X-tract: Structure extraction from botanical textual descriptions

Rocío Abascal and J. Alfredo Sánchez
CENTIA[1]
Laboratory of Interactive and Cooperative Technologies
Universidad de las Américas – Puebla
Cholula, Pue. 72820 México
{abascal,alfredo}@mail.udlap.mx

Abstract

Most available information today, both from printed books and digital repositories, is in the form of free-format texts. The task of retrieving information from these ever-growing repositories has become a challenge for information retrieval (IR) researchers. In some fields, such as Botany and Taxonomy, textual descriptions observe a set of rules and use a relatively limited vocabulary. This makes botanical textual descriptions an interesting area to explore IR techniques for finding structure and facilitating semantic analysis.

This paper presents X-tract, a solution to the problem of text analysis and structure extraction in a specific application domain, namely floristic morphologic descriptions. The solution demonstrates the potential of using a grammar in the determination of information structure in a botanical digital library. We have developed a prototype based on this approach in which given an HTML or plain text, X-tract analyzes it and presents results to the user so he or she can verify the proposed structure before updating the database. This transformation is useful also in the process of storing morphologic descriptions in a database with a preestablished format. The solution is implemented in the context of the Floristic Digital Library (FDL), a large digital library project comprising a wide variety of botanical documents, formats and services.

Subject areas: information extraction, X-tract, botanical digital libraries, FDL

1 INTRODUCTION

Digital libraries continue to perform important functions such as collecting, organizing, presenting and finding information. They also extend the services that are provided by conventional libraries by taking advantage of the digital media [Lesk 1997].

One of the digital libraries currently being constructed, the Floristic Digital Library (FDL), is a virtual distributed space comprising botanical information and a variety of services offered to users to facilitate the use and extension of knowledge about plants [Schnase et al 1997]. Several international research and development projects financed by the National Scientific Foundation (NSF), like the Flora of North America (FNA), the Flora of China (FOC) and the Flora Mesoamericana (FM) participate in the FDL.

The main objective of this project is to create a digital library with information about plants from various geographical areas. For example, FNA manages information of approximately 20,000 species of vascular plants and bryophytes of North America north of Mexico [Schnase et al. 1997]. This library will contain textual documents, maps, illustrations and will provide services for the general public and for over 800 scientists who are contributing to this project.

One of the major problems faced by projects such as FNA and FOC relates to the fact that most of the information managed does not follow any specific format. However, botanical descriptions do regularly adhere to generally accepted rules and are based on a relatively limited vocabulary. The FDL is developing an object-relational model to store botanical descriptions. We therefore need to extract information that is available in non-structured documents so that it can be incorporated into the FDL's database. Among other resulting benefits, on-line information can be presented in a uniform format, and information can be produced in many formats for its distribution in paper or via web.

1.1 The problem of information extraction

The collections maintained by a library represent the individual efforts of thousands of authors, working together and separately over hundreds or even thousands of years and using a tremendous range of composition tools to capture their thoughts [Furuta 1994]. The proliferation of on-line text motivates most current work in text interpretation. Although massive volumes of information are available at low cost in digital free text form, people cannot read and digest the information any faster than before; in fact, for the most part they can digest even less.

Information extraction (IE) systems analyze unrestricted text in order to extract specific types of information [Lehnert 1996]. IE systems do not attempt to understand all of the text in all input documents, but they do analyze those portions of each document that contain relevant information. Relevance is determined by pre-defined domain guidelines which must specify, as accurately as possible, exactly what types of information the system is expected to find. The problem of extracting information from data is not addressed by simply developing better classification schemes, organizing data collections using newer and better database schemata, nor simply making the data accessible to the entire world by

[1] Center for Research in Information and Automation Technologies

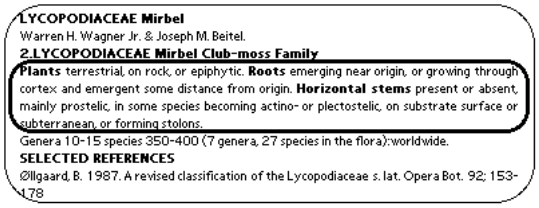
quickly transporting it across the evolving computer networks or data highways [Springer and Patrick 1994]. Filters are needed that can derive information or knowledge that can be extracted and analyzed from the massive collections of data stored in digital form. In this paper we demonstrate the importance of using a filter to analyze only the useful information.

2. X-TRACT: A HEURISTIC METHOD FOR STRUCTURE EXTRACTION

As mentioned previously, we have developed an approach to extract structure from botanical descriptions in the context of the Floristic Digital Library.

Biologists and other researchers write botanical descriptions referred to as "morphologic descriptions", which are included in "taxonomic treatments". Taxonomic treatments are comprehensive documents that include general discussions about plants, their toxicity, bibliographic references which refer to publications dealing with the plants, names of researchers and other information. Figure 1 is an example of a taxonomic treatment highlighting the morphologic description, as it appears (expanded) in FNA Volume 2 [Morin et al 1993]. Taxonomists consider the essence of a taxonomic description to be a list of properties possessed by a taxon [Taylor 1994].

Every word in bold type in the morphologic description is called a *structure*. Every structure may have a number of *characteristics* and characteristics may take one or more *values*.



LYCOPODIACEAE Mirbel

Warren H. Wagner Jr. & Joseph M. Beitel.

2.LYCOPODIACEAE Mirbel Club-moss Family

Plants terrestrial, on rock, or epiphytic. **Roots** emerging near origin, or growing through cortex and emergent some distance from origin. **Horizontal stems** present or absent, mainly prostelic, in some species becoming actino- or plectostelic, on substrate surface or subterranean, or forming stolons.

Genera 10-15 species 350-400 (7 genera, 27 species in the flora):worldwide.

SELECTED REFERENCES

Øllgaard, B. 1987. A revised classification of the Lycopodiaceae s. lat. Opera Bot. 92; 153-178

Figure 1. Example of a taxonomic treatment

2.1 Extraction heuristics

In order to illustrate our heuristic approach, we illustrate first how the morphological description in Figure 1 can be analyzed.

For example *Plants* is a structure that has the characteristic *location* and the value *terrestrial*. In this case, the characteristic "location" is implicitly understood by the reader. The manual process starts with the separation of the morphological description into sentences. Every sentence is analyzed by looking up to the first word. Almost every sentence starts with a structure. Then every word is looked up in a glossary to find the name of the characteristic or whether it is a value. Then the analyzer builds a table with the structures, characteristics and values to organize and classify the data. The problem arises when there is a substructure. Substructures are structures that are part of another structure. We note there are two types of substructures:

1. Substructures appearing in the same sentence.
2. Substructures appearing in separate sentences.

There can be substructures within other substructures, which adds complexity to the analysis. One of the main problems in structure determination occurs when characteristics or substructures are not explicitly stated. For example, in "Leaves 2.5--6cm, 3mm, 2-ranked to spiraled,...", 2.5 to 6 cm may refer to length or to width. A knowledgeable user will know when a description is talking about *width* or *length*.

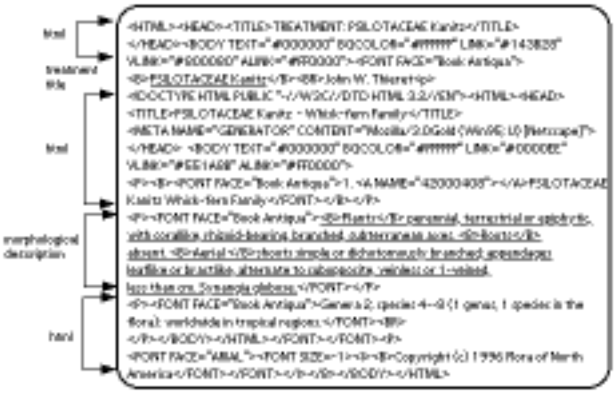
After the separation of the morphological description into sentences, and the look up of every word in the glossary, the manual process continues with the construction of a large table that has a column named structure, other substructure, characteristic and values. Every word is located in this table and then we find the relation of every word with the others. For example if we have "Plants terrestrial.", Plants will be the parent of terrestrial, and terrestrial will have the characteristic habit. The table for the example: "Plants terrestrial" will be like:

Structure	Substructure	Characteristic	Value
Plants			
		habit	terrestrial

The manual process ends with the construction of this table. For X-tract the process ends with the update of the database. In this case we have to consider the relationship of the words to assign a number and to organize it as a tree and save it in the FDL's database.

2.2 A grammar for morphologic descriptions

The grammar used by X-tract is based on the research of the current morphologic descriptions found in the on-line page of FNA. In general, this descriptions are of the form shown in Figure 2. In this figure, every part of the description is analyzed to comprehend the applied grammar.

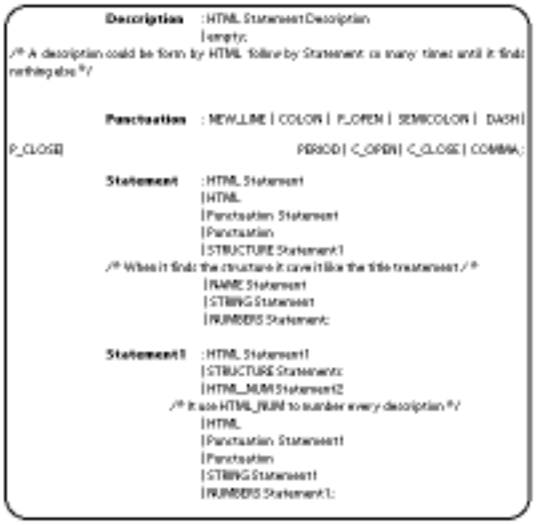


```
<HTML><HEAD><TITLE>TREATMENT: PSILOTACEAE Kanitz</TITLE>
</HEAD><BODY TEXT="#000000" BGCOLOR="#FFFFFF" LINK="#143828"
VLINK="#800080" ALINK="#FF0000"><FONT FACE="Book Antiqua">
<B>PSILOTACEAE Kanitz</B><BR>John W. Thieret<p>
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2//EN"><HTML><HEAD>
<TITLE>PSILOTACEAE Kanitz - Whisk-fern Family</TITLE>
<META NAME="GENERATOR" CONTENT="Mozilla/2.0Gold (Win95; I) [Netscape]">
</HEAD>  <BODY TEXT="#000000" BGCOLOR="#FFFFFF" LINK="#0000EE"
VLINK="#551A8B" ALINK="#FF0000">
<P><B><FONT FACE="Book Antiqua">1. <A NAME="42000408"></A>PSILOTACEAE
Kanitz Whisk-fern Family</FONT></B></P>
<P><FONT FACE="Book Antiqua"><B>Plants</B> perennial, terrestrial or epiphytic,
with corallike, rhizoid-bearing, branched, subterranean axes. <B>Roots</B>
absent. <B>Aerial </B>shoots simple or dichotomously branched; appendages
leaflike or bractlike, alternate to subopposite, veinless or 1-veined,
less than cm. Synangia globose.</FONT></P>
<P><FONT FACE="Book Antiqua">Genera 2, species 4--9 (1 genus, 1 species in the
flora): worldwide in tropical regions.</FONT><BR>
</P></BODY></HTML></FONT></FONT><P>
<FONT FACE="ARIAL"><FONT SIZE=-1><I><B>Copyright (c) 1996 Flora of North
America</FONT></FONT></I></B></BODY></HTML>
```

Figure 2. Example of taxonomic treatment in HTML

In the taxonomic treatment we have labeled *html* to all the irrelevant information to purposes of our analysis. In this example we have underlined the treatment title and the morphological description which we are going to analyze later. This way the taxonomic treatment is formed by the *html* label followed by the treatment title, the *html* label again and then the first description. A taxonomic treatment could have more than one description, and in this case the next description is next to the last *html* label, and so on.

Since we have to find the treatment title first, when we find the first "<B>" label we know this is the title treatment, and we save it; the next "<B>" label corresponds to the morphologic description. The complete grammar is in Figure 3. The analysis starts with a "Statement" block, which may include an HTML code followed by any punctuation sign. We then go to the next block after finding a structure in the "Statement" area. "Statement1" is formed like "Statement" but with the difference that if we find a structure, it corresponds to the morphologic description and we will start to save everything to analyze it afterwards.



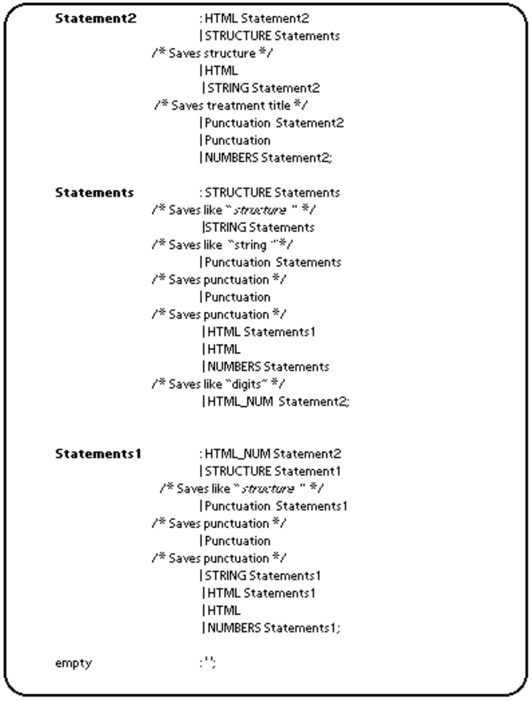
```
                Description     :HTML Statement Description
                                |empty;
/* A description could be form by HTML follow by Statement as many times until it finds
nothing else */


                Punctuation     : NEW_LINE | COLON | P_OPEN | SEMICOLON |  DASH|
P_CLOSE|                                   PERIOD | C_OPEN | C_CLOSE | COMMA;

                Statement       :HTML Statement
                                |HTML
                                |Punctuation Statement
                                |Punctuation
                                |STRUCTURE Statement1
                /* When it finds the structure it save it like the title treatement / *
                                 |NAME Statement
                                 |STRING Statement
                                 |NUMBERS Statement;

                Statement1      :HTML Statement1
                                |STRUCTURE Statements
                                |HTML_NUM Statement2
                           /* It use HTML_NUM to number every description */
                                |HTML
                                |Punctuation Statement1
                                |Punctuation
                                |STRING Statement1
                                |NUMBERS Statement1;
```

Figure 3. Complete grammar



```
Statement2              :HTML Statement2
                        |STRUCTURE Statements
                /* Saves structure */
                        |HTML
                         |STRING Statement2
                 /* Saves treatment title */
                        |Punctuation  Statement2
                        |Punctuation
                        |NUMBERS Statement2;

Statements              :STRUCTURE Statements
                /* Saves like " structure " */
                         |STRING Statements
                /* Saves like "string "*/
                        |Punctuation  Statements
                /* Saves punctuation */
                        |Punctuation
                /* Saves punctuation */
                         |HTML Statements1
                         |HTML
                         |NUMBERS Statements
                /* Saves like "digits" */
                         |HTML_NUM  Statement2;


Statements1             :HTML_NUM Statement2
                        |STRUCTURE Statement1
                  /* Saves like " structure " */
                        |Punctuation  Statements1
                /* Saves punctuation */
                        |Punctuation
                /* Saves punctuation */
                        |STRING Statements1
                         |HTML Statements1
                         |HTML
                         |NUMBERS Statements1;

empty                   :'';
```

Figure 3. (Continued) complete grammar

2.3 The X-tract approach

Jerry Hobbs [Hobbs 1994] defines an information extraction system as a cascade of transducers or modules that at each step add structure and often lose information, hopefully irrelevant, by applying rules that are acquired manually and/or automatically. X-tract includes modules for the grammatical analysis of a given text, the automatic extraction of the morphologic description located in the text, the use of glossaries to classify every word located in the description as a structure, characteristic or value and the update of the database. These modules are a preparser, a parser , a filter, a semantic interpreter and a structure resolution. Each of these modules is described next.

1. Preparser. This module takes a sequence of lexical items and attempts to identify small structures or small parts of the text. This way for example, we know if a word is a value or a structure.

2. Parser. The input for this module is a sequence of lexical items and phrases and its output is a set of structures that are part of a bigger structure. This analysis is done by using a grammar constructed for syntactic analysis, as described in section 2.2. The grammar was constructed by analyzing FNA descriptions. The grammar is also useful in finding the morphologic description within a given text. We also use a glossary to identify to what characteristic refers each word.

3. Filter. This module turns the sequence of sentences into smaller pieces by dropping irrelevant information. In this case the input text is composed by the title or name of the taxon, the name/s of the author/s and by a paragraph we named "html" label (see grammar in section 2.2). In this "html" paragraph we will start saving every word to be analyzed later.

4. Semantic Interpreter. This module generates a semantic structure from parse tree segments and with these segments we try to identify the complete document. The morphologic descriptions used accepted rules that are based in a vocabulary limited. The format and the vocabulary used can be consider like a sublanguage [Kittredge 1987], useful for the semantic analysis.

5. Structure Resolution. This module turns a tree-like structure into a network-like structure by identifying different descriptions of the same entity in different parts of text. Some given texts could have more than one description and X-tract can analyze them. Figure 4 shows the relationships among the components of X-tract.

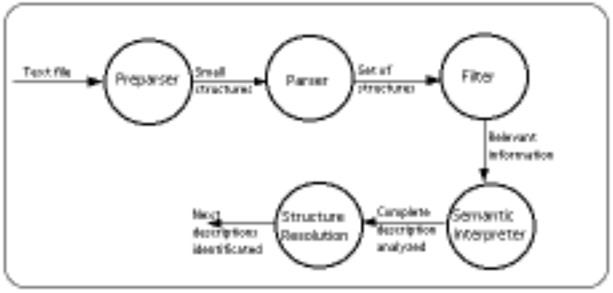


Figure 4. Relationships among X-tract components

3. X-TRACT PROTOTYPE AND EVALUATION

We have implemented a working prototype that demonstrates the viability of our approach for extracting structure from botanical descriptions.

X-tract is a prototype that uses a heuristic method to extract the attributes and the structure of morphologic descriptions written in free ASCII format or in HTML. The main objective is to facilitate the processes of entering new morphologic descriptions, and verifying the existing descriptions for format and parallelism. Two versions of X-tract have been developed, one using IIlustra DBMS and another using Informix Universal Server DBMS. The input to X-tract is an HTML text that contains a morphologic description with structures that are generally in bold typeface. For example, the "Plants" structure would be in the form "<B>Plants</B>". For the construction of the preparser, also named "scanner", we use lex . This allows programmers to create their own definitions and use this preparser to identify and label every part of the document. Also we use yacc to perform the syntactic analysis of taxonomic treatments. C is used as the main host language.

The prototype implementation is based in programs that use CGI (Common Gateway Interface). With them we manage the interaction between the user and Illustra and IUS using HTML forms.

3.1 X-tract operation

X-tract provides an interface for authors and the people in charge of the edition and revision of taxonomic treatments. Users must have a login and a password to use X-tract.

X-tract offers two options for introducing textual morphologic descriptions to FDL's database. One works by providing the name of the HTML file containing the description. The other alternative allows the user to introduce the description directly in plain text. Figure 5 shows the page that appears after the login and password are accepted by X-tract. In this page the user can input the name of any HTML text file, located in the local machine. Alternatively, the user may type the text in the input area at the bottom.

The process that X-tract follows after receiving the file name is the following:

· X-tract looks for the file in the local directory if the user types the name of the file name or looks for the file in the given path; if the file does not exist an error message is produced.

· If the file exists, then X-tract uses the grammar to find out where the morphologic description is located in the text. In this case it saves every part of the morphologic description to subsequently analyze it.

· Every string is saved according to its type. This means that the preparser recognizes when a word refers to a structure, a number or something else.

· X-tract uses a glossary to find out whether a string is a value or a structure. In case the string is a value, it assigns the corresponding characteristic. For example, if the value is "green", its characteristic is "color".

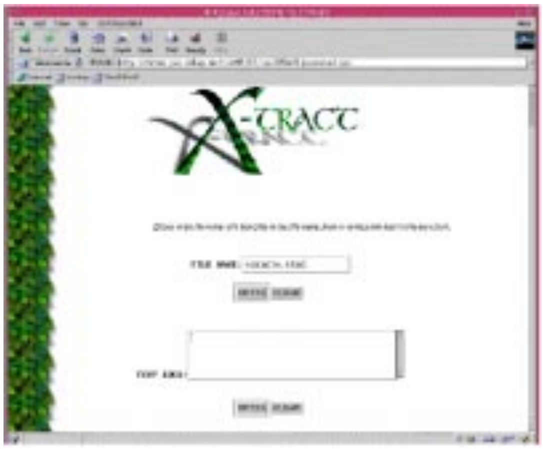
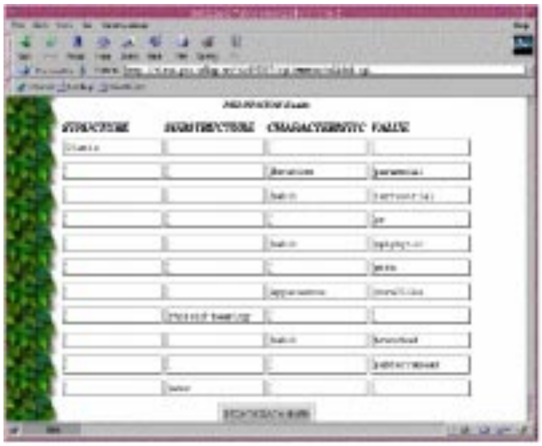


Figure 5. Options offer by X-tract

· X-tract creates a form to organize the document analyzed into structures, substructures, characteristics and values as shown in Figure 6. At this time the user can decide whether to update the database or to modify (and improve) the forms given by X-tract.



STRUCTURE	SUBSTRUCTURE	CHARACTERISTIC	VALUE
Plants			
		duration	perennial
		habit	terrestrial
			or
		habit	epiphytic
			[illegible]
		appearance	[illegible]
	[illegible] bearing		
		habit	branched
			[illegible]
	[illegible]		

Figure 6. Table created by X-tract

3.2 Xtract benefits

Currently, editors of taxonomic treatments use an electronic spreadsheet to construct enormous forms where they put information descriptions. The analysis is complicated because of the number of descriptions saved in these forms. X-tract organizes the information in a database, offering the user various specialized ways to present and look for information by using a DBMS.

One of the main problems in the FDL digital library is to extract the information found in a taxonomic treatment. X-tract facilitates several activities including:

· authors can verify their taxonomic treatments from everywhere in the world, thanks to the access of X-tract via web,

· introduction of new treatments to analyze the morphologic descriptions found in them,

· syntax verification by the authors before updating the database,

· automatically updating the database.

4. RELATED WORK

In the research area of natural language processing, a major challenge is in the area of information extraction. Some of the projects analyzed below are from lexicography, translation and information retrieval.

- Hector: Its focus is computer-aided lexicography and its purpose is to compile dictionary entries using corpus evidence and to sense-tag the corpus lines that have been used to create the dictionary entries [Kavanagh 1995].

- Translator's Workbench: designed to provide machine support for translators in the form of an integrated set of software tools designed to eliminate some of the tedious work of translation [Kavanagh 1995].

- TACT (Text Analysis Computing Tools): designed to do text retrieval and analysis on literary works in any language that uses the roman alphabet or in classical Greek.

-Xtract: developed by Frank Smadja that attempts to find collocations in a text. A collocation is an arbitrary and recurrent word combination. The collocations are interesting for translators and is important also for knowledge engineers trying to do a conceptual analysis of a domain [Kavanagh 1995].

- The Text Analyzer: combine methods from computational linguistics and artificial intelligence to provide the users with a variety of options for finding information in documents, verifying the consistency of this information, performing word and conceptual analyses and other operations [Kavanagh 1995].

-DELTA format (DEscription Language for Taxonomy): flexible method for encoding taxonomic descriptions for computer processing. DELTA-format data can be used to produce natural-language descriptions, conventional and interactive keys, and cladistic and phonetic classifications [Taylor 1994].

-Terminator/NEMISIYS: electronic tool called "terminator" because it looks for terms, parses files and picks up whatever description characters they include [Diederich et al. 1987]. It works with a schema, this means working with a formal list of morphological characteristics and related information organized for use in a DBMS.

- Macros for MS Office:copies the descriptions of accepted taxa within a fully formatted FOC document and parses them into tab delimited fields so that they can be cut and pasted into a spreadsheet or database.

5. ONGOING AND FUTURE WORK

Since FDL is in construction we used only taxonomic treatments located in FNA electronic page (http://www.fna.org) to test our X-tract prototype. The main structures of a morphologic description are located in HTML code between "<B>" and "</B>". Initial tests with real users are encouraging as only minimal intervention is needed.

The grammar used in X-tract can be applied to analyze morphologic descriptions and not other parts of taxonomic treatments such as the names of the authors and the references. The X-tract prototype only updates the database containing information about the morphologic description. Future work includes the analysis of other parts of the taxonomic treatment. An ongoing project [Navarrete 1999] is using X-tract as the basis for the automatic translation of botanical descriptions.

Also the grammar of X-tract is used actually for a project of digital thesis where we want to have the thesis storage in a database. The grammar is used for example to identify where a paragraph corresponds to the introduction or refers to a table or a figure.

X-tract is part of a more general system to support authors of morphologic descriptions in the construction of the FDL repository. With its use we expect to expedite the construction of the digital library and also to contribute to the information extraction research.

REFERENCES

[Diederich et al 1987] Diederich , I. Ruhmann and M.May. 1987. KRITON: A knowledge-acquisition tool for expert systems. International Journal of Man-Machine Studies, 26, 1987.

[Fox et al 1993] Fox E., Hix D., Nowell L. 1993. Users, user interfaces, objects: Envision, a digital library. Journal of Am. Soc Inf. Sci. 44, 8 (Sept), 480-491.

[Fox 1995] Fox A. E. 1995. Digital Libraries. Communications of the ACM, 38, 4 (April) 24-25.

[Furuta 1994] Furuta R. 1994. Defining and Using Structure in Digital Documents. Proceedings of Digital Libraries'94. (DL'94, College Station, TX, June)

[Hobbs 1994] Hobbs R. J. 1994 Generic Information Extraction System. Artificial Intelligence Center SRI International.

[Kavanagh 1995] Kavanagh, J. 1995 The Text Analyzer: A Tool for Extracting Knowledge From Text. Master of Computer Science Thesis, project of the Language Analysis &Knowledge Engineering (LAKE) Group Department of Computer Science, University of Ottawa.

[Kittredge 1987] Kittredge, R. 1987. The significance of sublanguage for automatic translation. In Machine Translation: Theoretical and Methodological Issues; S. Nirenburg (Ed), Cambridge University Press, 59-67.

[Lenhert 1996] Lenhert, W. 1996. Information Extraction. Computer Science Departament at the University of Massachusets. Morgan Kaufmann Publishers, Inc.

[Lesk 1997] Lesk, M., 1997. Practical digital libraries: books, bytes and bucks. Morgan Kaufmann Publishers, Inc.

[Meyer et al 1992] Meyer I, Skuce D, Bowker L and Eck K. Towards a new generation of terminological re-sources: An experiment in building a terminological knowledge base. 13th International Conference on Computational Linguistics (COLING) Nantes.

[Morin 1993] Morin N. 1993 Flora of North America, vol 2. Oxford University Press.

[Navarrete 1999] Navarrete D. 1999. Traductor de Descripciones Morfológicas de Especies Vegetales en el Contexto de una Biblioteca Digital Botánica. Tesis de Licenciatura, Departamento de Ingeniería en Sistemas Computacionales, Universidad de las Américas-Puebla. Cholula, Puebla, México.

[Sánchez 1994] Sánchez J. A. 1994. User agents in the interface to digital libraries. Proceedings of Digital Libraries' 94. (DL'94, College Station, TX, June) 217-218.

[Schnase et al 1997] Schnase, J. L., Kama, D.L.Tomlinson,K.L., Sánchez, J. A., Cunnius, E.L. y Morin, N. R. 1997: The Flora of North America Digital Library: a case study in biodiversity database publishing; Journal of Network and Computer Applications, 21, 20; enero.

[Springer and Patrick 1994] Springer K. G and Patrick B. T. 1994. Translating Data to Knowledge in Digital Libraries. Proceedings of Digital Libraries'94. (DL'94, College Station, TX, June)

[Taylor 1994] Taylor A. 1994. Extracting Knowledge from Biological Descriptions. Department of Computer Science and Engineering. University of New South Wales. Sydney, Australia.

[Wiederhold 1995] Wiederhold, G. April, 1995. Digital Libraries, Value, and Productivity. Communications of the ACM, 38, 4, 85-86.