not glaucous	Glaucous
stipule	inconspicuous, silvery to dull tan, triangular, 1.5-2 mm, apex acuminate	not silvery, triangular, lanceolate, short

In these examples, we see that the conflict description pairs fall into one of the following categories:

1. Existence of odd number of "not" and the following word/phrase can find a synonym in another description.

2. Existence of odd number antonym between the two descriptions.

In conflict detection, recall that it is more important than precision. Recall means the ability to pull ***all*** conflict cases out of the data, whereas precision means the ability to pull ***only*** conflict cases out of the data. Therefore, we need the rules that can pull out all of the conflict cases.

According to these two rules, we can identify 62 description pairs. Among the description pairs, 5 of the description pairs are true conflicts. The other 5 cases were not identified by the system, primarily because the system could not correctly identify some of the synonyms or antonyms mainly due to the dictionary we are using. Some instances are shown in Table 5.27.

Table 5.27 Instances of failures in conflict detection

Element	Description 1	Description 2
lobe	palmately lobed, sometimes unlobed;	3-7(-9)-lobed,
lobe	unlobed,	palmately lobed,
base	not swollen	expanded

The system does not consider "unlobed" and "lobed" to be antonyms. In addition, "swollen" and "expanded" are not synonyms in WordNet.

For the identified description pairs that are not true conflict cases, most of these identified description pairs look like this:

Table 5.28 Instances of false positives in conflict detection

Element	Description 1	Description 2
complexity	simple, palmate ternate or pinnate	simple or compound
apex	acute to short acuminate	acute, acuminate to obtuse

In the first case, WordNet classifies "simple" and "compound" as being antonyms. In the second case, WordNet classifies "acute" and "obtuse" as being antonyms.

5.4 Summary

This chapter defined the CDR relationship within our dataset. We also presented details related to the automation of CDR relationship identification. The output of this step will be the input used in the next step: fusion.

Chapter 6: Information Fusion

The goal of information fusion is to achieve both better data quality and broader view of the data simultaneously. The final stage, information fusion can be done by applying fusion aggregators to the result of CST-relationship identification. We emphasize that all of the functions and procedures that are used are task-specific and must be changed according to the task, application and domain. The core problem in information fusion is to define the fusion method and how it will be executed in the target information set. We will discuss the fusion methods in this chapter.

6.1 Data Types and Modeling

Before discussing specific fusion methods, we must examine what kinds of data we have. Three different types of variables are the targets of this study: numeric, numeric-like, and categorical. They are summarized in Table 6.1.

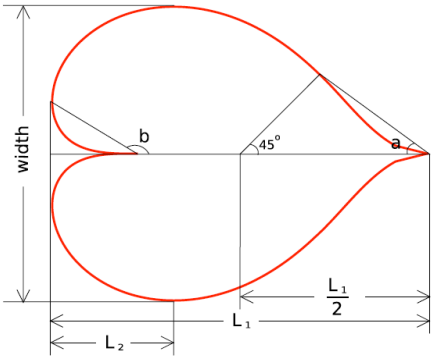
Table 6.1 Data types

Variable	Change	Distance measure	Fields	Examples
Numeric	Continuous	Yes	Bladelength, Bladewidth	[bladelength]1-2 cm
Numeric-like	Discrete	Yes	Bladeshape	[bladeshape] lanceolate, linear, or filiform
Categorical	Discrete	No	All other elements	[arrangement] alternate

For variables such as shape, although they are not continuous, they do have distance

measures. For example, ovate is closer to round than to being linear. They can also be modeled to become numeric-like variables. With domain knowledge, we are able to transform blade shape into a numeric-like variable and there are different approaches that can be taken.

In order to transform the leaf shape to be numeric-like, Wang & Pan (2006) used a four-feature vector to represent leaf shape. Figure 6.1 shows the features.



- length-width ratio: $f_1 = \frac{L_1}{width}$;
- the position of the widest part: $f_2 = \frac{L_2}{L_1}$;
- the apex angle: $f_3 = a$;
- the base angle: $f_4 = b$.

Figure 6.1 Features for blade shape (Wang & Pan, 2006)

For each feature, the value is not a number, but rather a range. As regards a simple shape, e.g., *elliptic*, each value in the four-feature vector will be a region with a small range, *i.e.*, (*rf*1, *rf*2, *rf*3, *rf*4), where *rfi* = [*fi**0.9, *fi**1.1], for *i* = 1, . . . , 4. In the case of a more complex shape such as "*narrowly elliptic,*" the vector will be based on the base shape "*elliptic:*"

"narrowly:" $f1' = f1*1.2$ and $fi' = fi*0.9$, for $i = 3, 4$

"broadly:" $f1' = f1*0.8$ and $fi' = fi*1.1$, for $i = 3, 4$

For hyphenated shapes, e.g., *oblong-ovate,* the vector will be based on both base shapes as shown in Figure 6.2.

$$hf_i = \frac{f_{X\,i} + f_{Y\,i}}{2}, \text{ for } \quad i = 1, \ldots, 4$$

Figure 6.2 Value for hyphenated shapes (Wang & Pan, 2006)

The distance of two shapes will then be computed based on ranges. This method offers the advantage of transforming the shapes into numbers, and results in a continuous variable. The problem is that the result we obtain from using this method will be a range of numbers that is not human-readable text. Figure 6.3 shows the final output of the transformation.

After examining the data obtained and consulting with some domain experts, we used an alternate method to transform the shape description. The transformation is based on Figure 6.4. The shapes are separated into three groups: the first group, the second group and the third group.

1) The first group (frequently appearing shapes): from elliptic, oblong, lanceolate to ovate;

2) The second group: from oblanceolate to obovate;

3) The third group (rarely appearing shapes): triangle, rhombic, trullate, obtrullate, obtriangle.

Species	Leaf Shape Descriptions	Integration Results			
		R_{f_1}	R_{f_2}	R_{f_3}	R_{f_4}
Salix pentandra (Laurel willow)	ovate or ovate-elliptical to elliptical- or obovate-lanceolate broadly lanceolate to ovate-oblong broadly elliptical broadly lanceolate, ovate-oblong, or elliptic-lanceolate	1.21–2.87	0.27–0.57	0.10–0.35	0.27–0.37
Glinus lotoides	obovate or orbiculate to broadly spatulate obovate to oblong-spatulate orbiculate or more or less cuneate	0.90–2.33	0.46–0.80	0.34–0.47	0.04–0.44
Spinacia oleracea	hastate to ovate ovate to triangular-hastate oblong	1.22–1.63 1.81–2.21	0.08–0.39 0.45–0.55	0.17–0.25 0.27–0.33	0.37–0.63 0.27–0.33
Alternanthera paronychioides	oblanceolate or spatulate elliptic, ovate-rhombic, or oval elliptic, oval or obovate	2.83–3.46 2.39–2.92 1.45–2.57	0.62–0.76 0.72–0.88 0.40–0.69	0.28–0.34 0.34–0.42 0.17–0.38	0.09–0.11 0.03–0.04 0.22–0.32

Figure 6.3 Examples of transformation results (Wang & Pan, 2006)

Below is the procedure for transforming the shapes in the descriptions into numeric-like variables and calculating the distance between two shapes. They have categorical distances but not numerical distances and this feature is true only if the two shapes are in the same group. It does not make sense to traverse across the groups since they have very different shapes. Therefore, for any two shapes in the table, they should be one of the following cases:

(a) They belong to different groups: in this case, the distance between them is the distance between the two shapes. In other words, we do nothing in this case.

(a) Both fall into a single group: in this case, we map the two shapes into the table. We

then include all of the shapes that are between them horizontally, vertically and diagonally.

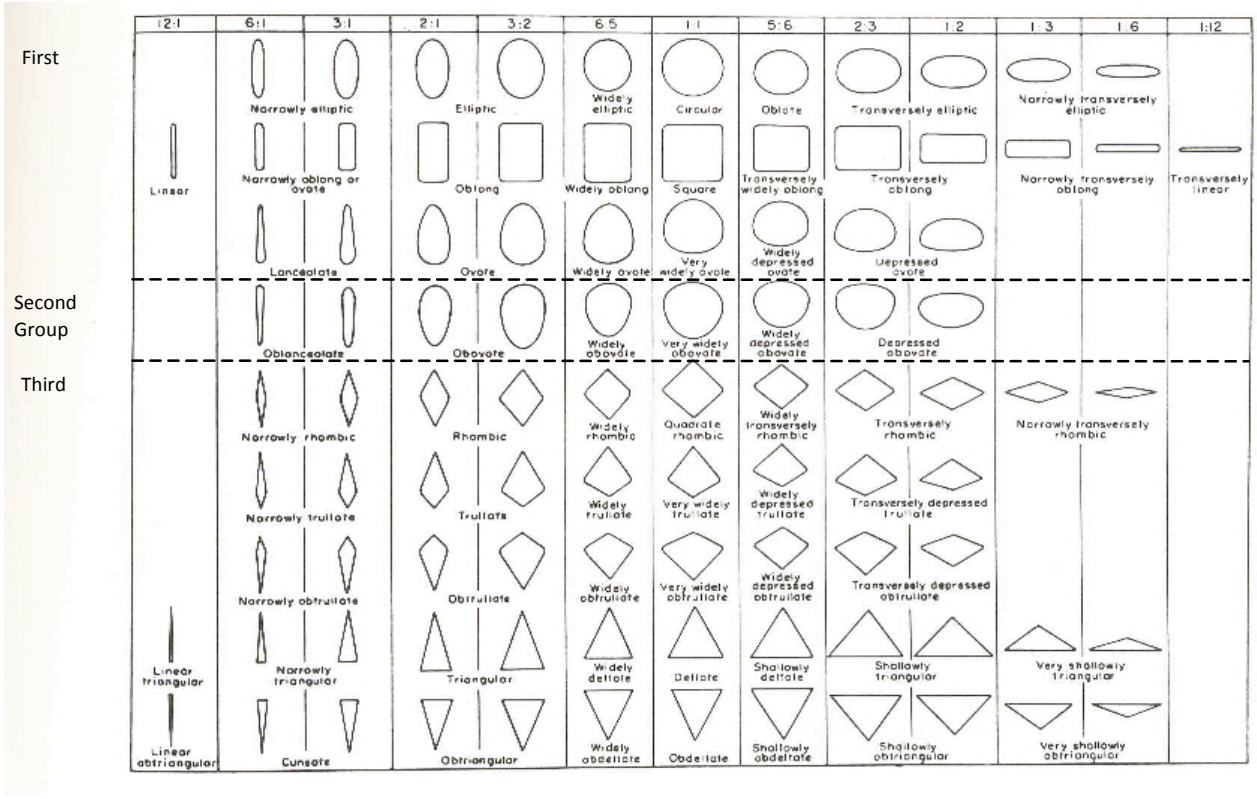


Figure 6.4 Standardized shapes for descriptions of outlines of symmetric leaf blades (Stuessy, 1990)

Here are some more examples showing the input and output of the transformation.

input: <bladeshape> linear to ovate<bladeshape>

output: <bladeshape> linear, ovate, lanceolate<bladeshape>

input: <bladeshape>narrowly to broadly elliptic or oblanceolate<bladeshape>

output: <bladeshape> elliptic, oblanceolate <bladeshape>

input: <bladeshape>elliptic, oblong-elliptic, or obovate-oblong <bladeshape>

output: <bladeshape>elliptic, oblong, obovate <bladeshape>

6.2 Fusion Methods

The fusion method is the combination of a set of fusion operators (which can also be called fusion aggregators, fusion functions, or aggregation operators) which is the operator that corresponds with particular mathematical or non-mathematical functions that are used for information fusion. Selecting a specific fusion operator should depend on the nature of the information to be fused. It is generally a function that is used to combine N values in a given domain D (e.g., N integers) and return a value (e.g. another integer) on the same level, or higher level in the same domain (Torra & Narukawa, 2007). The function *F* can be denoted as:

$$F(i\ldots\ldots n) = m \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(1)$$

Torra & Narukawa (2007) argued that the operations fuse input values by taking into account certain information about the sources that is usually called domain knowledge or background knowledge. The knowledge can be one of the following (Kokar, Tomasik and Weyman, 2004):

a. Knowledge about the source/public;

b. Knowledge of the goal of the fusion system;

c. Content domain knowledge.

The operators are parametric so that the domain knowledge obtained from the sources can be considered during the fusion process. We can express the knowledge *k* in function (1) as:

$$F_k(i \ldots\ldots n) = m \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2)$$

Those operators are traditionally used in database systems to aggregate data into another variable. Some common operators are shown in Table 6.2.

Table 6.2 Available operators

Data Type	Popular Operators
Numerical data	sum, minimum, maximum, arithmetic mean, weighted mean (when information from one source is more reliable than another source), fuzzy measure (when all of the sources are independent), etc.
Categorical data	the median and the majority rule

As discussed above, the selection of information fusion methods is highly dependent upon the data type, the task and domain of the application. There are always assumptions and meanings that are associated with a specific operator. The differences of the assumptions are highly reliant upon the sources and domains. When we make decisions about which operators to use, several factors must be considered. Torra & Narukawa (2007) summarized the factors including: information type, type of data representation, level of abstraction, and construction of fusion method. Table 6.3 shows the details.

Table 6.3 Factors to consider when constructing fusion methods

Factors	What	How
Information Type	The original information can be classified as different types by different criteria. Possible categories include: scholarly, professional, government, facts, entertainment, etc.	Two types of information are of importance in IF: (1) Redundant information. One objective of information fusion is to remove redundant information when we have multiple sources. The question then becomes how to recognize the redundant information when the data is not fully identical? (2) Complementary information. It is beneficial to have complementary information where fusion is applied. The question then becomes how to separate complementary information from conflicting information?
Data Representation	The data needs a representation that the methods can be applied to.	The most popular representations include: numerical data, ordinal scales, fuzzy sets, belief functions, dendrograms, etc.
Level of Abstraction	Fusion techniques can be applied at different levels of abstraction.	IF is a process where higher-level information is obtained by aggregating lower-level information. The selection of the appropriate level depends upon the information available.
Construction	This is a process in which aggregation operators are defined.	They should minimize a given expression following the same data and output the same result while the same input is fed. The starting point for defining the fusion method is a set of properties considered to be requirements for the method. Given these properties, the function is derived using mathematical tools.

As mentioned above, the definition of the fusion method greatly depends on the data and the domain to which the methods will be applied. As proposed in Chapter 3, we will define our fusion method based on cross-description relationships and the variables the operator intends to execute.

6.3 Assumptions

The assumptions provide us with knowledge about the sources and the goal of the fusion system. Without these assumptions, we will not be able to define the fusion rules in Section 6.4. The assumptions are:

- We assume that the two sources we use (FNA and FOC) exhibit the same degree of reliability. This means that the information from FNA and FOC have the same importance. We do not give preference to any source during processing.
- We assume that the information included on the higher (genus) level is true for all lower (species) levels. For example, if the genus level says that the arrangement is "alternate," all of the species' leaf arrangements should also be "alternate." The FNA handbook states "Character states common to all taxa are treated in the description of the taxon at the next higher rank. For example, if corolla color is yellow for all species treated within a genus, that character state is given in the generic description." [23]
- We assume that there is no conflicting information within a single description.

6.4 Fusion Scenarios

Before we proceed to the fusion method, some background information is needed in order to establish a clear idea what the fusing tasks are given the data we have. There are

[23] Same as 6.

different scenarios where fusion can occur.

Merge known with unknown. There are many cases (about 50%, as discussed in Chapter 5) where the statement about a specific character is available in only one source. In such cases, we merge the statement with an empty statement. The background assumption regarding the descriptions is open world that means all of the things that are not mentioned are unknown instead of "not." For example, if the description does not mention the leaf arrangement of a species it simply means that the author does not have knowledge of the leaf arrangement. It does not mean that the leaf arrangement does not exist or that the leaf arrangement is not alternate, for example. Thus, when we merge the character statements where only one statement is available, we simply regard the available statement as being the fusion result. In such cases, we do not need to evaluate the accuracy because we simply use what we have as the result.

Intra-level information fusion VS Inter-level information fusion. The system enables fusion between descriptions on the same level (either genus level or species level). The output of the fused result should be the union of the two treatments. In this sense, intra-level fusion is a process that adds variance to the existing characters. In other words, we want the fused result to contain information that is as broad as possible.

At the same time, the system also provides for inter-level information fusion. Inter-level fusion is different than intra-level fusion in the sense that we want more specific information instead of adding variance to the description. We want to reduce the variance

by adding information from another level and we want the information to be as exclusive as possible. Let us examine the following examples in Table 6.4. These examples show that we obtained different correct results in different fusion scenarios while using the same input. While constructing fusion methods, different fusion methods are needed for different fusion scenarios. The differences in fusion scenarios are captured in the condition functions (discussed in section 6.5).

Table 6.4 Different correct fusion results in different scenarios

Element	Description 1	Description 2	Intra-level Fused Result	Inter-Level Fused Result
Texture	papery to leathery	leathery	papery to leathery	leathery
Blade Length	7-12 cm	6-10 cm	6-12 cm	6-10 cm

6.5 Fusion Functions

When merging a set of descriptive treatments, we start with background knowledge, their CDR relationship, and the information in the treatments that will be merged and then apply fusion rules to this information. The application of the fusion rules then becomes a monotonic process that builds up the final results by applying the functions. The functions can be separated into three groups: feature functions, condition functions and action functions.

Feature Functions. Feature functions are those functions that capture the structures and features information of the input text. They are usually used to encode the information

regarding a single description. Examples include:

Element(“apex”, s) means that the string s is describing the apex of leaf.
Overlap(s1, s2) means that the CDR relationship between string s1 and s2 is overlap.
Source(“FNA genus”, s) means that description s comes from FNA genus level.

The information in feature functions is captured prior to the fusion process. Meanwhile, feature functions can also be considered to be one special category of condition functions.

Condition Functions. The condition functions relate the information in the descriptions to the domain knowledge. There are many possible functions that can be defined depending on how much domain knowledge we want to encode into the system. Given that this is an exploratory project, we currently have the following functions:

IsSynonym(s1, s2): determines whether string s1 and string s2 are synonyms. String s1 and s2 can either be phrases or words.
IsStop(s): determines whether string s is a stop word that is on the stop word list.
IsAntonym(s1, s2): determines whether string s1 and string s2 are antonyms. String s1 and s2 can either be phrases or words.
IsGreater(n1, n2): determines whether number n1 is greater than number n2.
IsEqual(n1, n2): determines whether number n1 is equal with number n2.
IsSameGroup(s1, s2): determines whether shape n1 and shape n2 are in the same group.
IsSameLevel(Source(s1),Source(s2)): determines whether description s1 and s2 are on

the same level (either species or genus)

More functions can be defined if we want to encode more detailed domain knowledge:
IsCoherent (“acute”, “mucronulate”) & Element(“apex”, “acute”) & Element(“apex”, “mucronulate”)
IsConflict (“green”, “red”) & Element(“color”, “green”) & Element(“color”, “red”)

As mentioned above, the construction of condition functions is heavily dependent upon the quality and quantity of the domain knowledge we have and how much we want to be encoded into the system. The quality of the condition functions directly impacts the performance of the fusion process. The accuracy of the functions is largely correlated with the dictionary (ontology) we are using. The better the dictionary we use, the better results we are likely to get.

Action Functions. Action functions specify the actions to take when one or more condition functions are met. It tells the computer how the fusion happens. We can also say that those action functions are aggregators that can be applied to the real dataset in order to get the fusion output.

We are processing numerical variables, numeric-like variables and categorical variables. Here are some illustrative action functions in our system. Action functions define the actions that are to be applied in the data.

Initialize(S): the function is to start the fusion process with the basic string which can be a description or a word.

Add(S, s): this function adds the string s to string S.

Remove(S, s): this function removes string s from S.

Replace(S, s1,s2): this function replaces string s1 in S with string s2.

Min(n1, n2): this function outputs the smaller number between number 1 and number 2.

Max(n1, n2): this function outputs the larger number between number 1 and number 2.

ShapesInBetween(s1, s2): this function outputs the shapes between shape 1 and shape 2.

6.6 Executing Fusion Functions

We defined and implemented a set of functions in our system and the next step is to execute the fusion rules in order to get our fusion result. The execution is achieved by a multi-step merging process. We will examine the following examples in order to see how the fusion is accomplished by using the functions that were defined in Section 6.4.

Before we progress to the results of fusion, the representation of the fusion result needs to be determined. There is always a tradeoff relationship between the efficiency of processing and the expressiveness of the representation. For this particular kind of text mining, we need to choose a representation that allows us future application such as information retrieval since fusion is the first step toward more efficient access.

The simplest representation is in the form of pairs, e.g. (attribute, value). Our intention here is to extend this simple representation strictly as needed. Although we are focusing on leaves right now, our future works will include flowers and stems, etc. In this case we would have a triple (part, attribute, value).

Although some of the information is not well represented in the representation of the result, the information is preserved in the knowledge base, e.g., the CDR relationship identified between the description pairs.

Example One: (Numeric Variables)

Consider the following descriptions on blade length.

<bladelength> 30 mm to 50 mm</bladelength>

<bladelength> to 40 mm</apex>

The intra-level fusion process will be:

Fuse (“30 mm to 50 mm”, “to 40 mm”) & Element (“bladelength”) & IsSameLevel (“FNA species”, “FOC species”)

-> Min(“30 mm”, “”) & Max(“40 mm”, “50 mm”)

Therefore, the final output is

(Leaf, Blade length, “30 mm to 50 mm”).

The inter-level fusion process will be:

Fuse (“30 mm to 50 mm”, “to 40 mm”) & Element (“bladelength”) & IsSameLevel

(“FOC genus”, “FOC species”)

-> Min(“30 mm”, “”) & IsGreater(“40 mm”, “50 mm”)

Therefore, the final output is

(Leaf, Blade length, “30 mm to 40 mm”). [Note, if 40 mm is larger than 50 mm, the system will report that a conflict is detected.]

Example Two: (Numeric-like Variables)

Suppose we have the following two descriptions on blade shape.

<bladeshape>elliptic, lanceolate, oblong, ovate<bladeshape>

<bladeshape>elliptic, oblong, obovate<bladeshape>

The intra-level fusion process will be:

Fuse (“elliptic, lanceolate, oblong, ovate”, “elliptic, oblong, obovate”) Element (“bladeshape”) & IsSameLevel (“FNA genus”, “FOC genus”)

-> IsSameGroup(“elliptic”, “oblong”) & IsSameGroup(“elliptic”, “obovate”) & IsSameGroup(“lanceolate”, “elliptic”)……[combine any two arbitrary shapes, one from each description]

-> ShapesInBetween(s1, s2) [if two shapes in the same group]

->remove duplicate shapes

The final output is:

(leaf, bladeshape, “elliptic, lanceolate, oblong, ovate, obovate”).

The inter-level fusion process will be:

Fuse (“elliptic, lanceolate, oblong, ovate”, “elliptic, oblong, obovate”) Element (“bladeshape”) & IsSameLevel (“FNA genus”, “FNA species”)

->Output shapes in species level

-> IsSynonym(“elliptic”, “elliptic”) & IsSynonym (“oblong”, “elliptic”)…… [detecting whether there are shapes in species level but not in genus level.]

The final output is:

(leaf, bladeshape, “elliptic, oblong, obovate”).

Example Three: (Categorical variables)

Consider the following overlapping descriptions of leaf apex. Note that the same descriptions might be assigned different CDR relationships while in different fusion scenarios.

<apex>obtuse, rounded, mucronulate .</apex>

<apex>obtuse or acute.</apex>

The Intra-level fusion process will be:

Fuse (“obtuse, rounded, mucronulate”, “obtuse or acute.”) & Element (“apex”) & Relationship (“overlap”) & IsSameLevel (“FNA species”, “FOC species”)

->Initialize (“obtuse, rounded, mucronulate”)

->IsSynonym (“obtuse”, “obtuse”) & IsSynonym(“obtuse”, “acute”)……[combine any two arbitrary terms, one from each descriptions to determine where the new information comes from]

->Add (“obtuse, rounded, mucronulate”, “acute”)

Therefore, the fused result is:

(Leaf, apex, obtuse, “rounded, mucronulate, acute”)

The Inter-level fusion process will be:

Fuse (“obtuse, rounded, mucronulate”, “obtuse or acute.”) & Element (“apex”) &

Relationship (“refinement”) & IsSameLevel (“FNA genus”, “FNA species”)

->Output species level description as final result

Therefore, the fused result is:

(Leaf, apex, obtuse, “obtuse or acute”)

Example Four: (Categorical variables)

Consider the following two descriptions about stipule.

<stipule>entire to serrate </stipule>

<stipule>2-3 cm, bugle shaped, deciduous</stipule>

In the case of two descriptions that have a complementary relationship, we add the descriptions together to produce the final output. The actions are the same for intra-level and inter-level fusion.

Fuse (“entire to serrate”, “2-3 cm, bugle shaped, deciduous”) & Element (“stipule”) &

Relationship(“complementary”) & IsSameLevel(“FNA species”, “FOC species”)

->Initialize (“entire to serrate”)

->Add (“entire to serrate”, “2-3 cm, bugle shaped, deciduous”)

Therefore, the final output will be:

(Leaf, stipule, “entire to serrate, 2-3 cm, bugle shaped, deciduous”).

6.7 Summary

This chapter describes the fusion methods we used and how these fusion methods are constructed. We will evaluate the performances of these fusion methods in the next chapter.

Chapter 7: Evaluation

The classical approach to evaluating an information retrieval system in information science is to conduct user-based evaluations. In order to evaluate our fusion system, an experiment was conducted by using the data from FNA and FOC. One of the goals of this dissertation study is to demonstrate that an automatic/semi-automatic information fusion system is feasible using currently available techniques. We are exploring the performance of this prototype system using human subjects. We are interested in both the performance itself as well as in ideas for future improvements that can be obtained by examining the failures of the current system. The results are reported and analyzed.

7.1 Evaluation Design and Subjects

The experiment involved randomly sampling (without replacement) 60 out of the 153 species in FNA and FOC. The 240 treatments (both FNA and FOC on the genus and species levels) were then processed through the fusion system.

7.1.1 Subjects

The subjects for this experiment were recruited in the second "Fine-Grained Semantic Markup of Descriptive Data for Knowledge Applications in Biodiversity Domains" conference. The conference was held on March 24-25, 2011 at University of Arizona, Tucson, AZ.

A total of 13 subjects were recruited, and only one botanist was among them. There were varying levels of botanical knowledge among the members of this group, but all of them had some understanding of biodiversity informatics. This group of subjects represents a broad range of potential users of flora data. Figure 7.1 presents the distribution of the subjects' profession.

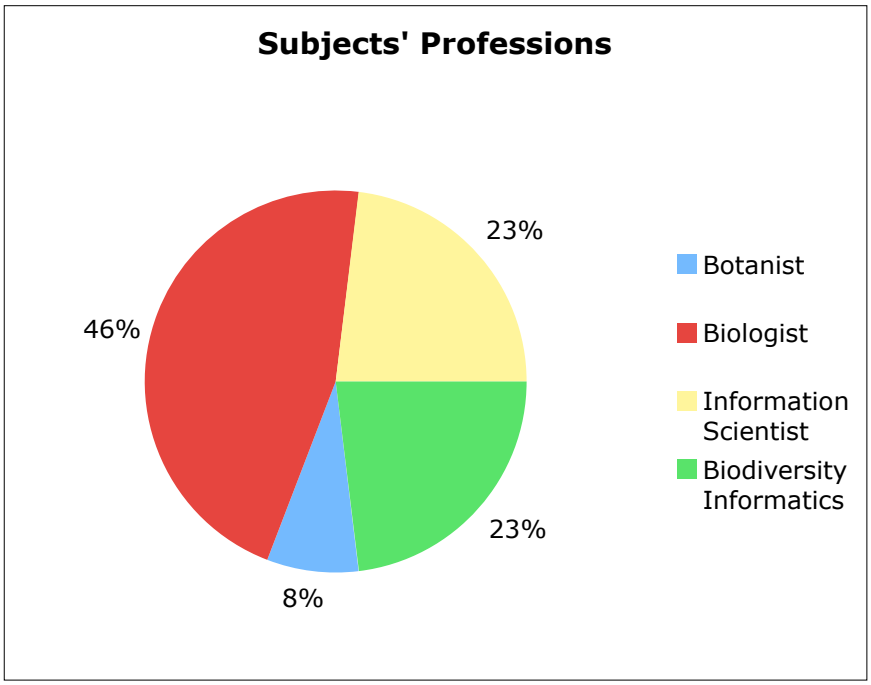


Figure 7.1 Subjects' Professions

The subjects were recruited in a professional workshop coming from several biodiversity domains. Most of the subjects have been in their chosen field for over 10 years. Figure 7.2 displays information about how many years the subjects have been in the profession that was identified in Figure 7.1.

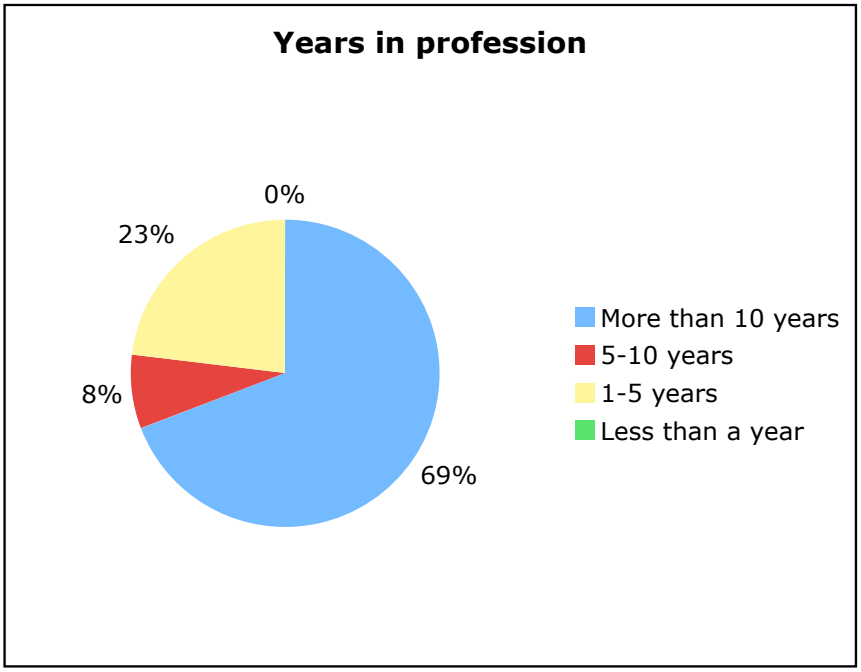


Figure 7.2 Years in Profession

Eight out of the 13 subjects identify themselves as being frequent users or amateur users of FNA and/or FOC with the remaining 5 being unfamiliar with FNA or FOC. Although most of the subjects claimed they were not very familiar with FNA or FOC, they were quite familiar with biological identification.

Most (69%) of them performed some kind of biological identification and used dichotomous keys either weekly or annually. Similar distributions are shown in Figure 7.3 and Figure 7.4. We also noticed that the spoken language used by most of our subjects' was English (11 out of 13).

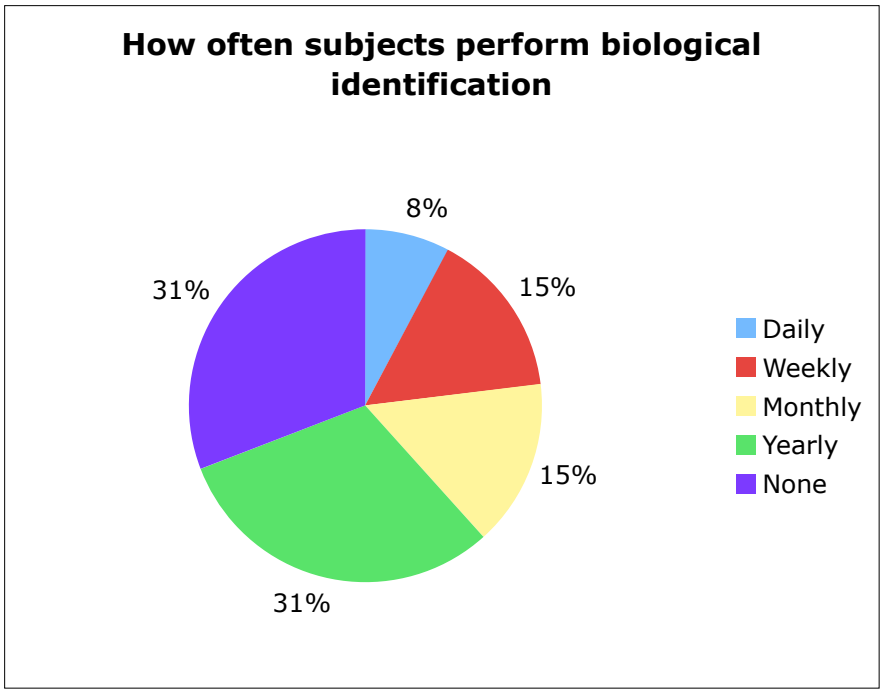


Figure 7.3 How often subjects perform biological identification

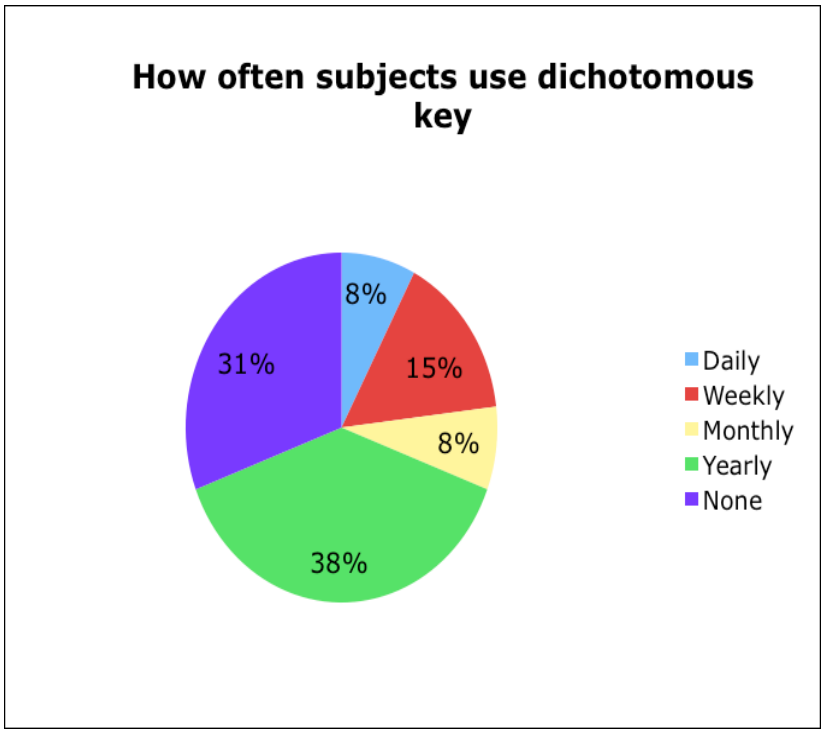


Figure 7.4 How often subjects use dichotomous keys

7.1.2 Evaluation Procedure

The evaluations were held at the "Fine Grained Semantic Markup of Biosystematics Literature" conference. A brief background presentation, which lasted about 10 minutes, was given prior to the evaluations. During the presentation, the experimenter gave a brief introduction to the system and explained the fusion system procedures. Some explanations were also offered concerning possible instances where the system made mistakes.

Following the presentation, the subjects were given printed questionnaires. A copy of the survey can be found in Appendix B. The questionnaire had six sections: information consent form, biographical information about the subject, the genus level fusion, the species level fusion, FNA genus species fusion and FOC genus species fusion. Each subject evaluated the fusion results of 4 different species. In total, the fusion results for 52 different species were evaluated by the subjects.

The evaluation lasted approximately 30 minutes. The evaluators were allowed to ask questions if they had difficulties in making decisions. The subjects were encouraged to answer the questions as accurately as possible instead of answering all of the questions. By the end of the 30-minute session, most (8 out of 13) of the participants were able to complete the questionnaires. The participants who have not completed the questionnaires were allowed to take their questionnaires home and return the finished questionnaires the next morning. A unique number was given to each subject that was marked on every set of evaluation questionnaires so we were able to associate biographical information with

respondent answers.

7.1.3 Evaluation Measures

In standard information retrieval system evaluations, two evaluation measures are popular: precision and recall (Manning et al., 2008).

Precision is defined as the ratio of retrieved documents that are relevant:

Precision = #(relevant items retrieved)/#(retrieved items) = P(relevant|retrieved)

Recall is the fraction of relevant documents that are retrieved:

Recall = #(relevant items retrieved)/#(relevant items) = P(retrieved|relevant)

In this fusion project, precision is used as the evaluation measure to determine the performance of the system and is defined similarly to information retrieval. Precision in this case is defined as the fraction of correct instances among all instances:

Precision = the number of correct instances / total number of instances =P(correct|all)

7.2 Evaluation Results

7.2.1 Results on Numeric and Numeric-like Variables

Different types of data exist in our dataset. A detailed discussion about transforming categorical variables into numeric data and numeric-like variables can be found in Chapter 3. We expect that the fusion results will always be correct for numeric and numeric-like variables except for some cases that are not considered during the modeling process. This expectation is seen in the evaluation results. In the 80 cases of blade length and blade width, the subjects agreed that 78 results are correct. In the 119 cases of blade

shape, 110 were marked as correct by the subjects. After carefully examining the 2 cases of blade length and 9 cases of blade shape, we found that the results were actually correct. The confusion was mainly caused by intra-level or inter-level fusion, and some subjects complained about the adverbs not being included in the results. Here a correct result means that the system's output turned out to be the precise result that we expected. We are not arguing that the output captures all the information included in the descriptions. It may be true that everyone has their own correct fusion result because different people have different perspectives and context of what a correct fusion result should be. Particularly in the case of blade shape, different people might have different opinions on whether lanceolate is closer to linear or closer to ovate. Therefore, what we found to be "correct" is that the system accurately did what it was supposed to do based on the assumptions.

We also note that the results of our system might be less accurate if we were to use a much larger data set. In these experiments, we used a limited dataset in order to engage in data cleaning and transformation on a very detailed scale that allows the system to take care of each case. This may not be possible if we were to use a larger dataset. For example, in the case of blade length, we were able to insert real numbers for each case where the length was referenced in other properties (e.g. petiole length). In the case of blade shape, we were able to identify every term in the vocabulary and pre-map these terms into the shapes in Figure 6.4 or identify them as separate groups. However, the error should not be considered fusion error but rather as error caused by unsuccessful data cleaning and transformation in our data set. The error rate is then determined by how

heterogeneous the data is and to what level the data cleaning and transformation can be accomplished. In an ideal world, once we transform the variables into numeric or numeric-like variables perfectly, there should not be any error involved in fusing them because fusion would be conducted following a set of rigid mathematical functions if they are defined appropriately. Although it is almost impossible that we can achieve 100% accuracy in data cleaning and transformation and fusion by automatic techniques, there are reasons for research in this area. First, the accuracy of automatic data cleaning and transformation and fusion could be estimated by a relative small sample. Second, the performance could be improved by better domain knowledge and the improved techniques. Finally, the techniques developed in this domain might be beneficial to other domain.

7.2.2 Results on Categorical Variables

The following sections will present the performances of our system on categorical variables other than blade shape, blade length and width.

Genus descriptions contain broader information than species level descriptions. We have 111 fusion cases and the subjects agreed that 77% of the fusion results are correct. Among the 17 (15%) cases where the fused results were incorrect, 11 (65%) of them are due to missing information and 5 (29%) of them are due to duplicate information. Figure 7.5 and Figure 7.6 present detailed information about the performance on genus level fusion.

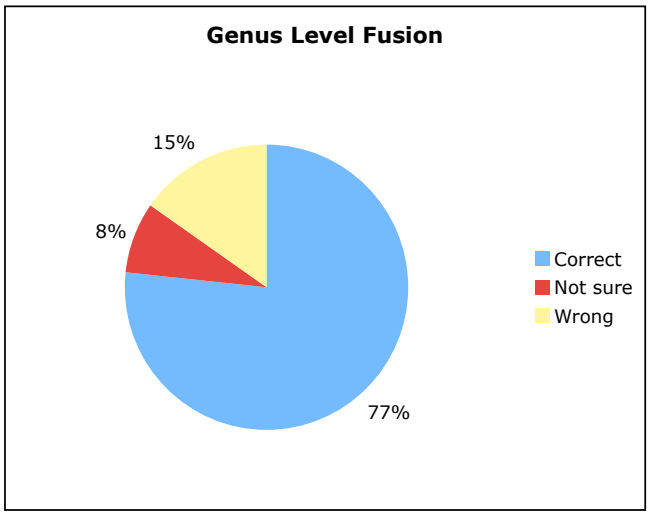


Figure 7.5 Genus level fusion

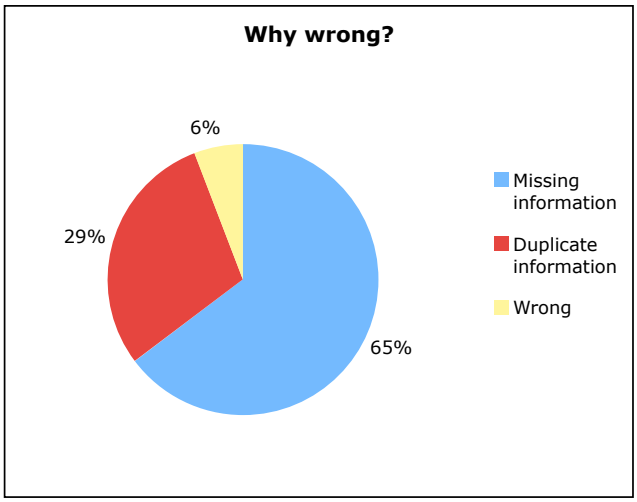


Figure 7.6 Why were the fused results wrong?

We have 124 species fusion cases and the judges agreed that 63% of the fusion results are correct. Among the 30 (24%) cases where the fused results were incorrect, 16 (53%) of them are due to missing information and 5 (17%) of them are due to duplicate information. Figure 7.7 and Figure 7.8 present detailed information about the performance on species level fusion.

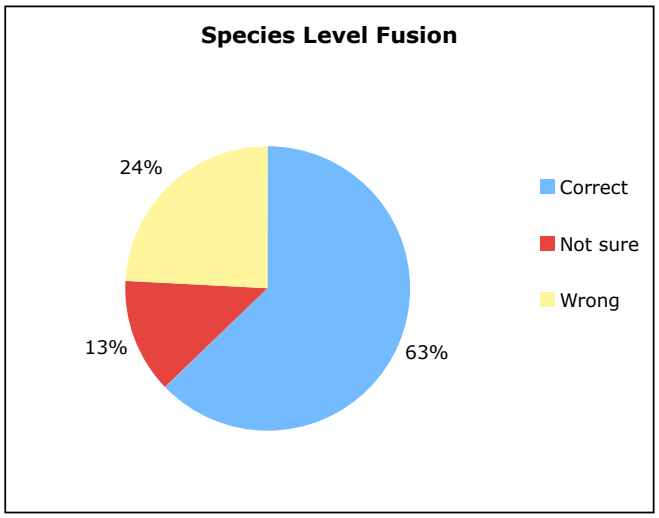


Figure 7.7 Species level fusion

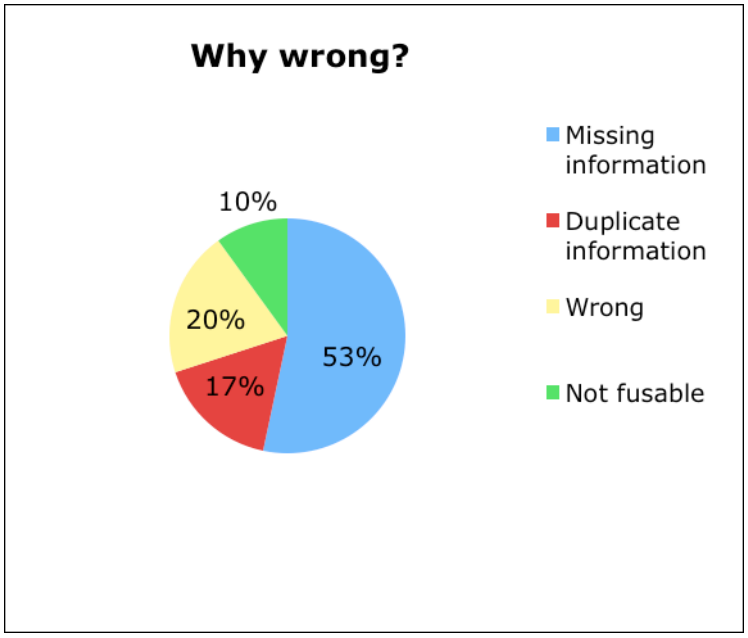


Figure 7.8 Why were the fused results wrong?

We have 90 FNA genus-species fusion cases and the judges agreed that 66% of the fusion results are correct. Among the 22 (24%) cases where the fused results were incorrect, 18 (82%) of them are due to missing information and 2 (9%) of them are due to duplicate

information. Figure 7.9 and Figure 7.10 present detailed information about the performance on FNA genus-species fusion.

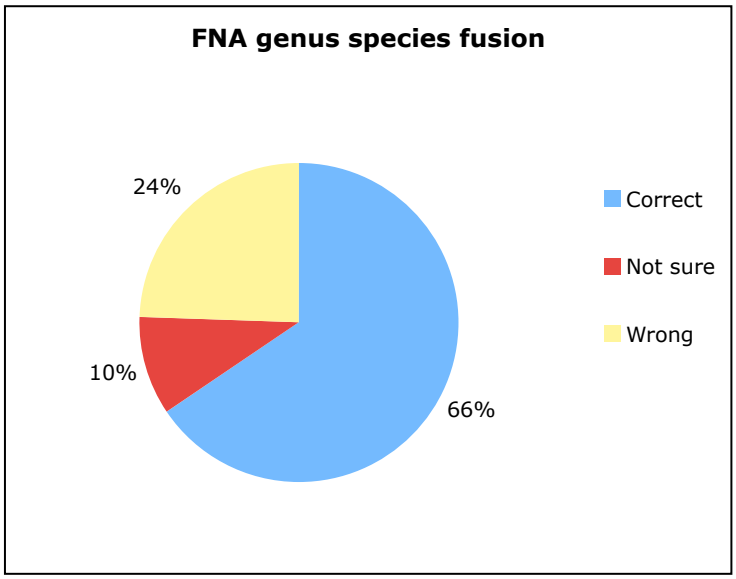


Figure 7.9 FNA genus species fusion

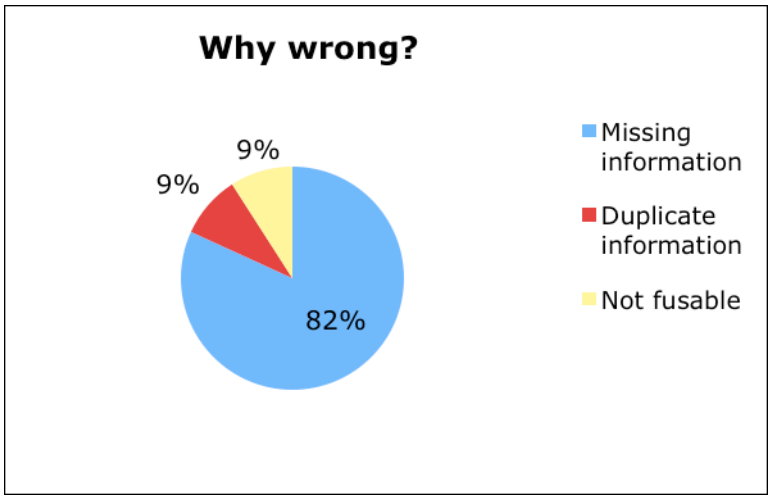


Figure 7.10 Why were the fused results wrong?

We have 80 FNA genus-species fusion cases and the judges agreed that 71% of the fusion

results are correct. Among the 17 (21%) cases where the fused results were incorrect, all of them are due to missing information. Figure 7.11 presents detailed information about the performance on FOC genus-species fusion.

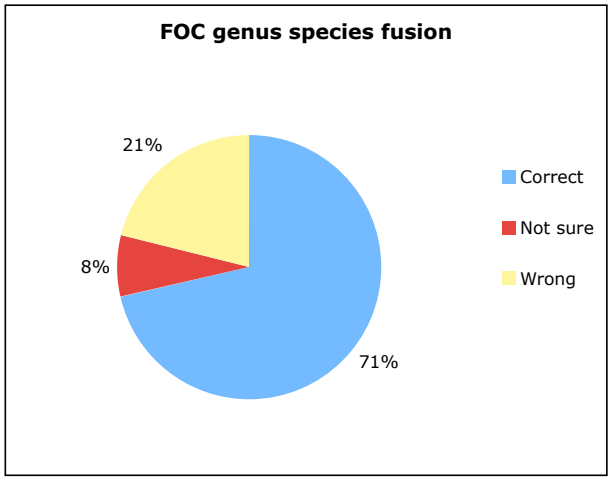


Figure 7.11 FOC genus species fusion

7.3 Result Analysis

The previous section presented our fusion results based on four fusion scenarios. This section will discuss some findings based on the results.

(1) There is no significant difference between the results given control group and experiment group. Inter-subject reliability tests the consistency of evaluation measurement and the repeatability of the evaluation results. We can argue that the results are reliable with some degree of confidence by conducting experiments using control groups. We will compare the results for two groups to determine whether any significant

difference exists between them.

Ten species were chosen from the total of 52 and were randomly selected to test the reliability of the results. The fusion results of the 10 species were evaluated by 8 different subjects in the experiment group. As a control group, a volunteer second-year doctoral student majoring in Plant Biology at the University of Illinois at Urbana-Champaign was asked to evaluate the fusion results for the 10 species. Similar procedures were conducted in the control group experiment. The following table shows the results of the experiments.

Table 7.1 Results for both groups

Number of Instances (total 82)	**Subject**	**Control**
Correct	57	62
Not Sure	7	6
Wrong	18	14

We find that the correlation coefficient between the two results for the two groups is 0.707. Therefore, we developed the following null hypothesis.

Null Hypothesis (a): *there is no significant difference between the results in the subjects' group and the control group.*

We did a paired t-test to determine whether there is significant difference between the results for the two groups. The results are presented in Table 7.2. Therefore, at the 0.05 level of significance, we cannot reject the null hypothesis (a).

Table 7.2 Paired t-test results

P value and statistical significance: The two-tailed P value equals 0.1064 Using conventional criteria, this difference is not considered to be statistically significant. **Confidence interval:** The mean for Group One minus Group Two equals 0.11 95% confidence interval for this difference: From -0.02 to 0.24 **Intermediate values used in calculations:** t = 1.6328 df = 81 standard error of difference = 0.067

(2) There is no significant relationship between the results and the subjects regardless of whether the subject is a biologist or non-biologist. In total, our samples have 406 instances that require fusion.

Null hypothesis (b): *There is no significant difference between the results judged by a biologist or non-biologist.*

Null Hypothesis (b) could be proven by two null hypotheses:

Null hypothesis (b-1): *There is no significant difference between the precision given by biologists or non-biologists.*

Null hypothesis (b-2): *There is no significant difference in the percentage of "not sure" instances given by biologists and non-biologists.*

Our subjects included 6 biologists and 7 non-biologists. Table 7.3 presents the overall results. We will conduct a student t-test to determine whether there is a significant difference in the precision between the biologists and non-biologists.

Table 7.3 Overall information on the subject groups and results

UID	Biologist	Correct	Not Sure	Wrong	Precision	Percentage of Not Sure
3	No	30	4	10	0.681818182	0.090909091
7	Yes	19	0	5	0.791666667	0
11	No	20	7	7	0.588235294	0.205882353
13	Yes	19	4	6	0.655172414	0.137931034
17	Yes	26	0	1	0.962962963	0
21	No	23	2	12	0.621621622	0.054054054
23	No	26	8	13	0.553191489	0.170212766
27	No	20	3	13	0.555555556	0.083333333
31	Yes	13	2	7	0.590909091	0.090909091
37	No	12	2	4	0.666666667	0.111111111
41	Yes	16	3	7	0.615384615	0.115384615
51	No	19	4	1	0.791666667	0.166666667
99	Yes	36	0	1	0.972972973	0

Table 7.4 shows the t-test results on the precision between control group and the experiment group. Table 7.5 presents the t-test results on percentage of "not sure" between control group and the experiment group.

Therefore, at the 0.05 level of significance, we cannot reject the null hypothesis (b-1) or hypothesis (b-2). We could not reject the null hypothesis (b). In summary, we found that there is no significant difference for the results between the judges regardless of whether the judge is a biologist or non-biologist.

Table 7.4 T-test results on the precision between groups

P value and statistical significance: The two-tailed P value equals 0.1134 Using conventional criteria, this difference is not considered to be statistically significant. **Confidence interval:** The mean of Group One minus Group Two equals -0.1248 95% confidence interval for this difference: From -0.2844 to 0.0349 **Intermediate values used in calculations:** t = 1.7198 df = 11 standard error of difference = 0.073

Table 7.5 T-test results on percentage of not sure between groups

P value and statistical significance: The two-tailed P value equals 0.0669 Using conventional criteria, this difference is not considered to be statistically significant. **Confidence interval:** The mean of Group One minus Group Two is 0.0676 95% confidence interval for this difference: From -0.0056 to 0.1408 **Intermediate values used in calculations:** t = 2.0333 df = 11 standard error of difference = 0.033

(3) There is no significant difference between the results of FNA and FOC genus and species fusion.

Null Hypothesis (d): *There is no significant difference between the results of FNA and FOC genus and species fusion.*

Within our sample, we have 31 species that exhibit instances of fusion between genus and species in both FNA and FOC. We did a paired t-test to determine whether there is a significant difference between them. The results in Table 7.6 show we cannot reject the null hypothesis (d).

Table 7.6 Paired t-test on the fusion results between FNA and FOC

P value and statistical significance:
The two-tailed P value equals 0.8943
Using conventional criteria, this difference is not considered to be statistically significant.

Confidence interval:
The mean of Group One minus Group Two equals -0.0129
95% confidence interval for this difference: From -0.2096 to 0.1838

Intermediate values used in calculations:
t = 0.1340
df = 30
standard error of difference = 0.096

(4) There is a significant difference between the results for genus-level fusion and species-level fusion.

Null hypothesis (d): *there is no significant difference between the results of genus-level fusion and species-level fusion.*

We did a paired t-test to determine whether there is a significant difference between the results from genus-level and species-level fusion. The results are shown in Table 7.7. Therefore, at the 0.05 level, we reject the null hypothesis (d). We also found that the

performance on genus-level is better than on the species-level. The reason can be explained by our next argument.

Table 7.7 Paired t-test on the results between genus-level and species-level fusion

P value and statistical significance:
The two-tailed P value equals 0.0277
Using conventional criteria, this difference is considered to be statistically significant.

Confidence interval:
The mean of Group One minus Group Two equals 0.1559
95% confidence interval for this difference: From 0.0180 to 0.2938

Intermediate values used in calculations:
t = 2.2890
df = 38
standard error of difference = 0.068

(5) There is an inverse relationship between the element vocabulary size and the fusion performance.

When we plotted the element vocabulary size and fusion performance for the same element (the elements which contain at least 10 instances of fusion), we found that there is an inverse relationship between them.

We conducted a linear regression between the two variables and found they exhibit a strong correlation coefficient (0.78 > 0.7). The results are presented in Table 7.8. We found the correlation coefficient between the vocabulary size and fusion performance to be -0.78, which means there exists a strong inverse relationship between the two variables. This explains why the genus-level fusion obtained better results than was the

case on the species-level fusion. Genus-level descriptions tend to be more concise and have less variation in their vocabularies for each of the elements.

Table 7.8 Linear regression between element vocabulary size and fusion performance

Sample size: 12 Correlation coefficient (r): -0.77539562794163 Mean x (x): 98.416666666667 Mean y (y): 72.413583399167 Intercept (a): 88.555154263359 Slope (b): -0.16401257440331 Regression line equation: y=88.555154263359-0.16401257440331x

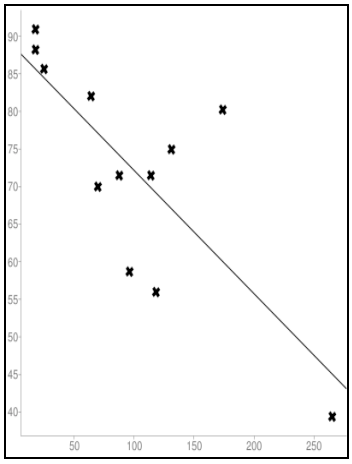


Figure 7.12 Linear regression line between element vocabulary size and fusion performance

(6) The major reason for incorrect fusion results is missing information. We had a total of 405 instances of fusion. Of these, 279 of them were correct and 40 of them were marked as not sure. For those incorrect instances (86 in total), 62 (72.1%) were due to missing information and 12 (14%) were due to duplicate information.

Missing information. When we analyzed the cases where the judges marked the fused results as incorrect due to missing information, we found that the missing information primarily occurred in elements such as surface, petiole, stipule, and margin. These elements were among those elements where mixed information exists and larger vocabulary sizes were observed. The previous section showed that there is an inverse relationship between the variance of the vocabulary size and the fusion performance. This indicates that there is a positive relationship between the level of granularity and the performance of the fusion result.

What is meant by different "levels of granularity"? For example, a description of stipule, "caducous, ovate to lanceolate, 5-9mm, pubescent" could be marked up on 3 different granularity levels:

a. <stipule>ovate to lanceolate, 5-9mm, pubescent.</stipule>

b. <shape> ovate to lanceolate,</shape>
<length> 5-9mm,</length>
<surface> pubescent.</surface>

c. <shape>
<from> ovate</from><conjunction> to</conjunction><to> lanceolate,</to>
</shape>

<length>

<from> 5</from><conjunction>-<conjunction><to>9</to><unit>mm,</unit>

</length>

<surface> pubescent.</surface>

In this situation, we say that c is marked up to a deeper level of granularity in comparison with a and b. When we have two statements regarding the same stipule, information extraction could be done at different levels.

a. <stipule>caducous<stipule>

<stipule>ovate to lanceolate, 5-9mm, pubescent.</stipule>

b.<duration>caducous<duration>

<shape> ovate to lanceolate,</shape>

<length> 5-9mm,</length>

<surface> pubescent.</surface>

These examples show that when we extract the information at a deeper level, we can compare the same input with more detailed information about the input. In the b case, the system will not miss the information about “caducous” because duration differs from shape, length and surface. The system can tell that “caducous” is new information and should be included in the final result. This argument is supported by the evaluation results that the elements that have smaller vocabularies exhibit better performances.

The question concerning what level of granularity is needed in order to obtain the best performance is important because there is cost involved in conducting deeper level information extraction. It is best to mark up every term in the description into its semantic class and obtain the background knowledge associated with it. For example, when we mark up a term, e.g., lanceolate, and we need to know both whether it is a leaf shape and be able to identify its semantic network: it is between linear and ovate, the ratio of length and width is approximately 6:1 to 3:1, and other information if needed.

In that ideal situation, we know exactly what we will fuse and how it can be fused. In real world situations, the costs of marking up at deeper-level granularity do not increase in a linear manner. A more realistic expectation is an exponential increase, because the semantic network for the terms in the vocabulary increases exponentially. In this sense, it is impossible to have a perfect information extraction result for fusion. It is also true that when we attempted to conduct information extraction, we were mapping a more expressive schema (natural language in the sense of having the most expressive power) into a less expressive schema. During this process, we always lose some information in the transition.

Thus, the question is what is the best combination of the level of information fusion and the fusion performance. More specifically, what is the level of robustness of the fusion algorithm such that even when there is some degree of error in the result of information extraction, the fusion can still perform with some level of confidence toward producing

the correct results?

The ultimate goal of information fusion is to model every variable into numeric or numeric-like variables using domain knowledge and available semantic network tools, e.g., domain ontologies. It will allow us to calculate the distance between the semantic classes and perform operations such as union, intersection precisely. Where modeling is possible is at the level of granularity we are seeking for efficient and effective fusion.

Duplicate Information. Duplicate information exists in the fused result because we failed to determine whether two phrases/words contain similar information.

Our project used WordNet as the dictionary for checking whether two terms are synonyms. However, WordNet is a "large lexical database"[24] which is popular in a general way but is not specifically intended to specifically describe plants, or more specifically, leaves.

A better dictionary or ontology would be helpful. For example, WordNet could not tell us whether "swollen" has the same meaning as "expanded" when referring to a leaf base. Given our dataset, a machine-readable plant ontology would be helpful that can be used to examine the semantic relationship between concepts in a more systematic manner. A Plant Ontology (PO)[25] is currently under construction. It is "developing controlled vocabularies (ontologies) that describe plant anatomical and morphological structures and

[24] About WordNet http://wordnet.princeton.edu/ Retrieved on April 4, 2011

[25] Plant Ontology Accessible at http://www.plantontology.org/

growth and developmental stages for all plants. The goal of the POC is to establish a semantic framework for meaningful cross-species queries across gene expression and phenotype datasets from plant genomics and genetics experiments"[26]. PO does not presently contain sufficiently detailed information for it to be useful for processing leaf description in our system. It focuses on genomics and genetics function and deemphasizes morphological descriptions. We expect that a better ontology or dictionary would significantly reduce duplicate information errors in fusion.

In other cases, confusion is caused by unsuccessful matching between a term and a phrase or matching between phrases. For example, "caducous" has the same meaning as "falling early" and the phrases "glabrous except for conspicuous axillary tufts of tomentum" and "glabrous or sparingly pubescent along major veins or in tufts in axils of principal lateral veins and midribs" have duplicate information that "the leaf is generally glabrous but with visible tufts." Since we do not allow matching more than two words, those duplications are not correctly identified.

In order to improve the performance on word-phrase/phrase-phrase synonymous relationship identification, one possible approach might involve loosening our matching criteria to allow for matching between one word and two or more terms.

In order to better identify the duplicate information between phrases, we need to find

[26] About Plant Ontology (PO) Retrieved from http://www.plantontology.org/ on April 11, 2011

some general patterns that would facilitate the matching process. RST was proposed by (Mann & Thompson, 1988), and made several assumptions about how text was constructed: how it organizes words, phrases, grammar, and other linguistic entities. In RST, each sentence (which sometimes can be a noun phrase or a verb phrase) is called a text span. The text span can either be a "nucleus" or "satellite." Nuclei are the most salient parts of a text, while satellites contribute to the nuclei and are secondary. RST claimed "this nuclearity principle is also the basis of hypotactic relations postulated for a lower level of organization in language (i.e. the main-subordinate distinction in complex clauses)" (Taboada and Mann, 2006). Therefore, we can identify the most important information (nucleus) and compare such information instead of comparing entire strings. In this sense, currently available dependency parsing tools, such as the Stanford dependencies tool[27] might be used for future improvements.

7.4 Summary

This chapter presented our fusion experiment results on FNA and FOC samples. We used human judges, and our evaluation data show that we obtained very promising results using our fusion system. The two major issues for current systems are missing information and duplicate information. We found that the error rate of the fusion results is closely related to variance in element vocabulary size. The next chapter will discuss some approaches that might lead to improved performance of the system based on the findings presented in this chapter.

[27] Stanford Dependencies Accessible at http://nlp.stanford.edu/software/stanford-dependencies.shtml

Chapter 8: Conclusions and Future Work

Information fusion is a relatively new field in information science. In this dissertation, information fusion was proposed as a method of improving information access for biodiversity researchers. This chapter will discuss the findings and contributions of this dissertation project and conclude with suggestions for future research.

8.1 Conclusions

This research involved testing the following hypothesis using sample descriptions from FNA and FOC:

There is non-trivial complementary information found in different floras for the same species. It is also true that genus-level (higher-level) descriptions contain non-trivial complementary information on the species-level (lower-level). Automatic/semi-automatic information fusion is useful both because it provides a single access point to the user and provides better data quality. It also provides us with opportunities to detect conflicting information.

We also have three research questions that are closely related to the hypothesis which were proposed in the Introduction Chapter. We sought to answer these research questions in this project. Below are the highlights:

Q1: Does non-trivial complementary information exist in different sources and on different levels? If so, what are their types and how frequently do they occur?

This question was examined in Chapter 5 CDR relationships and automatic CDR relationship identification. We did find that non-trivial complementary information exists in different sources and on different levels. Merging complementary information from different sources and levels allowed us to add 60-75% new information to the original description. 50% new information was added when only one source had the information about a specific character and 10-25% more could be obtained from other sources where complementary, overlapping and contradictory relationships hold. We also found that a small amount of conflicting information exists.

Q2: Is a semi-automatic or automatic information fusion feasible and how can it be evaluated?

Implementing a real fusion system allows us to argue that semi-automatic or automatic information fusion is feasible. The overall system design is presented in Chapter 4. Given the same output from information extraction and CDR relationship identification, the key factors that contribute to the performance of the system include: the quality of the dictionary (or the quality of the domain knowledge), the variance of the vocabulary and the quality of information extraction.

Q3: What challenges exist when applying the method to different texts?

The nature of this project is exploratory. Testing the system on the leaf descriptions from FNA and FOC successfully demonstrated that information fusion is feasible using current techniques. We could also argue that the method could be generalized to different types of texts.

- The method can be generalized to other organs in floras. The construction of a flora description follows some standard guidelines and shares the same domain knowledge. For example, FNA has clear instructions on the sequence for describing characters: presence, number, position, arrangement, orientation……internal parts[28]. Both leaves and other organs have this information organization structure. The methods (automatic information extraction, automatic CDR-relationship identification and different fusion methods) used here can be adapted for use with other organs with minimal change but some domain knowledge might be required in the process.

- The method can be generalized to other collections with certain changes. FNA and FOC aremore structured than smaller floras. Cui (2005) demonstrated that the knowledge obtained from one collection can be used to significantly improve the performance on less structured collections. This means that there exists some knowledge that is consistent among the floras that can be captured in the more structured collections. It sets down the foundation that similar methods might be

[28] Flora of North America Contributors Guide http://fna.huh.harvard.edu/files/FNA%20ContribGuide%202008.pdf Retrieved on

applicable for different collections. Knowledge such as our common understanding of shapes modeled in our method can be easily adapted for use in different situations.

We are arguing that our fusion system is adaptable to other domain if certain changes are made (more discussion in Chapter 8). This approach is generalizable to different texts that have the following features:

1. Sublanguage
2. Hierarchical Information organization property and/or describing the same object or concept
3. Available domain knowledge (dictionary)

Here we outline how the approaches might be useful for another domain by the example from weather reports. The example is from Hunter and Liu, 2006. Weather reports has the three features that we outlined. They are written in sublanguage. Different reports describe the weather for the same place (object). Dictionaries are available in weather domain, e.g. http://www.weather.com/glossary/.

Report 1: TV1: 19/3/02, London, showers, wind 1 kpm, temperature 12 C degree.
Report 2: TV3: 19 March 2002, London, inclement, wind 25 kph, temperature 8-12 C degree.

With the new text, we need to understand the quality and quantity of the data we have

(data profiling process). For example, here we have different data types, e.g. numbers, strings, units, location information, etc. We also want to understand the data length, value range, discrete values and frequency, variances and so on for each variable in the data. With the knowledge from the profiling process, the data preprocessing would include normalize the units and so on. Therefore, after this step the reports would become:

Report 1: TV1: 19/3/02, London, showers, wind 60 kph, temperature 12 C degree.

Report 2: TV3: 19/3/02, London, inclement, wind 25 kph, temperature 8-12 C degree.

The next step is schema design where we need to decide the data schema for information extraction. The schema determines how many variables we have. We also note that the first step (data profiling) and the second step (schema design) might be repeated one after the other till we have a schema appropriate for the text to be processed. There might be different schemas appropriate for the same text. For the weather report we have, we might have a flat schema that has the following variables: source, date, city, outlook, windspeed, and temperature. Therefore, after the information extraction process, we have the two reports as in table 8.1.

After the information extraction, we need to determine what CST relationship exist in the new data set. In weather reports, the same 7 CDR relationships that used in our fusion system might be applicable: Identity, Equivalence, Subsumption, Overlap, Complementary, Contradiction, Refinement.

Table 8.1 Weather reports after information extraction

Report 1	Report 2
<source>TV1</source> <date>19/3/02</date> <city>London</city> <outlook>showers</outlook> <windspeed>1 kpm</windspeed> <temperature> 12 C</temperature>	<source>TV3</source> <date>19 March 2002</date> <city>London</city> <outlook>inclement</outlook> <windspeed>25 kph</windspeed> <temperature> 8-12 C</temperature>

After we have the relationship set, the features and learning algorithm need to be determined for automatic CST-relationship identification. In this step, the knowledge gained from data profiling becomes important.

Table 8.2 Automatic CST relationship identification

Variable	Report 1	Report 2	Relationship
source	TV1	TV3	Complementary
date	19/3/02	19/3/02	Identity
city	London	London	Identity
outlook	showers	inclement	Complementary
windspeed	60 kph	25 kph	Contradiction
temperature	12 C	8 – 12 C	Subsumption

With the output from CST automatic identification, fusion methods needs to be defined based on each relationship. Some example functions used in our fusion system could also be used here: IsSynonym(s1, s2), IsGreater(n1, n2), Initialize(S), Add(S, s), Min(n1, n2), etc. Depending on the fusion methods, the system might output different fusion results. One possible final fusion output of the weather reports could be:

<report>

<source>TV1, TV3</source>

<date>19/3/02</date>

<city>London</city>

<outlook>showers, inclement</outlook>

<windspeed>25-60 kph</windspeed>

<temperature> 8-12 C</temperature>

</report>

With the ideas from the example, what changes and challenges will we face when processing new data? We want to separate the knowledge we obtain from this domain and the procedures we conducted due to the particular characteristics of our sample data. The changes that should be made are presented below.

Different data cleaning and normalization procedure. A similar data profiling procedure can be conducted to determine the quality and quantity of the data. The key issues include: data type, data length, value range, discrete values and frequency, variances, occurrence of null values, typical string patterns and other data properties when applicable. During this stage, we want to obtain as much information as possible about the data we will process. The knowledge obtained will impact decisions concerning data cleaning and normalization and the entire information fusion process.

Different data schema needed for information extraction. Information extraction is a

process where we are mapping the data to a less expressive schema. There are several factors to be considered when designing the schema: the variation of the data, the costs and performance of information extraction at different granularity levels, and the particular fusion task.

Different sets of CST relationships. CST proposed 24 different cross-document sentence relationships on 4 different levels. The 24 relationships cover most sentence pairs in generic data. In our dataset, only 7 out of the 24 relationships were identified. It is typical that different data sets will feature different relationships.

Different feature sets and learning algorithms for relationship identification. Different data sets can have very different features for the purpose of relationship identification. However, the features identified in this project serve as a good starting point. The features (e.g., number of tokens in sentence 1, percentage of common words in sentence 1, percentage of common words in sentence 2, differences in length of sentences) identified in this project are generic features that can be applied to different texts. Other learning algorithms are available, e.g., naïve bayes, and support vector machines. Different learning models can be used if better performance is obtained.

Different fusion methods. Chapter 6 noted that different fusion methods should be constructed and defined when dealing with different data sets and different tasks. The fusion methods defined in this project are of three different types: feature functions, condition functions and action functions. In the case of different datasets, it is possible

that some of the functions will be applicable to the new dataset, e.g., condition functions IsSynonym() and IsAntonym().

8.2 Contributions

This dissertation project makes the following contributions to the research community:

- Improves our understanding of the underlying organization of information in taxonomic literature by developing a taxonomy of the cross-description relationships.
- The features used in automatic cross-description relationship identification that can be useful for other similar types of research projects and tasks.
- Developed a semi-automatic information fusion approach (information fusion based on cross-description relationships).
- Built an information fusion system that can be generalized to other types of text and identified the challenges of adapting the system.

The main objective of information fusion is to achieve refined estimates of the object. There are wide applications of our fusion system. For example, our fusion system would be useful in following two scenarios:

Information fusion simulates what taxonomist do in real life. Our fusion system allows the experts to gain a perception of the level of agreement or disagreement between

the descriptions of the same taxon being investigated. If there is consensus then the user knows that this character must be important to describing the taxon; if there is consensus but the attributes of the characters differ then the user needs to do some research as to why; if there is conflict in the sense that the character only appears in one of the descriptions and not the other then there is value in knowing of its existence; if there is conflict because two characters do not agree/overlap then a user will want to explore why that is. In the case of conflict, the conflict may be due to just a lack of information provided because one treatment is more general than another; an error based on an editorial process; or may highlight an interesting scientific discovery that expands knowledge on the variability found in this taxon.

The fusion system would benefit the authors/editors as an author assistance system. The system could be used to help authors when writing/editing new treatments, they are able to compare their facts with the facts in other floras. They can also benefit from our fusion system on judging how well a treatment is written based on noticing characters that were not recorded or errors that have been made.

Writing or editing a new treatment description is a complex problem as there is a lot of natural language involved, possibly many languages, and the interpretation of qualitative characteristics (long, short, tall, small...) especially when compared to more quantitative measures. This is also a problem with shape as describing shape variation can be difficult especially if you do not know if an intermediate shape is possible.

8.3 Future Work

In addition to improving the performance of the information fusion system as discussed at the end of the last chapter, this research can be extended in the following directions:

- Improving fusion results by fusing with modifiers and conjunctions. We neglect all modifiers and conjunctions that were treated as stop words in our current fusion system. We admit that those modifiers contain important information that might be useful for our users, particularly the frequency and certainty modifier. The frequency modifier represents the distribution of the presence of a specific state of a character, and the certainty modifier represents the possibility of a state. The conjunctions "and," "to," and "or" represent different meaning in original descriptions. We would like to take the distribution, possibility and conjunctions into consideration in future fusion systems. Below are some ideas concerning how fusion might work in the cases where we need to deal with possibilities. For example, two statements about the arrangement of leaves:

<arrangement> often alternate, rarely opposite</arrangement>

<arrangement> often opposite, rarely alternate</arrangement>

Using our current system would produce a result that would be "alternate, opposite" without taking the possibility into account. We assume here that "rarely" represents

the possibility of 0.01% to 20%, and “often” represents 20% to 80%. We thus can have the statement marked up as follows:

<possibility value=“20% - 80%”> alternate</possibility>

<possibility value =“0.01% - 20%”> opposite</possibility>

<possibility value=“20% - 80%”> opposite</possibility>

<possibility value =“0.01% - 20%”> alternate</possibility>

There are different ways in which we can perform fusion. For example, below are two methods we might choose to use:

a. Take the smaller lower bound of the possibility and larger upper bound of the possibility as the possibility for a final result. Therefore, in our case, the possibility for alternate would be 0.01% to 80% and the same for opposite.
b. Take the average of the lower bound as the lower bound and the average of the upper bound as the upper bound of the possibility. Therefore, we have 10% to 50% for alternate and the same for opposite.

Additional fusion methods can be proposed to fuse the descriptions in the example. Handling the modifiers (particularly frequency and certainty modifiers) would be another challenge for fusion when constructing the fusion methods. The fusion methods are required to be both mathematically correct and semantically

meaningful.

- Improving fusion results by encoding more variables into numeric or numeric-like variables. As discussed above, the ultimate goal of information fusion is not to construct fusion methods that are sophisticated enough to handle each case in the data. More efforts should be made in the areas of data cleaning and modeling. That is how information fusion can be accomplished correctly and meaningfully. Our current system takes leaf blade shape as the exemplar variable to be modeled as a numeric-like variable. There is more than one way to do modeling. Future research might be done to compare different methods of modeling and/or modeling different variables, e.g., color, apex, base, etc. The modeling is useful both from the perspective of information fusion and for other text mining tasks.

- Improving CST-relationship identification by finding a better feature set and using a better learning algorithm. The performance of CST-relationship identification is determined by two factors: the feature set and the learning algorithm. Finding the best performance feature set is not a one step process but is instead a cumulative and iterative procedure. The feature set we now use could serve as a starting point. Improved feature sets can be identified using more knowledge about the data and more trial and error experiments. Finding better learning algorithms can also improve the performance. There are assumptions that are found behind each learning algorithm. The

performance of the learning algorithms depends on how well the data fit with the background assumptions and how the algorithm fits the learning task. Moreover, efforts can be made to find the best combinations of feature sets and algorithms.

- Improving inconsistency detection methods. Inconsistency detection methods were proposed to detect potential contradictions existing in the data. Different types and levels of inconsistency exist in our dataset. Future research should include investigating better inconsistency detection techniques.

- Extending the fusion system to different and larger biodiversity collections. Our current fusion system was built based on leaf description in FNA and FOC. We could test our system immediately on other organ descriptions in FNA and FOC using slightly changed XML schemas, for example, and different character sets should be defined. However, interesting findings might be obtained using larger and different collections. When processing different types of information, fusion requires the encoding of domain knowledge, particularly during the fusion method construction. Different data cleaning and transformation procedures can be expected when processing different and larger datasets, so it is reasonable to expect that the data will be more heterogeneous than smaller sets. Future research questions regarding this perspective should include how to efficiently encode the domain knowledge into the process of fusion.

- Investigating the possible applications that might be built on top of the fusion system. For example, an information retrieval system might be built on top of our fusion system. The ultimate goal of text mining in the biodiversity area is to facilitate the process of finding useful information for users. We would like to test whether information retrieval can be improved by background fusion as well as the improvement on recall and precision. The design of the information retrieval user experiment would be as follows. Let's take FNA and FOC as the exemplar datasets in the retrieval experiment. Two parallel information retrieval systems are going to be implemented with the same interface, index and retrieval algorithms. The only difference between the two systems is that one system only includes the keywords from the original description of FOC for each species and the other system also includes the keywords from the fusion results of genus and species level descriptions of FOC and FNA. We would give the users the same species and ask them to identify the scientific names for the species by using the two retrieval systems.

  Firstly, we would expect the recall of the retrieval system is improved by using the fusion results since we add terms that are not in the original FOC description. For example, the term ovate is not in the original FOC description but in the same FNA description, the user could not be able to identify the species since the searching result won't include the species. With the analysis in Chapter 5, by combining information in different sources and levels we are

able to add about 60-75% new information to the original descriptions, of which 50% could be added when the character is only described in another source and level and 10-25% can be gained while complementary, overlapping and contradictory relationships hold. Therefore, by fusion, the recall could be improved by 60-75% as the upper bound with the fusion results. And the lower bound would be 50% where the new information could only be found in another source or level.

Second, we would expect that precision is going to be impacted because more false positive results are returned with the same query. For example, the term ovate was not in the original FOC description but was in the FNA description for the same species, we will get more species with ovate leaves (false positives) with fusion results. But the impact should be minimized when combines with other terms in the search query.

References

Abascal, R., & Sanchez, J. A. (1999). X-tract: Structure extraction from botanical textual descriptions. In *Proceeding of the String Processing and Information Retrieval Symposium and International Workshop on Groupware*, Cancun, Mexico, September 21-24,1999, pp. 2-17.

Abney, S. (2002). Bootstrapping. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 360-367.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, New Orleans, Louisiana, USA, September 2001, pp. 276-284.

Atsuhiro, T., & Aihara, K. (2002). DVHMM: variable length text recognition error model. *In Proceedings of International Conference on Pattern Recognition (ICPR02)*, Vol.3. pp. 110–114.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Bagga, A., & Biermann, A. W. (1997). Analyzing the complexity of a domain with respect to an information extraction task. In *Proceedings of the Tenth International Conference on Research on Computational Linguistics (ROCLING X),* pp. 175-194.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech Communication*, (42): 93-108.

Bellegarda, J.R. (2000). Exploiting latent semantic information in statistical language modeling, In *Proceedings of the IEEE*, 88(8): 1279-1296.

Biodiversity Heritage Library (BHL). Accessible at http://www.biodiversitylibrary.org/.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology*, 37(1): 3-50.

Bloch, I., Hunter, A., Appriou, A., Ayoun, A., Benferhat, S., Besnard, P., Cholvy, L., Cooke, R., Cuppens, F., Dubois, D., Fargier, H., Grabisch, M., Kruse, R., Lang, J., Moral, S., Prade, H., Saffiotti, A., Smets, P. & Sossai, C., Fusion: general concepts and characteristics, *International Journal of Intelligent Systems,* 16 (2001) (10): 1107-1134.

Borkar, V., Deshmukh, K., & Sarawagi, S. (2001). Automatic segmentation of text into structured records. In *Proceedings of ACM SIGMOD international conference on Management of data*, Santa Barbara, California, 30(2): 175-186.

Bose, R. (2004). Natural language processing: current state and future directions. *International Journal of the Computer, the Internet and Management*, 12(1): 1-11.

Brach, A. R., & Song, H. (2006). eFloras: new directions for online floras exemplified by the Flora of China Project. *Taxon,* 55(1): 188-192.

Bratko, A., & Filipic, B. (2006). Exploiting structural information for semi-structured document categorization. *Information Processing and Management*, 42(3): 679-694.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, pp. 161-175.

Cholvy, L., & Hunter, A. (1997). Fusion in logic: A brief overview, In *Proceedings of the Fourth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'97),* Lecture Notes in Computer Science vol. 1244 (1997), pp. 86-95.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1): 51-89.

Cowie, J., Nirenburg, S., & Molina-Salgado, H. (2000). Generating personal profiles. In *Proceedings of The International Conference On MT And Multilingual NLP*. pp. 23(1-7). Retrieved from http://www.mt-archive.info/BCS-2000-Cowie.pdf on January, 2, 2010.

Cui, H. (2005). Automating semantic markup of semi-structured text via an induced knowledge base: a case-study using floras. *Ph. D Dissertation*. University of Illinois at Urbana-Champaign. 2005.

Cui, H. (2008). Unsupervised Learning for Semantic Markup of Biodiversity Literature. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, Pittsburgh, PA, June 16-20, 2008, pp. 25-28.

Cui, H., & Heidorn, P. B. (2007). The reusability of induced knowledge for the automatic semantic markup of taxonomic descriptions. *Journal of the American Society for Information Science and Technology*, 58(1): 133-149.

Culotta, A., Bekkerman, R., & McCallum, A. (2004). Extracting social networks and contact information from email and the Web. In *Proceedings of First CEAS (Conference on Email and Anti-Spam)*, Retrieved from http://www2.selu.edu/Academics/Faculty/aculotta/pubs/culotta04extracting.pdf on January 23, 2010.

Cunningham, H., Maynard, D., Bontcheva, C. K., Tablan, V., & Dimitrov, M. (2002). Developing language processing components with GATE. *Technical Report*, University of Sheffield, Sheffield, UK.

Cunningham, S. J., Witten, I. H., & Littin, J. (2000). Applications of machine learning in information retrieval. *Annual Review of Information Science and Technology*, 34(6): 341-384.

Curran, J.R. (2003). Blueprint for a high performance NLP infrastructure. In *Proceedings of Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*. Edmonton, Canada, 2003, pp. 40-45. Retrieved from http://acl.ldc.upenn.edu/W/W03/W03-0806.pdf on January 2, 2010.

Dallwitz, M. J. (1993). DELTA and INTKEY, In Fortuner, R., editor, *Advances in Computer Methods for Systematic Biology*, Johns Hopkins University Press.

Dubois, D. & Prade, H. (2004). Possibilistic logic: a retrospective and prospective view. *Fuzzy Sets and Systems*, 144: 3-23.

Efron, M. (2009). Generative model-based metasearch for data fusion in information retrieval. In *Proceedings of the 2009 Joint International Conference on Digital Libraries (JCDL 09)*. ACM, New York, pp. 153-162.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Foster, J. (2007). Treebanks gone bad: parser evaluation and retraining using a Treebank of ungrammatical sentences. *IJDAR (2007),* pp. 129-145.

Fox, B. A. (1987). *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge: Cambridge University Press.

Fox, E. A., Koushik, M. P., Shaw, J., Modlin, R. & Rao, D. (1993). Combining evidence from multiple searches. In *Proceedings of The First Text REtrieval Conference (TREC-1)*, Gaitherburg, MD, USA, March 1993, pp. 319-328.

Frasconi, P., Soda, G., & Vullo, A. (2002). Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, 18 (2-3): 195-217.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1): 9-42.

Greenberg, J., Spurgin, K., & Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of experts' opinions. *International Journal of Metadata, Semantics and Ontologies*, 1(1): 3–20.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2): 199-220.

Hahn, U., & Wermter., J. (2006). Levels of natural language processing in text mining. In Ananiadou, S., & Mcnaught, J., editors, *Text Mining for Biology and Biomedicine*, Boston, Artech House, pp. 13-41.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

Hatzivassiloglou, V., Klavans, J., & Eskin, E. (1999). Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* College Park, Maryland, June 1999. Association for Computational Linguistics, pp. 203-212.

Hunter, A. (2000). Merging potentially inconsistent items of structured text, *Data and Knowledge Engineering*, 34(3): 305-332.

Hunter, A. (2002). Logical fusion rules for merging structured news reports, *Data and Knowledge Engineering*, 42:23-56.

Hunter, A., & Liu, W. (2006). Fusion rules for merging uncertain information, *Information Fusion,* 7(1): 97-134.

Hunter, A., & Summerton, R. (2004). Fusion rules for context-dependent aggregation of structured news reports, *Journal of Applied Non-classical Logic*, 14(3): 329-366.

Jardine, N. (1969). A logical basis for biological classification. *Systematic Zoology,* 18: 37-52.

Kludas, J., Bruno, E., & Marchand-Maillet, S. (2007). Information fusion in multimedia information retrieval. In *Proceedings of 5th international Workshop on Adaptive Multimedia Retrieval (AMR)*, Paris, France, July 5-6, 2007. Retrieved from http://www.multimatch.eu/docs/publications/Kludas.amr07.pdf on January 2, 2010.

Kokar, M. M., Tomasik J. A., & Weyman, J. (2004). Formalizing classes of information fusion systems. *Information Fusion: An International Journal on Multi-Sensor, Multi-Source Information Fusion*, 5(3): 189-202.

Koller, D. & Sahami, M. (1996) *Toward Optimal Feature Selection*. Morgan Kaufmann.

Langseth, H., & Nielsen, T.D. (2006) Classification using hierarchical naive bayes models. *Machine Learning*, (63): 135-159

Lehrberger, J. (1982). Automatic translation and the concept of sublanguage. In R. Kittredge and J. Lehrberger, editors, *Sublanguage: Studies of Language in Restricted Semantic Domain*, Walter de Gruyter, pp. 9-26.

Lewis, D. D. (1998). Naive (Bayes) at forty: the independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning,* April 21-23, 1998, pp. 4-15.

Lewis, D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93.

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of Library and Information Science*. NY, Marcel Decker, Inc.

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., & White, F. (2004). Revisiting the jdl data fusion model ii. In *Proceedings of the Seventh International Conference on Information Fusion (FUSION)*, Stockholm, Sweden, 2004, pp. 1218-1230.

Lydon, S., Wood, M., Huxley, R., & Sutton, D. (2003). Data patterns in multiple botanical descriptions: implications for automatic processing of legacy data. *Systematics and Biodiversity*, 1(2): 151–157.

Mann, G. S., & Yarowsky, D. (2005). Multi-field information extraction and cross-document fusion, In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, June 25-30, 2005, pp. 483-490.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization, *Text*, 8(3): 243-281.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Marcu, D. & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting, Association for Computational Linguistics*, University of Pennsylvania, Philadelphia, pp. 368-375.

Maziero, E. G., Jorge, M. L. C., & Pardo, T. A. S. (2010). Identifying Multidocument Relations. In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science*. June 8-12, 2010, Funchal/Madeira, Portugal.

McGuinness, D.L. (2003), Ontologies come of age, In: Fensel, D., Hendler, J.A., Lieberman, H. & Wahlster, W., editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, MIT Press (2003), pp. 171-194.

Mitchell, T. M. (1997). *Machine Learning*, McGraw-Hill.

Miyabe, Y., Takamura, H., & Okumura, M. (2008). Identifying cross-document relations between sentences. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08),* pp. 141-148.

Montague, M., & Aslam, J.A. (2002). Condorcet fusion for improved retrieval. In *Proceedings of 11th ACM International Conference on Information and Knowledge Management (CIKM),* pp. 538-548.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models, *Information Retrieval*, 7(3-4): 317-345.

Radev, D. R. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure, In *the 1st SIGdial workshop on Discourse and dialogue*, pp. 9-26. Retrieved from http://portal.acm.org/ft_gateway.cfm?id=1117745&type=pdf&CFID=16234159&CFTOKEN=98788572 on April 4, 2010.

Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4): 1-11.

Ravn, T., & Høedholt, M. (2009). How to measure and monitor the quality of master data. *Information Management Magazine.* Retrieved from http://www.information-management.com/issues/2007_58/master_data_management_mdm_quality-10015358-1.html?ET=informationmgmt:e963:2046487a:&st=email on April 4, 2011.

Richardson, S. D. (1994). Bootstrapping statistical processing into a rule-based natural language parser. In *Proceedings of ACL Workshop On The Balancing Act: Combining Symbolic And Statistical Approaches To Language*, pp. 96-103.

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? In *Proceedings of the IEEE*, 88(8): 1270-1278.

Steinberg, A. & Bowman, C. (2004). Rethinking the JDL Data Fusion Levels, In *Proceedings of the Military Sensing Symposia (MSS) National Symposium on Sensor and Data Fusion (NSSDF)*, Retrieved from https://svn.v2.nl/andres/Documentation/Multisensor%20data%20fusion/Rethinking%20JDL%20Data%20Fusion%20Levels_BowmanSteinberg.pdf on January 23, 2010.

Stuessy, T. F. (1990). *Plant Taxonomy: the Systematic Evaluation of Comparative Data*, Columbia University Press.

Subramaniam, L., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V. S., Kamesam, P. V., & Kothari, R. (2003). Information extraction from biomedical literature: methodology, evaluation and an application. In *the Proceedings of the Twelfth International Conference on Information and Knowledge Management*. New Orleans, LA, pp. 410-417.

Taboada, M. & Mann, W. C. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8 (3): 423-459.

Tang, X. (2007). Extraction and use of structured information in full-text retrieval: a case study. *Ph.D. Desertation*, University of Illinois at Urbana-Champaign, 2007.

Tang, X. (2008). Enhancing keyword-based botanical information retrieval with information extraction. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08),* New York, NY, 2008, pp. 789-790.

Taylor, A. (1995). Extracting knowledge from biological descriptions. Presented at *2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases,* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9081&rep=rep1&type=pdf on January 23, 2010.

Taylor, A. G. (1999). *The Organization of Information*, Libraries Unlimited, Inc., Third Edition.

Torra, V., & Narukawa, Y. (2007). *Modeling Decisions: Information Fusion and Aggregation Operators*, Springer.

Ulicny, B., Kokar, M. M., & Matheus, C. J. (2010). Uses of ontologies in open source blog mining. In Obrst, L., & Janssen, T., editors, *Ontologies and Semantic Technologies for the Intelligence Community: Frontiers in Artificial Intelligence and Applications*, IOS Press Amsterdam, 2010, pp. 37-55.

Valet, L., Mauris, G., & Bolon, P. (2000) A statistical overview of recent literature in information fusion, In *Proceedings of the Third International Conference on Information Fusion*, July 10-13, 2000, pp. MOC3(22-29), Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=862457&isnumber=18695 on April 4, 2010.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Huebner, S. (2001). Ontology-based integration of information - a survey of existing approaches. In *Proceedings of the workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI),* pp. 108-117

Wan, X. (2008). Using only cross document relationships for both generic summarizations. *Information Retrieval*, 11(1): 25-49.

Wang, S., & Pan, J. Z. (2005) Ontology-based representation and query of colour descriptions from botanical documents. *OTM Conferences* 2005(2): 1279-1295.

Wang, S., & Pan, J. Z. (2006). Integrating and querying parallel leaf shape descriptions. In *Proceedings of International Semantic Web Conference ISWC2006,* Springer, 2006, pp. 668–681.

Wang, S., Schuurmans, D., Peng, F., & Zhao, Y. (2005). Combining statistical language models via the latent maximum entropy principle. *Machine Learning*, 60(1): 1-22.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques,* Morgan Kaufmann, Second Edition.

Wood, M. & Wang, S. (2004). Motivation for "ontology" in parallel-text information extraction. In *Proceedings of ECAI-2004 Workshop on Ontology Learning and Population (ECAI-OLP)*, Valencia, Spain, August 2004.

Wood, M., Lydon, S., Tablan, V., Maynard, D., & Cunningham, H. (2004). Populating a database from parallel texts using ontology-based information extraction. In Meziane, F. & M´tais, E., editors, *Proceedings of Natural Language Processing and Information Systems, 9th International Conference on Applications of Natural Languages to Information Systems*, pp. 254-264.

Wu, S., & McClean, S. (2006). Improving high accuracy retrieval by eliminating the uneven correlation effect in data fusion. *Journal of the American Society for Information Science and Technology (JASIST),* 57(14): 1962–1973.

Zhang, Z., Blair-Goldensohn, S., & Radev, D. R. (2002). Towards CST-enhanced summarization. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002)*, Edmonton, pp. 439-445.

Zhang, Z., Otterbacher, J., & Radev, D. (2003) Learning cross-document structural relationships using boosting. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management.* November 03-08, 2003, New Orleans, LA, USA.

Zhao, R., & Grosky, W. I. (2002). Narrowing the semantic gap - improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, 4(2): 189-200.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5): 358-75.

Appendix A: Leaf Sub-organ Level Markup Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<element name="Leaves" xmlns="http://relaxng.org/ns/structure/1.0">
  <interleave>
    <zeroOrMore>
      <element name="FNA_genus">
        <interleave>
          <zeroOrMore>
            <element name="leavesla">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="solidshape">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="texture">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="color">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="ns">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="bladewidth">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="petiole">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
```

```
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
```

```
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="basela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apexla">
    <text/>
  </element>
```

```
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="fnagenus">
    <interleave>
      <zeroOrMore>
        <element name="origtext">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="focspecies">
    <interleave>
      <zeroOrMore>
        <element name="origtext">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="focgenus">
    <interleave>
      <zeroOrMore>
        <element name="origtext">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="fnaspecies">
    <interleave>
      <zeroOrMore>
```

```
        <element name="origtext">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="genus">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="bladewidth">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="petiole">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
```

```
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
```

```
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="basela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apexla">
    <text/>
  </element>
```

```
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="FNA_FNA">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="bladewidth">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="petiole">
          <text/>
        </element>
```

```
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
```

```
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
```

```
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="basela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apexla">
```

```
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="species">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="bladewidth">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="petiole">
```

```
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
```

```
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="basela">
    <text/>
  </element>
</zeroOrMore>
```

```
      <zeroOrMore>
        <element name="apexla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>

</zeroOrMore>
<zeroOrMore>
  <element name="FOC_FOC">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="bladewidth">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
        </element>
```

```
</zeroOrMore>
<zeroOrMore>
  <element name="petiole">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
```

```
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
```

```
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="basela">
```

```
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="apexla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="FNA_species">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="bladewidth">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
```

```
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiole">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
```

```
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobela">
    <text/>
  </element>
</zeroOrMore>
```

```
      <zeroOrMore>
        <element name="basela">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="apexla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>

</zeroOrMore>

<zeroOrMore>
  <element name="FOC_genus">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="color">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladewidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiole">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
    <text/>
  </element>
```

```
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
  </element>
</zeroOrMore>
```

```
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
```

```
        <element name="lobela">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="basela">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="apexla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="toothla">
          <text/>
        </element>
      </zeroOrMore>
    </interleave>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="FOC_species">
    <interleave>
      <zeroOrMore>
        <element name="leavesla">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="ns">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="solidshape">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
        <element name="texture">
          <text/>
        </element>
      </zeroOrMore>
      <zeroOrMore>
```

```
  <element name="color">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladewidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiole">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladelength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletlength">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="leafletwidth">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="vein">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipule">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surface">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="duration">
```

```
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="arrangement">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="attachment">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexity">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladeshape">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="margin">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="lobe">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="base">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="apex">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="tooth">
    <text/>
```

```
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="other">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="petiolela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="bladela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="veinla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="stipulela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="surfacela">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="durationla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="complexityla">
    <text/>
  </element>
</zeroOrMore>
<zeroOrMore>
  <element name="marginla">
    <text/>
  </element>
```

```
          </zeroOrMore>
          <zeroOrMore>
            <element name="lobela">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="basela">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="apexla">
              <text/>
            </element>
          </zeroOrMore>
          <zeroOrMore>
            <element name="toothla">
              <text/>
            </element>
          </zeroOrMore>
        </interleave>
      </element>
    </zeroOrMore>
  </interleave>
</element>
```

Appendix B: Sample Questionnaire

Informed Consent Form

Title of Project: Information Fusion in Taxonomic Descriptions

Responsible Principal Investigator:

Qin Wei, Doctoral student, University of Illinois at Urbana-Champaign
Dr. Bryan Heidorn, University of Arizona

Purpose of the Study:

This study seeks to address the problem of information access in the biodiversity area by providing a single access point to multiple sources. The main objectives of this work are to: (a) Gain in-depth insight into the information properties of taxonomic descriptions and the constancy of fact-based information in those texts; (b) Build a cross-sentence relationship taxonomy based on leaves descriptions from multi-sources; (c) Study the implications of the cross-sentence relationships to text mining in general and information fusion in particular; (d) Develop an information fusion system that provides a single access point for multiple floras.

Procedures to be followed:

You will need to fill out a questionnaire regarding the output of the information fusion system.

Discomforts and Risks:

There are no foreseeable risks to you beyond those normally encountered in your daily life while participating in this study.

Benefits:

The benefits from participating in this study will include gaining more insights into the information properties of the floras and becoming more familiar with the information systems in this area.

Statement of Confidentiality:

We will keep your participation in this research study confidential. All data will be kept in a password-protected computer that is locked up so that only the

researchers involved in this project will have access to it. If we write a report or article that is either published in a journal or in a professional conference presentation about this study, we will describe the study results in a summarized manner so that you will not be personally identified.

Voluntariness:

Taking part in this research study is completely voluntary. You may choose not to take part at all. If you decide to be in this study, you may stop participating at any time. If you decide not to be in this study, you can stop participating at any time.

Contact information:

If you have further questions, please contact Qin Wei, qinwei2@illinois.edu. If you have questions about your rights as a research subject please contact the University of Illinois Institutional Review Board at 217-333-2670 or by e-mail at irb@uiuc.edu. You may call the IRB if you identify yourself as a research subject.

Your signature indicates that this research study has been explained to you, that your questions have been answered, and that you agree to take part in this study.

- I have read and understand the above consent form and voluntarily agree to participate in this study.

_____ _____

Participant Signature Date

Section 1 - About you

1.1 Are you a _____.

- Botanist
- Biologist other than botanist, please specify your field_____
- Information Scientist
- Other, please specify_____

Your specialty in the field _____.

1.2 Your years of experiences in the profession identified in 1.1?

- More than 10 years
- 5 - 10 years
- 1 - 5 years
- Less than a year

1.3 Are you familiar with Flora of North America and/or Flora of China.

- Expert(writer and/or editor)
- Frequent User
- Amateur
- None

1.4 How often do you perform biological identification?

- Daily
- Weekly
- Monthly
- Yearly
- None

1.5 How frequently do you use dichotomous key?

- Daily
- Weekly
- Monthly
- Yearly
- None

1.6 Your native language is_____.

- English
- Not English, please specify_____

Section 2 - Genus Level Fusion

FNA_Genus: Leaves sometimes tardily deciduous; stipules falling early. Leaf blade ovate to obovate or elliptic, base usually oblique, sometimes cordate or rounded to cuneate, margins serrate to doubly serrate; venation pinnate.
FOC_Genus: Leaves distichous, blade base ± oblique, margin doubly or simply serrate; venation pinnate; secondary veins extending to margin, each ending in a tooth.

Table B.1 Genus level fusion

Element	FNA genus	FOC genus	Fused Result	Correct?	Why wrong
Base	usually oblique, sometimes cordate or rounded to cuneate,	oblique,	usually oblique, sometimes cordate or rounded to cuneate,	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Margin	serrate to doubly serrate;	doubly or simply serrate;	serrate to doubly serrate;	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.

Table B.1 Genus level fusion (cont.)

Element	FNA genus	FOC genus	Fused Result	Correct?	Why wrong
Stipule	falling early.	2, lanceolate-ovate to linear, membranous, caducous, leaving a short transverse scar on each side of leaf base.	2, lanceolate-ovate to linear, membranous, caducous, leaving a short transverse scar on each side of leaf base.	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Vein	pinnate.	pinnate; secondary veins extending to margin, each ending in a tooth.	pinnate; secondary veins extending to margin, each ending in a tooth.	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.

Section 3 - Species Level Fusion

FNA_Species:Leaves: petiole 2-6(-8) mm, glabrous or sparsely pubescent with short hairs. Leaf blade elliptic to ovate-obovate, (3.5-)4-5(-6) x 1.5-2.5 cm, base oblique, margins mostly singly serrate (some doubly serrate), apex acute; surfaces abaxially pale, glabrate, adaxially dark green, lustrous, glabrous; lateral veins forking 5 or more times per side.
FOC_Species:Petiole 2-6 mm, pubescent; leaf blade lanceolate-ovate to narrowly elliptic, lamina on two sides of midvein unequal in length and width, 2.5-5 x 1-2 cm, thick, abaxially pea green and pubescent when young, adaxially dark green, lustrous, and pubescent only on midvein, base oblique, margin obtusely and irregularly simply serrate, apex acute to obtuse; midvein depressed; secondary veins 10-15 on each side of midvein.

Table B.2 Species level fusion

Element	FNA Species	FOC Species	Fused Result	Correct?	Why wrong
Apex	acute;	acute to obtuse;	acute to obtuse;	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Blade Length	(3.5-)4-5(-6)	2.5-5	25.0 to 60.0 mm	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Blade Width	1.5-2.5 cm	1-2 cm,	10.0 to 25.0 mm	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.

Table B.2 (cont.)

Element	FNA Species	FOC Species	Fused Result	Correct?	Why wrong
Blade Shape	elliptic to ovate-obovate,	lanceolate-ovate to narrowly elliptic,	elliptic, ovate, obovate, lanceolate	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Base	oblique,	oblique,	oblique,	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Margin	mostly singly serrate (some doubly serrate),	obtusely and irregularly simply serrate,	mostly singly serrate (some doubly serrate),	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.

Table B.2 (cont.)

Element	FNA Species	FOC Species	Fused Result	Correct?	Why wrong
Petiole	2-6(-8) mm, glabrous or sparsely pubescent with short hairs.	2-6 mm, pubescent;	2-6(-8) mm, glabrous or sparsely pubescent with short hairs.	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.
Surface	abaxially pale, glabrate, adaxially dark green, lustrous, glabrous;	abaxially pea green and pubescent when young, adaxially dark green, lustrous, and pubescent only on midvein,	abaxially pea green and pubescent when young, adaxially dark green, lustrous, and pubescent only on midvein,	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.
Vein	forking 5 or more times per side.	depressed; 10-15 on each side of midvein.	forking 5 or more times per side. depressed; 10-15 on each side of midvein.	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.

Section 4 - FNA Genus-Species Fusion

FNA_Genus: Leaves sometimes tardily deciduous; stipules falling early. Leaf blade ovate to obovate or elliptic, base usually oblique, sometimes cordate or rounded to cuneate, margins serrate to doubly serrate; venation pinnate.
FNA_Species: Leaves: petiole 2-6(-8) mm, glabrous or sparsely pubescent with short hairs. Leaf blade elliptic to ovate-obovate, (3.5-)4-5(-6) x 1.5-2.5 cm, base oblique, margins mostly singly serrate (some doubly serrate), apex acute; surfaces abaxially pale, glabrate, adaxially dark green, lustrous, glabrous; lateral veins forking 5 or more times per side.

Table B.3 FNA genus species fusion

Element	FNA genus	FNA species	Fused Result	Correct?	Why wrong
Base	usually oblique, sometimes cordate or rounded to cuneate,	oblique,	oblique,	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.
Blade Shape	ovate to obovate or elliptic,	elliptic to ovate-obovate,	elliptic, ovate, obovate	○ Agree ○ Not sure ○ Disagree	○ Missing information ○ Duplicate information ○ Wrong information ○ Other, possible reasons: _____ The correct result should be _____.

Table B.3 (cont.)

Element	FNA Species	FOC Species	Fused Result	Correct?	Why wrong
Margin	serrate to doubly serrate;	mostly singly serrate (some doubly serrate),	mostly singly serrate (some doubly serrate),	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.
Vein	pinnate.	forking 5 or more times per side.	forking 5 or more times per side.	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.

Section 5: FOC Genus-Species Fusion

FOC_Genus: Leaves distichous, blade base ¬± oblique, margin doubly or simply serrate; venation pinnate; secondary veins extending to margin, each ending in a tooth.
FOC_Species: Petiole 2-6 mm, pubescent; leaf blade lanceolate-ovate to narrowly elliptic, lamina on two sides of midvein unequal in length and width, 2.5-5 x 1-2 cm, thick, abaxially pea green and pubescent when young, adaxially dark green, lustrous, and pubescent only on midvein, base oblique, margin obtusely and irregularly simply serrate, apex acute to obtuse; midvein depressed; secondary veins 10-15 on each side of midvein.

Table B.4 FOC genus species fusion

Element	FOC genus	FOC species	Fused Result	Correct?	Why wrong
Base	oblique,	oblique,	oblique,	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.
Margin	doubly or simply serrate;	obtusely and irregularly simply serrate,	obtusely and irregularly simply serrate,	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.
Vein	pinnate; secondary veins extending to margin, each ending in a tooth.	depressed; 10-15 on each side of midvein.	pinnate; secondary veins extending to margin, each ending in a tooth. depressed; 10-15 on each side of midvein.	o Agree o Not sure o Disagree	o Missing information o Duplicate information o Wrong information o Other, possible reasons: _____ The correct result should be _____.

End of Evaluation