Capítulo 1. Introducción

Se han estado generando bibliotecas tradicionales a lo largo de cientos de años. Hoy en día es posible proveer nuevos servicios gracias a la colaboración de escritores, investigadores y especialistas en tecnología de información. Actualmente podemos hacer uso de la información, en su formato original, desde cualquier parte del mundo a través de la computadora. De este modo, el conocimiento puede ser accesado desde cualquier lugar y por cualquiera que tenga acceso a la red global de computadoras y comunicaciones, principalmente a través de World Wide Web.

Las bibliotecas digitales continúan llevando a cabo importantes funciones como recolección de información, organización, presentación y localización de información. Además extienden los servicios que actualmente proporcionan las bibliotecas convencionales haciendo uso de las ventajas que ofrece el medio digital [Lesk 1997].

El presente trabajo se ubica en el contexto de una biblioteca digital accesible vía web. El problema principal a tratar se refiere a la dificultad de extraer información que se encuentra en diversos formatos para organizarla en una base de datos.

A continuación se introducen a las bibliotecas digitales con el fin de tener un panorama de lo que son y sus objetivos principales.

1.1 Bibliotecas Digitales

Esta era y la que estamos construyendo, conocida como *ciberespacio* , la *era de la información* , la *supercarretera de la información* o el *interespacio* , por nombrar algunos términos populares, están soportadas por un sistema de red. Sin embargo, su esencia es la información. Información es lo que viaja en la red, lo que se presenta en distintos modos electrónicos, lo que se manipula en nuestras

computadoras, o lo que se ofrece a través de las bibliotecas digitales [Fox et al. 1993].

Con el crecimiento de Internet y la gran demanda de los usuarios por accesar información de manera rápida y eficiente, se ha generado el concepto de biblioteca digital. Una biblioteca digital se considera popularmente como una versión electrónica de una biblioteca convencional. Sin embargo, la sustitución del papel por documentos electrónicos nos lleva a tres grandes diferencias: acceso y almacenamiento de información en forma digital, comunicación directa con la biblioteca desde cualquier parte del mundo para obtener material, y obtención de una copia del material de la versión original [Wiederhold 1995].

Actualmente, el concepto de "biblioteca digital" suele dar una impresión distinta a cada lector. Para algunos significa nuevos métodos para almacenar la información y preservarla. Para otros significa nuevas técnicas para clasificar, catalogar, para interactuar con el usuario, mayor seguridad en sistemas electrónicos, y cambios dramáticos en lo organizacional [Fox 1995]. Para los profesionales en la computación, una biblioteca digital es simplemente una colección de servicios distribuidos de información; es un espacio distribuido de información interrelacionada [Slonim 1995].

Algunas de las ventajas que ofrece el uso de bibliotecas digitales, a diferencia de las bibliotecas convencionales son: uso de multimedios, hiper-referencias, mejor control de la información, trabajo académico cooperativo y remoto, ordenamientos múltiples, flujo más libre de información [Sánchez 1994]. Otras ventajas de las bibliotecas digitales incluyen un mejor uso de los recursos existentes, mayor protección de libros y documentos históricos contra el rápido deterioro y vandalismo, solución a problemas masivos de almacenamiento y por último un acceso instantáneo a cualquiera desde cualquier lugar del mundo.

Las bibliotecas digitales se enfrentan a problemas tales como el manejo de mucha información en distintos formatos, seguridad, actualizaciones, recuperación de información, búsquedas, manejo de la base de datos de la biblioteca, costos elevados en software y hardware, entre otros.

Este proyecto se desarrolla dentro de una de las bibliotecas digitales que actualmente se están construyendo: la Biblioteca Digital Florística. En la sección que sigue se explicará el propósito de su creación.

1.2 Biblioteca Digital Florística

Existe gran preocupación por parte de los investigadores en

biodiversidad acerca de la extinción de especies botánicas antes de que hayan podido estudiarse o incluso identificarse dadas las limitaciones de los métodos tradicionales de recolección y de intercambio de información. Las tecnologías de información y comunicación pueden apoyar actividades científicas para ofrecer lo que se conoce como bibliotecas digitales.

La Biblioteca Digital Florística (FDL) es un espacio virtual distribuido que comprende información botánica y una variedad de servicios ofrecidos a los usuarios para facilitar el uso y la extensión del conocimiento acerca de las plantas [Schnase et al. 1997]. La Biblioteca Digital Florística abre un espacio a investigadores y expertos en el área para compartir conocimiento y para tener información actualizada de manera accesible.

En la Biblioteca Digital Florística participan proyectos internacionales de investigación y desarrollo financiados por la Fundación Nacional para la Ciencia (NSF) como la Flora de Norte América (FNA), la Flora de China (FOC) y la Flora de Mesoamérica (FM) bajo la dirección del Centro de Informática Botánica (CBI) del Jardín Botánico de Missouri y con la participación del Laboratorio de Tecnologías Interactivas y Cooperativas (ICT) de la Universidad de las Américas Puebla (UDLAP).

FNA es un proyecto que tiene como objetivo tener documentos, mapas, ilustraciones y sobre todo la base de datos de la Flora de Norte América, la cual se refiere a aproximadamente 20,000 especies de plantas de Norteamérica al norte de México [Schnase et al. 1997]. FOC y FM persiguen objetivos similares para especies de China y de Mesoamérica respectivamente. FNA se comenzó en 1987 y se pretende que esté terminado alrededor del 2006 [Tranter 1994]. Aproximadamente se cuenta con la colaboración de 800 investigadores contribuyendo a este proyecto.

Para los proyectos de Flora de Norteamérica (FNA) y Flora de China (FOC) se ha manejado información de forma masiva, la cual se encuentra en formato libre. La Biblioteca Digital Florística (FDL) se encuentra en desarrollo y está basada en un modelo objeto-relacional. Es deseable que la información que ya se encuentra en documentos no estructurados sea incorporada a la base de datos de FDL. Esto conllevaría múltiples beneficios, entre los que cabe mencionar:

la información puede ser accesada en línea en un formato uniforme;

facilidad en búsquedas más precisas al utilizar mecanismos de consulta disponibles para una base de datos;

la información puede extraerse de varios formatos para su distribución en papel o de manera electrónica (como HTML);

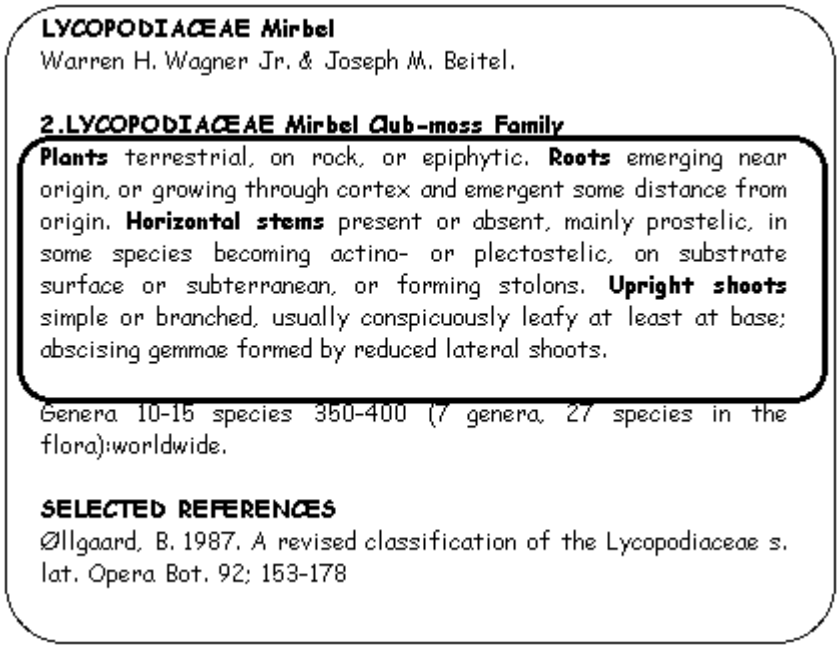
la validación y corrección de la información puede hacerse de manera más eficiente por los editores al utilizar una forma electrónica con campos donde el usuario puede corregir antes de actualizar la base de datos; y

es posible aplicar técnicas de traducción para interfaces multilingües a FDL.

La transformación de descripciones morfológicas en un formato libre a un formato estructurado presenta varios problemas, como se describe a continuación.

1.3 El problema de extraer información en FDL

Para la Biblioteca Digital Florística (FDL), los investigadores escriben descripciones de las especies botánicas llamadas "descripciones morfológicas" que se encuentran dentro de un "tratamiento taxonómico", el cual entre otras cosas contiene discusiones sobre su uso, toxicidad, referencias bibliográficas y los nombres de los investigadores. Los taxonomistas consideran como la esencia de una descripción taxonómica a una lista de propiedades que posee una taxonomía [Taylor 1994].



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. **Upright shoots** simple or branched, usually conspicuously leafy at least at base; abscising gemmae formed by reduced lateral shoots.

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

Figura 1. Ejemplo de un tratamiento taxonómico

La Figura 1 muestra un ejemplo de un tratamiento taxonómico el cual contiene una descripción morfológica (en el recuadro). Este tratamiento para su presentación ha sido condensado y se ha tomado del volumen 2 de FNA [Morin et al 1993]. Esta descripción utiliza un lenguaje natural restringido, en este caso el de los botánicos, lo cual nos lleva al estudio del lenguaje que comúnmente ocupan estos investigadores. Podemos observar que un tratamiento taxonómico, además de contener las características principales de la planta descrita, también maneja información sobre las referencias utilizadas

y sobre los autores de dicha descripción. Para el caso de FDL, nos interesa tener en la base de datos, por ejemplo, información como la bibliografía citada en cada tratamiento, los autores, las localidades y la descripción morfológica. Este proyecto se enfoca sólo a la extracción de la descripción morfológica.

La información de FDL se almacenará en un esquema objeto-relacional. Sin embargo, existen ya gran cantidad de descripciones que no se encuentran en esta forma, incluyendo algunas en FNA y FOC. Además de que existen descripciones archivadas en papel, las cuales no han sido almacenadas en la base de datos.

El proceso para obtener la información que nos interesa con el objetivo de almacenarla en una base de datos requiere de un proceso de análisis. Este análisis generalmente comienza con la identificación de términos, en este caso como el nombre, origen, tamaño, forma y color de cada planta o grupo de plantas. Para la Biblioteca Digital Florística requerimos de un método que nos permita identificar lenguaje natural restringido utilizado en las descripciones morfológicas.

1.4 Estructura en las Bibliotecas Digitales

La proliferación de textos en WWW ha motivado a muchos de los proyectos de extracción de información ya sea a una base de datos para poder llevar a cabo consultas. No obstante que la mayoría de los volúmenes de información se encuentran accesibles a bajos costos en texto de formato libre, la gente no puede leer y asimilar tanta información al mismo tiempo. De esta manera se requiere que la información se almacene en un formato estructurado, por ejemplo una base de datos relacional, o indexada sistemáticamente y ligada hacia otros textos en que contengan información relacionada. La mayoría de los procesos de extracción de información consisten en responder a preguntas a las que el usuario desea dar respuesta. Un diagrama de este proceso de extracción se muestra en la Figura 2 , donde se responden ciertas preguntas a partir de su búsqueda en un texto.

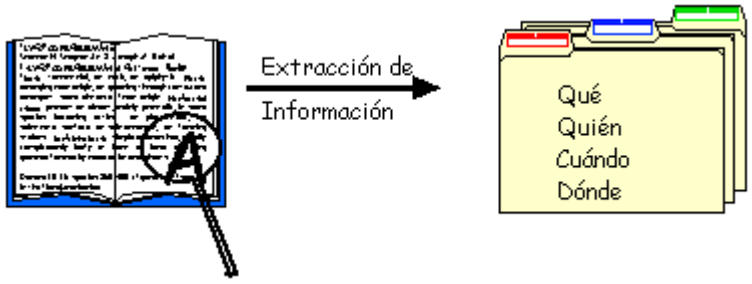


Figura 2. Proceso de extracción de información

Entender la estructura es un paso crítico en el proceso de desarrollo en el diseño de una biblioteca digital [Furuta 1994]. Entender la estructura implica analizar una gran cantidad de información para crear un método mediante el cual pueda extraerse la mayoría de la información. Cuando hablamos acerca de estructura nos referimos a la organización del texto.

La colección de una biblioteca representa esfuerzos de miles de autores, trabajando juntos y separados a través de muchos años y utilizando un amplio rango de herramientas para capturar sus pensamientos [Furuta 1994]. Existen distintas formas de representar lo que cada uno piensa, de esta manera se obtiene un trabajo con una organización en especial para cada autor.

La extracción de la estructura en la Biblioteca Digital Florística se basa en la identificación de las palabras contenidas en las descripciones morfológicas. Sin embargo, la utilización de las palabras puede variar de acuerdo a la descripción en la qué se está aplicando. Por ejemplo, la palabra "bundled" puede referirse a *arquitectura* o a *distribución* . Además, cada descripción morfológica contiene un tipo y número de características en especial. Es indudable que entonces, se complica la definición de una estructura o gramática que se aplique a todas las descripciones.

1.5 X-TRACT: Un Método heurístico de extracción de estructura

En este documento se presenta una alternativa de solución al proceso de extracción de información a partir de descripciones textuales de especies botánicas. Se plantea un sistema llamado X-tract que a partir de una descripción dada en HTML (HyperText Markup Language) extrae la información hacia la base de datos. Este sistema fue creado en base a las descripciones que se encuentran en HTML en el catalogo en línea de FNA.

Esta alternativa resulta ser un mecanismo menos tedioso y cansado para el usuario que un proceso manual en el que se debe estar buscando cada palabra en un glosario con el fin de saber a qué característica se refiere y obtener así el valor correspondiente. De esta manera, se agiliza el proceso manual de introducción de datos a la base de datos.

A continuación se proporcionan algunas definiciones y una descripción general de la solución propuesta.

Los investigadores escriben descripciones morfológicas que contienen términos que pueden ser características tales como *arquitectura* , *color* , *orientación* o *maduración* ; o bien puede tratarse de *estructuras* . Una estructura se refiere a una parte de la planta descrita

y es esta precisamente la que se describe mediante características como las antes mencionadas, además de que contiene uno o más valores. Finalmente, los valores pueden ser un adjetivo calificativo como *verde* , *flexible* , *plano* , *delgado* , entre otros; o bien puede ser un valor numérico. Basándonos en esto se desarrolló X-tract, un método heurístico para extraer los atributos e inferir la estructura de descripciones morfológicas expresadas en formato libre.

1.6 Objetivos del proyecto

Los objetivos del proyecto se definieron de la siguiente manera:

Creación de una gramática para poder analizar las descripciones localizadas dentro de un tratamiento taxonómico cuando se trata de un texto en formato HTML como las descripciones existentes en el catalogo en línea de FNA y cuando se trata de formato libre.

Investigación de técnicas para el análisis de textos. Existen actualmente analizadores de texto que nos permiten una variedad de opciones en la búsqueda de información dentro de un documento.

Construcción de un sistema capaz de identificar el conocimiento descrito en descripciones morfológicas para guardarlo en una base de datos. En este caso, algunas palabras de importancia para la descripción aparecerán con ”<” Y ”>” como si se tratara de HTML, entonces debe de identificarse cuando estas palabras deben de almacenarse para su análisis posterior o si se trata de texto irrelevante. Por ejemplo: *<B>Plants</B>* .

Desarrollo de programas para inferir la estructura implícita en descripciones morfológicas textuales. Es decir, el análisis de cómo está organizada generalmente un texto dado.

Una vez que se ha analizado la descripción morfológica su integración a la base de datos. Cada palabra debe de guardarse de acuerdo a su tipo y a su jerarquía dentro de la descripción.

Desarrollo de interfaces de usuario para facilitar el análisis de descripciones morfológicas. Esto con el fin de ayudar al usuario experto en la verificación del análisis antes de actualizar la base de datos.

1.7 Organización del documento

Las secciones restantes de este documento están organizadas de la siguiente manera: En el capítulo 2 se introduce a la investigación realizada de los sistemas y programas que actualmente existen enfocados al área de extracción de información, mostrando el resultado de dicho análisis. En el capítulo 3 se describe el diseño conceptual del sistema desarrollado en este trabajo. En el capítulo 4 se describe un prototipo del sistema, se hace una evaluación en términos de los objetivos planteados al inicio del proyecto y se evalúa su funcionalidad en comparación con otros sistemas. Finalmente, el capítulo 5 provee una síntesis del trabajo realizado y conclusiones derivadas del desarrollo, así como una descripción del trabajo a realizar en el futuro.

Abascal Mena, M. R. 1998. Extracción de estructura a partir de descripciones textuales botánicas. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.