UNIVERSIDAD DE GRANADA

E.T.S. INGENIERIA INFORMATICA

Departamento de Ciencias de la Computación e Inteligencia Artificial

MODELOS DE REPRESENTACIÓN DEL CONOCIMIENTO PARA LA IDENTIFICACIÓN TAXONÓMICA Y APLICACIONES

TESIS DOCTORAL

Eva Lucrecia Gibaja Galindo

Granada, septiembre de 2004

Editor: Editorial de la Universidad de Granada
Autor: Eva Lucrecia Gibaja Galindo
D.L.: Gr. 1712 - 2005
ISBN: 84-338-3597-1

MODELOS DE REPRESENTACIÓN DEL CONOCIMIENTO PARA LA IDENTIFICACIÓN TAXONÓMICA Y APLICACIONES

Eva Lucrecia Gibaja Galindo

TESIS DOCTORAL

Directores: Dr. D. Waldo Fajardo Contreras
Dra. Dña. Carmen Quesada Ochoa

Septiembre de 2004

DPTO. CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL. E.T.S.
INGENIERÍA INFORMÁTICA. UNIVERSIDAD DE GRANADA

La memoria *Modelos de Representación del Conocimiento para la Identificación Taxonómica y Aplicaciones*, que presenta Eva Lucrecia Gibaja Galindo, para optar al grado de Doctora en Informática, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, bajo la dirección del Profesor Dr. D. Waldo Fajardo Contreras, Profesor titular de Universidad y la Dra. Dña. Carmen Quesada Ochoa, Conservadora del Herbario de la universidad de Granada.

Granada, Septiembre de 2004.

Fmdo.: Dr. D. Waldo Fajardo Conteras Fado. Dra. Dña. Carmen Quesada Ochoa

Fmdo.: Dña. Eva Lucrecia Gibaja Galindo

A mis padres, Miguel y Lucrecia, por su esfuerzo y dedicación durante todos estos años.

MODELOS DE REPRESENTACIÓN DEL CONOCIMIENTO PARA LA IDENTIFICACIÓN TAXONÓMICA Y APLICACIONES

Eva Lucrecia Gibaja Galindo

Agradecimientos

Quiero manifestar mi agradecimiento a mis directores, Carmen y Waldo, Waldo y Carmen, por su interés, apoyo, e implicación en mi trabajo y, sobre todo, por su paciencia durante este tiempo. Porque su apuesta incondicional por mi y su tesón ha sido el impulso decisivo para este trabajo. A ellos les debo no solo la madurez científica, sino personal. Muchas gracias por acompañarme en este camino.

Así mismo, les debo mucho a las personas con las que he desarrollado mi trabajo en el seno del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, en particular a los integrantes del grupo de investigación ARAI y a mis compañeros de Mecenas que durante casi dos años han sido como una familia para mí.

Tampoco me puedo olvidar del equipo-H, porque gracias a ellos he descubierto otra forma de admirar la belleza de lo que me rodea y el valor de las cosas pequeñas. Por último, tengo que recordar a mis amigos, especialmente a Jose, por su comprensión y paciencia, que ha sido mucha.

Índice General.

Introducción general.

El término Inteligencia Artificial se acuña en el Dartmouth College , en la pequeña ciudad universitaria de Hannover. En este lugar, en 1956, se reunieron diez científicos representantes de diferentes disciplinas, como Matemáticas, Neurología, Sicología e Ingeniería Eléctrica. La base común que conectaba a un grupo tan disperso era que todos ellos utilizaban los ordenadores para sus investigaciones. Estas investigaciones perseguían, desde diferentes puntos de vista, simular distintos aspectos de la inteligencia humana. En esta conferencia situamos el nacimiento de la Inteligencia Artificial[1].

Uno de los objetivos de aquella primitiva Inteligencia Artificial era desarrollar un programa capaz de imitar todas las habilidades humanas para la resolución de problemas. Pero las expectativas no se hicieron realidad y, salvo para los casos más simples, los programas genéricos de Inteligencia Artificial se vieron desbordados por la complejidad de los problemas. Después de esta etapa de excesivo optimismo, la Inteligencia Artificial decayó y la investigación quedó

[1] Esta denominación fue sugerida por John McCarthy, uno de los organizadores de la conferencia, y ha llegado a asociarse irrevocablemente con esta área de investigación.

relegada a unas pocas universidades y grupos de investigación respaldados por subvenciones gubernamentales.

En la década de los ochenta, la Inteligencia Artificial vuelve a resurgir. Por primera vez en su historia, el hardware disponible tenía memoria suficiente para soportar los sistemas *inteligentes*. Además, estos empezaron a demostrar resultados prácticos, como Prospector [Duda *et al.*, 1978] o Mycin [Shortlife, 1976; Buchanan & Shortlife, 1984], y en algunos casos pudieron comercializarse, llegando a ser hoy en día una industria que maneja miles de millones de euros.

Uno de los principales factores que han intervenido en esta revolución, ha sido la importancia otorgada al *conocimiento* en el desarrollo de sistemas inteligentes. Así, uno de los principales exponentes de esta tecnología son los sistemas expertos que, de forma muy general, pueden definirse como sistemas que exhiben un nivel de productividad análogo al de un experto en un área concreta.

Si atendemos a la definición de inteligencia dada por Lenat y Feigenbaum, como la capacidad de encontrar rápidamente una solución adecuada en lo que, *a priori*, es un inmenso espacio de alternativas, vemos que no nos basta con tener algoritmos de búsqueda más o menos eficientes, sino que tenemos que recurrir al conocimiento sobre el dominio en cuestión. De hecho, suele enunciarse cómo el *Principio del Conocimiento* el siguiente: *Un sistema exhibe un comportamiento inteligente, debido principalmente al conocimiento que puede manejar: conceptos, hechos, representaciones, métodos, modelos, metáforas y heurísticas en su dominio de actuación*. Así pues, debe llegarse a un compromiso entre conocimiento y búsqueda. La búsqueda es costosa, y, aún en dominios muy restringidos, es prácticamente imposible tener almacenados todos los posibles casos y situaciones con las que nos podemos encontrar.

El *Principio del Conocimiento*, que acabamos de introducir, se completa con una intervención de Feigenbaum en la *International Joint Conference on Artificial Intelligence* de 1977 en la que enunció lo siguiente: "*El conocimiento experto proporciona la clave de la alta productividad, mientras que las representaciones del conocimiento y los esquemas de inferencia proveen de los*

mecanismos necesarios para su uso". Con esta idea como base se ha dado el nombre de Sistema Basado en el Conocimiento (*Knowledge-Based System*) a un conjunto de recursos especializados en la resolución de problemas acerca de un de cierto dominio de discurso que emplean de modo masivo conocimiento específico acerca de dicho dominio.

Los sistemas basados en el conocimiento difieren de forma importante de los sistemas convencionales de proceso de datos. Los principales rasgos que los caracterizan son: representación simbólica del conocimiento acerca del dominio, inferencia simbólica sobre ese conocimiento y búsqueda heurística como herramienta para la estrategia de control. De aquí que se haya acuñado el nombre de Ingeniería de Conocimiento para designar la rama dedicada al análisis, diseño y construcción de sistemas basados en el conocimiento.

La tecnología de sistemas basados en el conocimiento se ha convertido en una línea de investigación muy fructífera que ha prosperado en una gran variedad de campos de aplicación que nos llevan desde el asesoramiento en banca [Matsatsinis *et al.*, 1997; Vranes *et al.*, 1996] al diagnóstico médico [Cabrero-Carnosa *et al.*, 2003; Robertson & Noren, 2001].

A medida que la ciencia evoluciona, aparecen nuevos focos de interés, y con ellos nuevas posibilidades y ámbitos de aplicación para los sistemas basados en el conocimiento. Así, en la actualidad tienen un alto impacto los sistemas expertos diseñados para el diagnóstico y tratamiento de plagas en cultivos y granjas de animales [Mahaman *et al.*, 2002; Mahaman *et al.*, 2003; Daoliang *et al.*, 2002] y el asesoramiento en la gestión sostenible de recursos naturales [Li *et al.*, 2001].

De forma más específica, en el área de la Biología, las nuevas necesidades de los investigadores les hace interesarse por tecnologías para grandes bases de datos, modelos de predicción y simulación, extracción de conocimiento, etc. No es posible un correcto aprovechamiento de las herramientas tecnológicas sin un trabajo multidisciplinar de biólogos e informáticos que acompase las técnicas para

el desarrollo de programas inteligentes, escalables y robustos con las particularidades de los problemas biológicos.

Uno de los proyectos internacionales más ambiciosos es la elaboración del catálogo de todos los seres vivos que se conocen. No es un trabajo trivial si tenemos en cuenta que los casi dos millones de especies catalogadas actualmente suponen tan solo entre el 10% y el 30% de las que se estima alberga la Tierra. Incluso en un grupo tan estudiado como el de las plantas con flores, se describen más de 2000 especies nuevas cada año. Para hacer este inventario sobre la riqueza biológica del planeta hay que identificar correctamente a cada uno de los individuos (vegetales o animales) que componen las muestras de estudio.

En Botánica, cada planta se considera que pertenece a una serie de unidades de clasificación (que se denominan grupos taxonómicos o *taxa*) subordinadas unas a otras, cada una de las cuales tiene un nombre determinado. La categoría básica de estas unidades de clasificación es la especie. El género es la categoría superior de especie y sirve para agrupar a todas las especies que son suficientemente próximas entre sí desde un punto de vista evolutivo, especies que por lo general son además bastante parecidas en algunos de sus caracteres. La otra gran unidad de clasificación botánica es la familia, que de forma análoga a como lo hace el género con las especies, agrupa a un cierto número de géneros emparentados evolutivamente. Las familias se agrupan en unidades superiores que son, en secuencia ascendente: órdenes, clases, divisiones o *filum* y reinos [López, 2001].

Para abordar el estudio de un grupo taxonómico, debemos fijar previamente la localidad de referencia. El siguiente paso, es determinar el grupo o grupos taxonómicos de estudio y establecer una terminología de referencia. Para nuestros estudios nos hemos centrado en Gimnospermas (algunas de ellas cultivadas) presentes de en la Península Ibérica[2]. También es importante, definir una terminología adecuada, precisa y adaptada al perfil de los usuarios de esta

[2] La Península Ibérica reúne una serie de condiciones que han hecho de su flora una de las más ricas y variadas de toda Europa. Se calcula que incluye entre 7500 y 8000 especies distintas de plantas vasculares, cerca del 15% exclusivas (endemismos) [Castroviejo, 2002].

información; el conocimiento biológico se caracteriza por utilizar grandes volúmenes de información y su interpretación depende, en muchas ocasiones, del criterio del experto.

Los testimonios de las investigaciones vegetales recogidas en un determinado lugar, se organizan en colecciones en los herbarios. A grandes rasgos, un herbario es una colección de plantas secas, identificadas y conservadas para su estudio científico. Generalmente estas muestras se guardan en "pliegos", que incluyen la muestra y una etiqueta de identificación.

En Botánica, el problema de la clasificación y posterior identificación tiene características singulares: un pliego no cuenta en todos los casos con el individuo completo para su identificación, y las muestras recogidas dependen en muchas ocasiones de la estacionalidad, la edad del individuo, etc.

Como hemos anticipado, las investigaciones actuales persiguen la globalización de la información para que pueda ser utilizada por diferentes autores y con diferentes fines (identificación, elaboración de claves, floras, guías de campo). De ahí la importancia de establecer marcos de referencia comunes para representar el conocimiento taxonómico.

Al amparo del contexto general que acabamos de describir, nuestro trabajo pretende:

> Diseñar y desarrollar un sistema eficiente para identificar especimenes vegetales que sea correcto desde el punto de vista de la Inteligencia Artificial y la Ingeniería del Software. El sistema debe acomodar las particularidades de la Biología vegetal, trabajar con modelos de representación del conocimiento estándar y permitir la determinación a partir de las bases de conocimiento y claves de identificación generadas por el propio sistema.

De acuerdo con este objetivo, en este trabajo hemos desarrollado una *shell* de sistema experto para la identificación biológica, y particularmente vegetal, siguiendo los principios de la Ingeniería del Conocimiento. Para acomodar las

necesidades del problema biológico hemos aplicado técnicas de representación del conocimiento, modularidad, inferencia y tratamiento de la incertidumbre, sin olvidarnos de la utilización de modelos de representación de conocimiento taxonómico estándar. El proceso de identificación lleva apareado la construcción de modelos con los que poder realizar las determinaciones. Hemos desarrollado una herramienta que admite la intervención del experto en la generación de dichos modelos. Esta característica combinada con técnicas de Inteligencia Artificial produce modelos para la identificación simples, robustos y capaces de reflejar de forma fidedigna el significado biológico del conocimiento taxonómico.

Para esto comenzaremos introduciendo conceptos sobre la investigación biológica relacionados con las Ciencias de la Computación y nos centraremos, de forma específica, en la identificación taxonómica y la generación de modelos para la identificación. Una vez descritos estos dos problemas, analizaremos algunos enfoques propuestos para su solución. Dada la importancia de la utilización de modelos de representación del conocimiento taxonómico estándar, haremos un inciso en los más destacados y nos centraremos en un modelo de reciente aparición, que aspira a convertirse en un estándar aceptado por toda la comunidad científica relacionada con los estudios en Biología. Con esta perspectiva global del problema, pasaremos a describir nuestra propuesta de solución al problema de la identificación botánica. Finalizaremos con unos ejemplos de aplicación de las herramientas desarrolladas donde se aprecian las ventajas de su utilización.

En concreto, esta memoria se organiza del siguiente modo:

- En el capítulo 1 se introducen las principales áreas de investigación en biodiversidad y se sitúa la identificación taxonómica dentro de este marco general. Para ello, introducimos el concepto de identificación taxonómica y las diferentes aproximaciones que se han planteado para abordarla. Terminamos el capítulo describiendo algunos de los sistemas para identificación taxonómica más conocidos.
- En el capítulo 2 se introduce el concepto de modelo de representación de conocimiento taxonómico y posteriormente describimos aquellos modelos que han tenido más trascendencia en las investigaciones. También presentamos

una herramienta que hemos desarrollado para facilitar el intercambio de información entre ellos.

- El capítulo 3 muestra nuestra propuesta de sistema experto para la identificación taxonómica y nuestra propuesta para hacer compatible el sistema con los modelos de representación taxonómica descritos en el capítulo anterior. También analizamos el funcionamiento del sistema experto y planteamos una serie de mejoras. El capítulo termina con la descripción de la herramienta utilizada para la generación de claves dicotómicas.
- El capítulo 4 muestra e interpreta los resultados experimentales obtenidos mediante la prueba, por parte de expertos botánicos, de los sistemas desarrollados.
- A continuación se exponen las conclusiones finales que se derivan de esta memoria y los principales caminos de investigación que han aparecido durante el desarrollo de la misma y que pensamos continuar en el futuro.
- Se incluye una serie de apéndices como glosarios, protocolos de estandarización, y las claves de identificación generadas con la herramienta que hemos desarrollado.
- La memoria concluye con la bibliografía empleada en su confección o de interés sobre el tema.

Capítulo 1. Modelos para la identificación de especímenes biológicos.

El uso de computadores para el apoyo y la automatización de las tareas desarrolladas por los Biólogos no es reciente. En lo que a la Taxonomía se refiere, encontramos referencias ya en los años 70 que plantean la utilización de los ordenadores como apoyo en el proceso de identificación de especímenes biológicos [Morse, 1970; Pankhurst, 1970; Dallwitz, 1974]. De especial importancia han sido las investigaciones desarrolladas por Pankhurst y Dallwitz que se han materializado en el desarrollo de los sistemas *Pandora* y *Delta* utilizados en la actualidad.

La "*biodiversidad es la ciencia que trata del estudio de tendencias históricas y presentes en la riqueza biológica de los ambientes y, aún estando enlazada con la Biología evolutiva, la Taxonomía y la Ecología se centra fundamentalmente en la recolección y análisis de información útil para el manejo científico de recursos naturales y su conservación*" [Stockwell, 1997]. A continuación realizamos una revisión de la situación actual de la Informática dentro del ámbito de la Biología y la gestión de la biodiversidad centrándonos en

un aspecto muy concreto: la representación de información taxonómica en el ordenador y sus aplicaciones a la generación automática de claves y la identificación interactiva de especímenes biológicos.

Para comenzar, analizamos las principales áreas de investigación en biodiversidad computacional y describimos, de forma general, el problema de la identificación taxonómica y los enfoques planteados para su resolución. Haremos hincapié en los más aceptados por la comunidad científica. En concreto, las propuestas más interesantes se centran en dos líneas de trabajo muy bien diferenciadas en cuanto al tipo de sistemas desarrollados. Por un lado, la comunidad informática plantea el desarrollo de sistemas expertos y por otro la comunidad botánica plantea el desarrollo de sistemas basados en matrices. En este apartado analizamos detenidamente estos dos puntos de vista para determinar las ventajas e inconvenientes de los mismos y estudiar la conveniencia de elaborar una propuesta que integre estos dos enfoques e incorpore aspectos como las tecnologías web y XML.

Los sistemas de identificación interactiva utilizan conocimiento de forma masiva para alcanzar sus conclusiones. Este conocimiento queda completamente diferenciado del proceso utilizado para determinar el grupo taxonómico al que pertenece un individuo. Son, por tanto, sistemas basados en el conocimiento. En el ámbito de aplicación de la Biología "*tenemos que tener en cuenta que la diversidad, la falta de completitud y la excepción es la única regla válida*" [Conruyt & Grosser 1999 b]. Lo habitual es tratar con atributos imprecisos, continuos, desconocidos, dependientes del valor de otros atributos, e incluso inaplicables a un determinado grupo taxonómico. Esta casuística tan especial hace que la aplicación de técnicas de Inteligencia Artificial y la determinación de los requisitos de funcionalidad de este tipo de sistemas sean importantes frentes de trabajo.

1. Aplicación de las nuevas tecnologías a la investigación en biodiversidad.

A continuación describimos las principales líneas de investigación en la aplicación de las nuevas tecnologías a los estudios sobre biodiversidad (bases de datos, sistemas de información geográfica, distribución y paralelismo, predicción etc.). Dentro del área de los supercomputadores, el almacenamiento masivo de datos y las redes de computación masiva, distinguimos entre recursos básicos, aplicaciones avanzadas dependientes de dichos recursos y aplicaciones de carácter general [Stockwell, 1997]. En el grupo de recursos básicos destacamos:

- BASES DE DATOS BIOLÓGICAS.
- CONJUNTOS DE DATOS RELEVANTES DESDE EL PUNTO DE VISTA BIOLÓGICO.
- MODELADO DE LA DISTRIBUCIÓN DE ESPECIES.

Analicemos cada uno de ellos.

BASES DE DATOS BIOLÓGICAS (*SPECIES DATABASES*).

Las bases de datos biológicas recogen la localización de determinadas especies en un entorno concreto. Esta información es útil para monitorizar tendencias y realizar el recuento y gestión general de la biodiversidad. Incluso permite dar respuesta a preguntas como *¿qué especies están amenazadas?* Los problemas que hacen objeto de investigación a estas bases de datos son:

- Su tamaño. En potencia puede ser muy grande. A groso modo podríamos estimar dicho tamaño como el producto del número de especies descritas (1.5 millones) por el número de registros por especie (unos 1000 registros por especie para obtener una cobertura global de la misma, que se incrementaría durante la fase de monitorización). Esto hace estimar en un billón el número de registros de una base de datos internacional de estas características.

- Se trata de bases de datos distribuidas. Se presentan como grandes retos el desarrollo de una base de datos centralizada y consistente a partir de las fuentes de datos distribuidas y el desarrollo de una base de datos distribuida formada por fuentes de información interconectadas.

CONJUNTOS DE DATOS RELEVANTES DESDE EL PUNTO DE VISTA BIOLÓGICO.

Existen variables especialmente relevantes desde el punto de vista biológico porque permiten responder preguntas sobre el *por qué* y *dónde* viven los organismos (ejemplos de estas variables son el clima, la geología y los procesos humanos y biológicos). En la actualidad son muy escasas las fuentes de información biológica de este tipo, más aún si pretendemos que esté referida a cualquier escala. El tamaño de los datos a almacenar se dispara a medida que su granularidad es más fina. Por ejemplo, una capa climática del ecuador con una resolución de 100 *Km* puede ocupar unos 4 *MB*, pero en ocasiones se necesita una resolución de menos de 100 *m*, lo que aumenta los requerimientos 4 *TB*. Por esto, es importante diseñar herramientas que suministren conjuntos de datos relevantes desde el punto de vista biológico, referidos a una escala y una región geográfica definida por el usuario partiendo de cualquier región y utilizando los recursos computacionales y de memoria masiva de los supercomputadores.

MODELADO DE LA DISTRIBUCIÓN DE ESPECIES.

La información sobre la distribución de las especies es casi siempre incompleta porque no es viable realizar un muestreo que cubra una gran área y un gran número de especies. La investigación en modelos no lineales para la respuesta de las especies ante el ambiente pretende ofrecer respuestas fiables ante datos escasos. Para esto, se necesita el apoyo de métodos más complejos y computacionalmente costosos como, por ejemplo, algoritmos genéticos.

Vistos los recursos básicos, pasemos a describir el grupo de aplicaciones avanzadas. Dentro de este grupo distinguimos las siguientes aplicaciones:

- MODELADO DE LA BIODIVERSIDAD.
- MODELADO DE PROCESOS ECOLÓGICOS.
- SISTEMAS PARA LA PLANIFICACIÓN DE RESERVAS.

MODELADO DE LA BIODIVERSIDAD.

Los estudios sobre la biodiversidad están íntimamente ligados al estudio de las variaciones en las poblaciones en función de los cambios ambientales. Por lo general, la representación de la biodiversidad de una región se simplifica utilizando variables agregadas denominadas bioindicadores (*environmental surrogates*). Ejemplos de estas variables son el tipo de vegetación, o especies animales cuya presencia en un entorno indica la presencia de otras. Esta simplificación puede llevar en ocasiones a sistemas algo pobres. Una alternativa es desarrollar mapas predictivos para cada una de las especies para luego recombinarlos, lo que requiere grandes recursos computacionales.

MODELADO DE PROCESOS ECOLÓGICOS.

La Ecología incorpora un aspecto dinámico fundamental para los estudios en biodiversidad (migraciones estacionales, invasiones de organismos exóticos y enfermedades, etc.). La combinación de modelos ecológicos con el componente espacial introduce un mayor grado de exactitud y realismo y, puesto que los estudios en biodiversidad son *per-se* computacionalmente costosos, la adición de estos aspectos dinámicos repercute en un incremento adicional de la complejidad.

SISTEMAS PARA LA PLANIFICACIÓN DE RESERVAS.

Esta disciplina se dedica a la planificación de áreas naturales que contengan de forma adecuada y eficiente una muestra extensa del ambiente natural. Incluso los enfoques más sencillos para acometer este problema son computacionalmente costosos, lo que ha limitado el análisis a simplificaciones

bastante toscas que sólo garantizan el óptimo bajo condiciones muy estrictas que generalmente no se dan.

Tras el análisis de los recursos básicos y avanzados, describimos un conjunto de aplicaciones muy estudiadas que son de uso general. Dentro de este grupo incluimos la investigación en:

- LÍMITES DE CONFIANZA ESTADÍSTICA.
- VISUALIZACIÓN / REALIZACIÓN DE MAPAS.
- INTERACTIVIDAD.

LÍMITES DE CONFIANZA ESTADÍSTICA.

La estimación de la confianza o los límites de confianza estadística de una predicción es tan importante como la precisión de la misma. Cuando tratamos con variables estadísticas muy bien definidas podemos calcular los límites de confianza de forma muy sencilla, pero cuando los datos no son perfectos debemos aplicar otros métodos que requieren generalmente un gran esfuerzo computacional. Los conjuntos de datos biológicos y ambientales nunca cumplen las suposiciones utilizadas por la metodología estadística, esto es: las distribuciones no son normales y existen dependencias entre variables. Por esto cobra gran importancia la utilización de métodos de remuestreo (*resampling methods*) para obtener estimaciones robustas de la confianza en las predicciones.

VISUALIZACIÓN / REALIZACIÓN DE MAPAS.

El mapa es la herramienta de visualización más utilizada en Ecología y conservación. El medio ambiente es multidimensional y contiene complejas interrelaciones, lo que hace necesarios nuevos métodos de visualización.

INTERACTIVIDAD.

La utilización de Sistemas de Información Geográfica (GIS) ha cobrado un gran interés. Las interfaces para mapas interactivos en la web facilitan el acceso a bases de datos de biodiversidad y añaden flexibilidad a la visualización de mapas (rotaciones, cambios de escala, etc.).

La Tabla 1-1 resume las áreas de investigación en biodiversidad que acabamos de describir.

ÁREA DE INVESTIGACIÓN	OBJETIVOS
Bases de datos de especies	Base de datos distribuida de carácter internacional que almacene el catálogo de todas las especies conocidas y la localización de especímenes. Se preveían para 2001 un billón de registros georeferenciados.
Conjuntos de datos relevantes desde el punto de vista biológico	Diseño de herramientas que suministren conjuntos de datos relevantes desde el punto de vista biológico a una escala y una región geográfica definida por el usuario, utilizando los recursos computacionales y de memoria masiva de los supercomputadores.
Modelado de la distribución de especies	Incrementar la exactitud y disponibilidad de modelos para la distribución de especies. Servicios como el web permitirán acceder a mapas más exactos o a especies de forma más rápida.
Modelado de la biodiversidad	Predicción de la biodiversidad para grupos taxonómicos a partir de la predicción de especies individuales.
Modelado de procesos ecológicos	Incorporación de procesos ecológicos dentro de modelos de biodiversidad de gran resolución espacial.
Sistemas para la planificación de reservas	Delimitación de áreas que contengan de forma adecuada y eficiente una muestra extensa del ambiente natural.
Límites de confianza estadística	Mejorar la robustez del modelado utilizando métodos de remuestreo.
Visualización / Realización de mapas	Métodos mejorados para visualizar datos sobre la biodiversidad complejos y de múltiples dimensiones.
Interactividad	Provisión de información interactiva vía www.

Tabla 1-1. Áreas de investigación en biodiversidad computacional [Stockwell, 1997].

Hasta ahora nos hemos centrado en las áreas de la predicción y supercomputación. En la Taxonomía y la Sistemática encontramos otro frente de trabajo importante, pues la información que se maneja es compleja tanto en lo referente al diseño de una estructura de almacenamiento como al vocabulario con el que se trabaja (muy diverso y variable) [Berendsohn, 2001]. Por esto, uno de los retos actuales reside en la creación de modelos descriptivos consensuados y herramientas basadas en dichos modelos para identificar especies tanto en el

campo como en el laboratorio basadas en dichos modelos: *"Puesto que la biodiversidad trata del estudio de la variedad de los seres vivos, el primer paso en el estudio de la biodiversidad es la identificación de los diferentes tipos de organismos presentes en el biotopo que se pretende estudiar*[3]" [Diederich *et al.,* 2000].

Las colecciones biológicas son otra aplicación de gran utilidad, ya que cada especimen depositado en ellas es una muestra de lo que hubo en un lugar y un momento determinado. Ya se han realizado algunos estudios interesantes en esta dirección, como la predicción de la influencia del cambio climático sobre más de 1800 especies de la fauna mejicana [Peterson *et al.,* 2002], cuyos resultados fueron publicados en la revista *Nature*[4].

En cuanto a la estandarización de las bases de datos biológicas, el IUBIS-TDWG (*International Union of Biological Sciences-Taxonomic Database Working Group*) realiza muchos esfuerzos para desarrollar y promover el uso de estándares que faciliten el intercambio de información. En esta línea se ha llevado a cabo una intensa investigación en el campo del diseño de modelos de bases de datos para la representación de información taxonómica [Berendsohn, 1997; Pullan *et al.,* 2000].

En lo que se refiere a los nombres de *taxa* y las colecciones biológicas, aunque existen proyectos para recopilar los nombres, por ejemplo, de plantas superiores [IPNI, 1999], se persigue unificar todas las bases de datos de especies. En esta línea se encuentra el proyecto *Species 2000* [Bisby, 1998].

Por último citamos una iniciativa internacional muy ambiciosa y reciente: GBIF (*Global Biodiversity Information Facility*), que pretende poner accesible en

---

[3] En el campo el investigador encuentra un organismo, pero lo que se almacena en un registro en una base de datos es su identificación. Si esta determinación no se realiza de forma correcta, se degrada la calidad de la información almacenada en la base de datos.

4 Peterson, A. T.; Ortega-Huerta, M. A.; Bartley, J.; Sanchez-Cordero, V.; Soberon, J.; Buddemeier, R. H.; Stockwell, D. R. B. 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature* 416 (6881): 626-629.

Internet la información disponible sobre todos los organismos vivos conocidos. Por su importancia como entidades gestoras y vanguardistas en los estudios sobre biodiversidad, incluimos a continuación una pequeña reseña sobre TDWG, SDD y GBIF.

INSTITUCIÓN	CONTACTO
DTWG, *International Union of Biological Sciences Taxonomic Database Working Group*	http://www.tdwg.org/
SDD, *Structure of Descriptive Data*	http://www.tdwg.org/
SDSC, *San Diego Supercomputer Center*	http://www.sdsc.edu/
GBIF, *Global Biodiversity Information Facility*	http://www.gbif.org
ETI, *Expert Center for Taxonomic Identification*	http://www.eti.uva.nl/
RBGE, *Royal Botanic Garden Edinburgh*	http://www.rbge.org.uk/rbge/web/
BBA, *Federal Biological Research Centre for Agriculture and Forestry*	http://www.bba.de
CSIRO, *Division of Entomology. Commonwealth Scientific and Industrial Research Organization*	http://biodiversity.uno.edu/delta/

Tabla 1-2. *Contacto con algunas instituciones de reconocido prestigio por sus investigaciones en biodiversidad.*

1.1 ¿Qué son TDWG, SDD y GBIF?

TDWG es el acrónimo del *International Working Group on Taxonomic Databases*. Se trata de un grupo de trabajo formado por investigadores de reconocido prestigio cuyos objetivos son:

1. Proporcionar un foro internacional para la discusión de proyectos relacionados con información de tipo biológico.
2. Desarrollar y promover el uso de estándares.
3. Facilitar el intercambio de datos.

Históricamente, dentro de IUBIS (*International Union for Biological Sciences*), el grupo de trabajo TDWG (*Taxonomic Database Working Group*) ha abordado el desarrollo de estándares para representar información taxonómica en bases de datos botánicas. Entre sus miembros se encuentran las principales instituciones botánicas de todo el mundo. En 1994, el grupo de trabajo asumió el

rol de abarcar todo tipo de base de datos taxonómicas, y actualmente también cuenta con miembros especializados en zoología y microbiología.

Biological Collections Data
Economic Botany
Geography
Structure of Descriptive Data
TDWG Process
Spatial Data Standards

Tabla 1-3. *Grupos de trabajo de TDWG.*

El grupo de trabajo SDD (*Structure of Descriptive Data*) se inició en 1998. En el encuentro de 1999 celebrado en Harvard se concluyó que el subgrupo debía analizar los requerimientos para un nuevo estándar interoperativo de información descriptiva. El estándar (que veremos con posterioridad) se basa en XML y esquemas XML y se espera que alcance reconocimiento mundial y se convierta en el sucesor de otros como *Delta* o *Nexus*, que serán descritos en el Capítulo 1.4.8 y Capítulo 2.1.3 respectivamente. El nuevo estándar debe ser diseñado cuidadosamente y aceptado por consenso.

GBIF (*Global Biodiversity Information Facility- Infraestructura Mundial de Información sobre biodiversidad*) es una organización independiente cuyos miembros son países u organizaciones internacionales. Se crea formalmente en 2001 con el propósito de poner en Internet, de forma gratuita, toda la información disponible sobre los organismos vivos mundialmente conocidos. Las estimaciones indican que las colecciones del mundo albergan de 1,5 a 2 mil millones de especímenes repartidos en miles de centros. Es como el proyecto "Genoma Humano" de la biodiversidad. Para ello, GBIF pretende:

- Informatizar las colecciones de historia natural.
- Desarrollar la tecnología necesaria para permitir la interoperabilidad de la red GBIF.
- Construir el catálogo unificado de nombres científicos.
- Promover la formación y cooperación.
- Compilar bases de datos con información sobre especies (descripciones, claves, imágenes, etc.).
- Crear una biblioteca virtual de biodiversidad.

Su misión fundamental es apoyar a las colecciones, centros y proyectos relevantes sobre biodiversidad:

- Proporcionando soporte técnico: información, formación, estándares, software y asesoramiento.
- Asegurando la coherencia entre las distintas iniciativas para garantizar la interoperabilidad.
- Investigando como maximizar el valor de los datos al desarrollar herramientas de análisis, validación y visualización de los mismos.
- Recopilando y difundiendo información relevante para las colecciones y para el conocimiento y gestión de información sobre biodiversidad.

GBIF plantea una estructura de red interoperativa, no centralizada y abierta de bases de datos para poner al alcance de toda la comunidad científica información actualizada, representativa y científicamente validada (ver Figura 1-1). Esto ayudará significativamente a la realización de estudios a una escala hasta ahora inabordable: modelos que expliquen la distribución de las especies, modelos predictivos en función del cambio climático, etc., y a la toma de mejores decisiones sobre conservación y uso de la biodiversidad en el planeta.

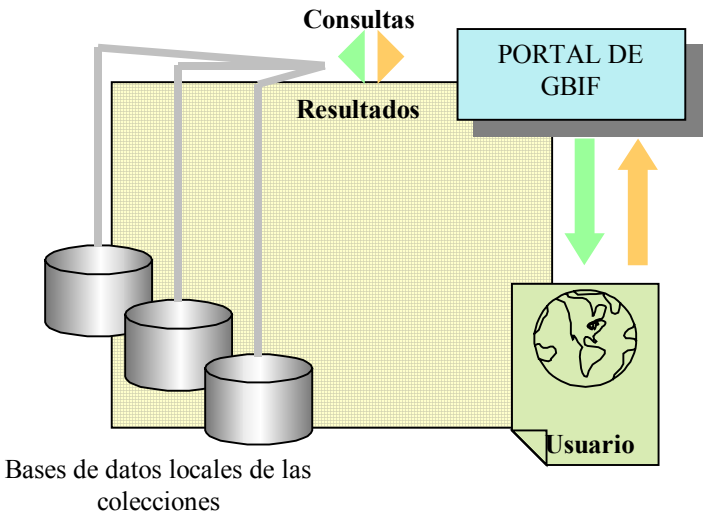


Figura 1-1. *Organización de GBIF.*

2. El problema de la identificación taxonómica.

Hemos visto que las investigaciones pretenden el diseño de sistemas globales de información para la biodiversidad capaces de relacionar información geográfica, climática y medioambiental con información relacionada con el aspecto molecular o fisiológico de los organismos, valores bioindicadores, utilización humana, etc. Dentro de este contexto, la identificación taxonómica juega un papel fundamental como punto de partida para los estudios sobre biodiversidad.

La identificación taxonómica es necesaria para biólogos, especialistas en conservación y en general cualquier persona relacionada con la biodiversidad y el análisis de impactos. En todos estos casos se necesita identificar correctamente los organismos implicados en el análisis del problema concreto al que nos enfrentemos.

Esta identificación debe ser realizada por especialistas en un grupo taxonómico determinado, pero no siempre es posible disponer de un experto. Este factor puede llevar al retraso del proyecto del cual depende dicha identificación. Por esto es importante disponer de herramientas que sirvan de apoyo a esta actividad. Tradicionalmente la herramienta utilizada por el investigador para llevar a cabo la determinación ha sido la clave dicotómica, una estructura con forma de árbol que presenta al usuario una serie de elecciones en cada paso (ver Figura 1-2).

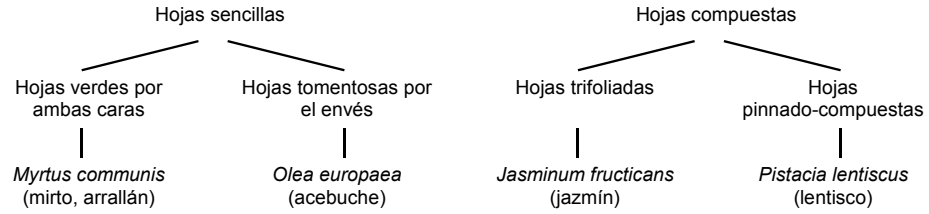


Figura 1-2. Ejemplo de clave dicotómica [Morales *et al.,* 2001].

No obstante, y a pesar de ser la herramienta más utilizada, hay otros métodos para determinar a qué especie o grupo taxonómico pertenece un

individuo. El objetivo de este apartado es, precisamente, hacer un recorrido por las diferentes herramientas con que el biólogo cuenta para realizar esta tarea.

2.1 Enfoques para la identificación taxonómica.

Podemos considerar los siguientes enfoques para llevar a cabo la identificación taxonómica [Diederich *et al.*, 2000]:

- DISCRIMINACIÓN INSTANTÁNEA.
- SISTEMAS DE IDENTIFICACIÓN MOLECULAR.
- CLAVES IMPRESAS.
- HERRAMIENTAS AUTOMATIZADAS.

DISCRIMINACIÓN INSTANTÁNEA.

Es el método de identificación que el ser humano lleva a cabo de forma inconsciente. Por ejemplo, no necesitamos utilizar ningún tipo de herramienta para identificar un árbol o un pez. Aunque generalmente este método no nos conducirá a una identificación completa, esta idea puede ayudarnos a eliminar gran parte del abanico de caminos de búsqueda.

SISTEMAS DE IDENTIFICACIÓN MOLECULAR.

Este método de identificación es, probablemente, el único enfoque posible para alcanzar una identificación en casos especiales en los que, por ejemplo, dos virus son prácticamente idénticos en morfología pero infectan diferentes plantas. A pesar de esto, no se trata de un enfoque demasiado prometedor pues se trata de un proceso largo, difícil y costoso. Se da el problema de que para determinar las moléculas marcadoras en las especies necesitamos un especimen patrón bien identificado. La pregunta que se plantea es *¿cómo identificamos el patrón?*

CLAVES IMPRESAS.

Es el segundo método más antiguo para acometer el proceso de identificación, no en vano "*ha sido el método utilizado durante los últimos 200 años*" [Dallwitz, 1992]. Un problema, relacionado en general con cualquier tipo de clave en papel, es lo difícil que resulta su actualización.

Claves dicotómicas.

La mayoría de las claves impresas son claves dicotómicas. Son de gran utilidad pero tienen sus inconvenientes. El mayor de ellos consiste en que proceden por eliminación, por lo que no disponer de la información necesaria para pasar de nivel, impide llevar a cabo la identificación. Por ejemplo, las claves para géneros de un tamaño medio requieren gran cantidad de caracteres que no están siempre disponibles. Las claves son herramientas muy poderosas cuando tratamos con caracteres de identificación primarios [Fortuner, 1989] esto es, caracteres capaces de diferenciar especies o grupos de especies con un riesgo de error muy pequeño.

Claves multientrada o tabulares.

Para evitar que la falta de un dato impida la identificación, se utilizan claves multientrada o tabulares en las que la eliminación de especies y grupos de especies depende de varios caracteres. Este tipo de claves son fáciles de utilizar con géneros pequeños pero resultan demasiado incómodas cuando el número de especies se incrementa.

HERRAMIENTAS AUTOMATIZADAS.

Los primeros indicios de identificación computerizada datan de principios de los años 70, pero no se han utilizado de forma regular por usuarios distintos de los autores que han desarrollado sus propias herramientas en los últimos 30 años.

En función del método utilizado para llevar a cabo la determinación, distinguimos los siguientes enfoques:

- SIMILITUD.
- ESTADÍSTICA MULTIVARIABLE.
- CLAVES COMPUTERIZADAS.
- SISTEMAS EXPERTOS.
- REDES NEURONALES.
- SISTEMAS PROBABILÍSTICOS.

Similitud.

Este enfoque se basa en la utilización de una medida de similitud entre un especimen desconocido y cada una de las especies de un género. Por lo general, se calcula a partir de 20-25 caracteres, aproximadamente el mismo número de caracteres que tiene una clave dicotómica, con la diferencia de que consideramos todos estos caracteres a la vez para obtener una única medida de similitud. Esto significa que las especies se organizan en función de todos los datos disponibles, de forma que si se ha introducido un dato erróneo el coeficiente calculado pierde exactitud, pero puede conducir a una identificación correcta. De este modo, el rendimiento del sistema decrece con los errores de forma gradual. La similitud no da una única respuesta, sencillamente organiza las especies candidatas en orden a su parecido con el especimen. La identificación debe concluir un único resultado por lo que este método no es suficiente y tiene que ser acompañado de otros. Existen diversos coeficientes de similitud, pero el más utilizado es el coeficiente de similitud de Gower [Gower, 1971].

Estadística Multivariable (Multivariate Statistics).

Es un enfoque estadístico basado en técnicas de agrupamiento (*clustering)* en el que todos los caracteres se utilizan de forma simultánea para determinar la proximidad relativa del especimen desconocido a todas las especies consideradas. Es más poderoso que la similitud porque facilita una medida de la distancia entre diversas entidades. El problema fundamental es que el hecho de que un especimen

tenga una distancia corta a una especie no quiere decir que pertenezca a esa especie. Lamberti y Ciancio utilizan este método para analizar las relaciones entre especies de un género de nemátodos y concluyen que no es un método práctico cuando el objetivo es la identificación [Lamberti & Ciancio, 1994].

Claves computerizadas.

Un ordenador permite utilizar con facilidad una clave multientrada, de forma que el usuario puede decidir qué caracteres introducir y en qué orden. Además hace posible utilizar una medida de la degradación que permite introducir uno o más caracteres erróneos antes de desechar una opción, lo que reduce el riesgo de obtener una respuesta incorrecta. Dentro del grupo de las claves computerizadas incluimos también la clave gráfica, una clave dicotómica en la que las elecciones se presentan en forma de imágenes que ilustran las opciones. Este tipo de clave se puede utilizar de dos formas: el usuario busca a través de la clave hasta encontrar una imagen que se parezca al especimen a identificar o bien puede utilizarla de forma dicotómica.

Sistemas expertos.

El término *sistema experto* se utiliza de forma indiscriminada dentro de la Biología. Por ejemplo, en ocasiones son considerados sistemas expertos el sistema *Pankey* [Pankhurst, 1970] o *Confor* [Dallwitz *et al.*, 2000] cuando realmente se trata de claves tabulares computerizadas. Los sistemas expertos ofrecen importantes ventajas con respecto a las claves: permiten utilizar todos los caracteres de forma simultánea en lugar de hacerlo de forma secuencial, proporcionan funciones extra para explicar sus razonamientos y un grado de confianza en la respuesta que devuelven.

Redes neuronales.

Se han desarrollado algunos trabajos, como los de Boddy [Boddy *et al.*, 1998] con redes neuronales artificiales para la identificación taxonómica. Pero este enfoque tiene inconvenientes como: la escasez de conjuntos de datos de entrenamiento, la necesidad de introducir valores para todos los caracteres que se utilizan para entrenar la red y no facilitar la justificación de sus razonamientos.

Sistemas bayesianos y probabilísticos.

Los sistemas bayesianos se basan en la regla de Bayes. Esta regla nos da la probabilidad de tener una especie *x* dada la evidencia disponible (ver Fórmula 1-1).

$$P(especie / evidencia) = \frac{P(evidencia / especie) * P(especie)}{P(evidencia)}$$

Fórmula 1-1. *Regla de Bayes aplicada a la identificación taxonómica.*

Uno de los principales problemas de la utilización de este enfoque es la necesidad de proporcionar una distribución de probabilidad *a priori* de la observación de una determinada especie (*P(especie)*). Estas probabilidades sólo se pueden proporcionar cuando las identificaciones se realizan en ambientes muy bien conocidos por los expertos.

La Tabla 1-4 presenta un resumen de los diferentes enfoques para la identificación taxonómica que acabamos de describir.

ENFOQUE	LIMITACIONES
Discriminación instantánea	No puede conducir a una identificación completa, pero sirve para podar caminos a explorar.
Sistemas de identificación molecular	Proceso largo, difícil y costoso del que hay que partir de especímenes bien identificados, el problema es elaborar el catálogo de todas las especies.
Claves dicotómicas	Su utilización nos hace caer en ocasiones en respuestas erróneas porque proceden por eliminación. Son de difícil actualización y generalmente requieren caracteres que no están

ENFOQUE	LIMITACIONES
	siempre disponibles. Son útiles cuando utilizan caracteres capaces de diferenciar grupos con un riesgo de error muy pequeño.
Claves tabulares o multientrada	Fáciles de utilizar con géneros pequeños, pero se hacen incómodas cuando el número de especies se incrementa. Son de difícil actualización.
Similitud	No da una única respuesta sino que organiza las especies candidatas en orden a su parecido con el especimen. La identificación debe concluir dando un único resultado por lo que este método no es suficiente y tiene que ser acompañado de otros
Estadística multivariable (*Multivariate statistics*)	Estudios concluyen que se trata de un método poco práctico cuando el objetivo es la identificación
Sistemas expertos	A pesar de proporcionar funciones para explicar sus razonamientos y expresar un grado de confianza en la respuesta que devuelven son una forma de escribir una clave.
Redes neuronales	Escasez de conjuntos de datos para el entrenamiento, necesidad de introducir los valores para todos los caracteres que se utilizan para entrenar la red y no existir una forma sencilla de explicar sus razonamientos.
Sistemas probabilísticos	Hay que proporcionar una distribución de probabilidad *a priori* de la observación de una determinada especie. Estas probabilidades sólo se pueden proporcionar cuando las identificaciones se realizan en ambientes muy bien conocidos por los expertos.

Tabla 1-4. *Enfoques para la identificación taxonómica y sus limitaciones [Diederich et al., 2000].*

Una vez determinadas las técnicas aplicables a la identificación por ordenador y las particularidades de la representación de información taxonómica, pasemos a analizar los problemas y requerimientos funcionales de las herramientas informáticas desarrolladas con este propósito.

2.2 Problemas y requisitos de un sistema para la identificación interactiva.

PROBLEMAS ASOCIADOS A LA IDENTIFICACIÓN POR ORDENADOR.

Podemos distinguir dos tipos de problemas asociados con la identificación asistida por computador [Diederich *et al.*, 2000]:

- PROBLEMAS RELACIONADOS CON LA BASE DE DATOS. Cada base de datos se asocia a una herramienta particular e incluye sólo un pequeño subconjunto de

características seleccionadas por el autor de la misma. Además, el formato en que se recogen los caracteres no es en absoluto uniforme pues cada autor ha creado su propio modelo. Un inconveniente añadido es que la mayoría de los modelos no son capaces de representar metadatos (por ejemplo, entradas de glosario) necesarios para crear herramientas de identificación que puedan ser utilizadas por no expertos.

- PROBLEMAS ASOCIADOS CON LAS HERRAMIENTAS DE IDENTIFICACIÓN. Cada herramienta aplica un único enfoque, esto es, eliminación (claves), comparación (coeficientes de similitud), sistemas expertos, redes neuronales, etc. Un sistema de identificación general debería acomodar las preferencias de todos los usuarios y no forzar la utilización de un enfoque determinado.

REQUISITOS FUNCIONALES PARA LA IDENTIFICACIÓN INTERACTIVA.

Es deseable que el sistema guíe al usuario durante la identificación mediante la sugerencia de los caracteres más prometedores y su ordenación de acuerdo a su capacidad de separación o a un factor de utilidad (*reliability*) determinado por el experto. Todo ello, sin perjuicio de que sea el usuario el que decida finalmente qué carácter seleccionar y en qué orden utilizarlo.

También son útiles las capacidades de eliminar de la lista de caracteres y estados disponibles aquellos que no son útiles en un determinado paso de la identificación y fijar valores de caracteres para mantener esta información cuando queremos comenzar a identificar a partir de un determinado nivel taxonómico. Además es útil el uso de probabilidades para los valores de los estados de los *taxa* y probabilidades de error de los usuarios que se pueden utilizar para calcular las probabilidades de que un especimen pertenezca a un determinado taxon.

Un sistema de identificación taxonómica debe incluir ayuda en línea. Esta ayuda comprende manuales de usuario y conocimiento sobre el dominio de aplicación, por ejemplo, glosarios, texto explicativo, ilustraciones, descripciones y subconjuntos de caracteres para facilitar su selección.

Hay algunos requerimientos obvios, que a veces se pueden olvidar: los requerimientos de memoria dentro de un rango normal, la ejecución rápida, la capacidad de funcionar en Internet y la posibilidad de concluir respuestas correctas a pesar de la presencia de errores.

Dallwitz [Dallwitz, 2000 a] sugiere que el sistema debe importar y exportar formato *Delta*, a esto tenemos que añadir que este requerimiento ha evolucionado y hay que incorporar la capacidad de importar y exportar el nuevo estándar, *SDD* (que será descrito en el Capítulo 2.1.4).

Por la naturaleza del dominio de aplicación al que nos enfrentamos, las claves enlazadas son muy comunes. Distinguimos dos tipos de claves enlazadas:

- Claves jerárquicas integrales. Posibilidad de tener claves para diferentes niveles taxonómicos en una misma matriz de datos.
- Claves jerárquicas separadas. Posibilidad de tener *taxa* enlazados con claves (por ejemplo, un género enlazado con la clave correspondiente a las especies de dicho género).

Otros requerimientos, derivados de las características de la información taxonómica son la capacidad de expresar dependencia entre caracteres, distinguir entre valores nulos desconocidos e inaplicables, valores numéricos y permitir la selección de más de un estado o un rango de valores numéricos.

Por último, incluimos las facilidades para recuperar información, aspecto que, aunque no es identificación taxonómica *sensu-estricto*, incrementa la utilidad del sistema. Aquí se incluye la posibilidad de encontrar cadenas de caracteres en los nombres de taxon, sinónimos y nombres comunes y todos los *taxa* que presentan determinados atributos. La capacidad de mostrar descripciones de calidad generadas directamente de los datos utilizados en la identificación y las diferencias y similitudes entre *taxa* es otro valor añadido.

HERRAMIENTAS NECESARIAS PARA UN SISTEMA DE IDENTIFICACIÓN GENERAL.

El proyecto *Genisys* (*GENeral Identification SYStem)* pretende definir los requerimientos para un sistema de identificación general compuesto por un conjunto de herramientas genéricas para asistir al usuario en diferentes tareas relacionadas con la identificación (entrada de datos, selección de un grupo biológico, eliminación, comparación, verificación) y que puede ser utilizado con cualquier grupo biológico [Diederich *et al.,* 2000].

Aunque no existe ningún prototipo implementado, el proyecto *Genisys* analiza los diferentes enfoques que se han dado al problema y propone un conjunto de utilidades necesarias para construir un sistema de identificación general en el que el usuario siempre tiene la libertad de escoger la herramienta más adecuada, los datos a introducir, las especies candidatas a considerar, los valores para los umbrales y cualquier otro aspecto relacionado con la identificación. Estas utilidades son:

- Una herramienta para exportar datos de una base de datos general a los modelos de datos utilizados por las herramientas de identificación actuales. Así se asegura la compatibilidad con las herramientas más populares.
- Una herramienta basada en la "discriminación instantánea" para reducir el número de especies candidatas, disminuir la cantidad de datos de entrada y garantizar que solo se solicita al usuario la información necesaria.
- Incluir enfoques que necesitan muy poca información del usuario, por ejemplo, identificación gráfica y búsqueda textual.
- Una herramienta de eliminación (por ejemplo, un sistema experto o una clave multientrada) basada en caracteres de identificación primarios [Fortuner, 1989], caracterizados por su facilidad de observación (por ejemplo, órganos muy visibles con propiedades no ambiguas que no varían dentro de un taxon). Permiten eliminar las especies diferentes de forma obvia del especimen.
- Además resulta de interés incorporar una herramienta de eliminación basada en un factor de utilidad de la información (*endorsement*) [Diederich & Fortuner, 1996]. Este factor puede ser utilizado, por ejemplo, para seleccionar

los caracteres más útiles en una clave o asignar un peso a los caracteres para el cálculo de un coeficiente de similitud. Se calcula según la expresión Fórmula 1-2. En [Diederich & Fortuner, 1996] también podemos encontrar un enfoque basado en reglas difusas.

$$Endorsement = Expertise * Pif + (1 - Espertise) * CPif$$

Fórmula 1-2. *Cálculo algorítmico del factor de utilidad (endorsement)[Diederich & Fortuner, 1996].*

Donde:

Expertise. Es el nivel de conocimientos del usuario. Cuanto mayor sea este valor, mayor será la importancia de la intuición del usuario para calcular el factor de utilidad.

Pif (*Personal Intituitive Feeling*) indica cuan bueno considera el usuario que es el dato.

CPif: (*calculated Pif*) es una combinación aritmética de la visibilidad de la estructura biológica (*conspicuity*), la ambigüedad (*ambiguity*) y la variabilidad del carácter en un taxon particular (*variability*).

- Tras la eliminación de las especies irrelevantes, una herramienta de comparación podría ordenar las especies restantes de acuerdo con su similitud con el especimen.
- Con una herramienta de búsqueda el usuario podría mirar las descripciones (texto o imágenes) de las especies ordenadas.
- Por último, una herramienta de diagnóstico podría verificar que los caracteres clave de la especie seleccionada están presentes.

2.3 Sinopsis.

Tras este estudio de los enfoques para llevar a cabo la identificación taxonómica observamos que la herramienta automatizada más utilizada dentro de la Biología es la clave (dicotómica o multientrada). Concretamente las claves

tabulares permiten al investigador manejar la información de forma cómoda, eficiente y facilitan su mantenimiento.

Por otro lado, la representatividad de los sistemas basados en redes neuronales y probabilísticos es prácticamente nula dentro del total de los sistemas desarrollados para la identificación. El número de sistemas para la identificación que incorporan técnicas de Inteligencia Artificial es muy pequeño y la mayoría son sistemas expertos. No obstante estos sistemas se han desarrollado fuera del ámbito de la Biología o como proyectos muy específicos y concretos.

La utilización de sistemas expertos puede resultar una idea prometedora debido, por ejemplo, a las capacidades de diálogo con el usuario, de justificar su razonamiento y de acomodar la incertidumbre inherente al problema de la identificación taxonómica.

A modo de resumen, al estudiar las diferentes propuestas de los autores para la solución de este problema, apreciamos dos líneas de trabajo muy claras:

1. Por un lado, gran parte de la comunidad de biólogos y, de forma particular, botánicos ha adoptado el sistema *Delta* (Capítulo 1.4.8) como herramienta de trabajo.
2. Por otro, se han desarrollado sistemas de identificación de forma más aislada (que han sido analizados en el apartado anterior) cuyo objetivo era la determinación de grupos taxonómicos específicos.

En los siguientes apartados nos detenemos, de forma especial, en estos dos enfoques para analizar cuáles han sido las principales herramientas desarrolladas en los últimos años así como en sus características más relevantes.

3. Revisión de sistemas expertos para la identificación taxonómica.

Al realizar la búsqueda bibliográfica de sistemas expertos para la identificación biológica detectamos que este tipo de sistemas no abunda, y mucho menos integrados dentro de Internet. También encontramos que generalmente se trata de trabajos aislados y desarrollados para proyectos muy concretos, lo que hace suponer que no existe una línea de investigación definida en este aspecto. En este análisis estudiamos tres sistemas:

- *IKBS*.
- *RIH*.
- *Agassistant*.

3.1 *IKBS*.

Conruyt y Grosser han diseñado *IKBS* (*Iterative Knowledge Based System*) para el tratamiento específico de la información sobre Taxonomía y biodiversidad [Conruyt *et al.*, 1997; Conruyt & Grosser, 1999 a; Conruyt & Grosser, 1999 b]. Este sistema pretende construir bases de conocimiento más adaptadas a las ciencias naturales y tratar con los problemas habituales de estructuración del conocimiento, dependencia entre objetos y la presencia de ruido y valores variables, desconocidos e imprecisos en los datos.

ADQUISICIÓN Y REPRESENTACIÓN DEL CONOCIMIENTO.

El proceso de construcción de la base de conocimiento comprende al proceso de adquisición, procesamiento y validación y refinamiento del conocimiento.

Durante la adquisición del conocimiento se obtiene un modelo descriptivo en forma de árbol que representa las características observables (objetos, atributos y valores) de los individuos pertenecientes a un dominio específico. La raíz del árbol es el nombre del dominio y cada nodo del árbol es un objeto definido por una lista de atributos con sus posibles valores. A partir de este modelo, el programa genera de forma automática un cuestionario que facilita al experto crear una base de casos. Este proceso genera subárboles del modelo descriptivo que permiten una comparación entre casos muy sencilla.

A partir de esta descripción se generan reglas mediante la aplicación de un C4.5 que permite trabajar con atributos continuos y se comporta de forma diferente según trate con objetos estructurados, atributos multievaluados o pares atributo / valor taxonómicos. La validación y refinamiento se realiza de forma manual.

El proceso de inducción con *IKBS* consiste en un algoritmo clásico de generación de reglas que se comporta de forma diferente según el tipo de atributo con que se trate:

- Atributos estructurados. Si se selecciona un atributo estructurado, el algoritmo tiene en cuenta que al generar un nodo con un atributo del que dependen otros habrá que incluir también estos otros atributos en el árbol. Además, si un atributo del que dependen otros tiene un valor desconocido, estos otros no serán visitados para evitar atributos inaplicables como clasificadores.

- Pares atributo / valor taxonómicos. Cuando se selecciona como carácter de ramificación un atributo taxonómico estructurado con relaciones de tipo jerárquico, el método consiste en crear un conjunto de particiones correspondientes al primer nivel de jerarquía y asignar cada caso a la partición que generaliza su valor.

- Atributos multievaluados. Un atributo discreto (nominal o taxonómico) puede ser multievaluado. En función de la semántica, *IKBS* puede aplicar uno de los siguientes métodos de procesamiento:

- Duda. Considera que se trata de una disyunción.
- Presencia simultánea o conjunción de estados. En este caso se crean *k* particiones correspondientes a cada uno de los posibles estados repartiendo los casos con dicho valor en cada partición.
- Establecer un umbral de similitud. De esta forma para repartir los casos, el valor de este tipo de atributos no tiene que coincidir exactamente con el correspondiente al del nodo.

PROCESO DE CONSULTA.

Podemos consultar *IKBS* con dos objetivos.

- Clasificación[5]. En este caso se construye un árbol de decisión a partir del modelo de forma que cada camino del árbol es una regla de clasificación o diagnóstico cuando hablamos de Biología.
- Identificación. Dado un conjunto de casos (el modelo) el sistema selecciona de forma dinámica el criterio más eficiente (utiliza como medida la ganancia de información) y presenta al usuario una lista ordenada de atributos. De su respuesta se seleccionan los casos que concuerden.

ARQUITECTURA.

El sistema fue desarrollado utilizando tres entornos de desarrollo diferentes:

- HyperCard. Para desarrollar las herramientas de adquisición del conocimiento locales. El resultado es HyperQuest, una herramienta visual para la adquisición del conocimiento.

[5] Aclarar que dentro de la Biología, el término *clasificación* atiende a la elaboración de un sistema lógico que agrupe a las plantas próximas entre si desde un punto de vista evolutivo. Las especies similares se agrupan en géneros, estos en familias, órdenes, etc. Las Ciencias de la Computación utilizan este término de forma diferente; en este caso se refieren a la asignación de un conjunto de ejemplos a un conjunto de categorías predefinidas.

- Java. Para que el sistema sea multiplataforma y accesible desde Internet se prevé el desarrollo de una herramienta de adquisición del conocimiento en Java. La comunicación con HyperQuest se realiza con un procesador de lenguajes escrito en JavaCC, un generador de procesadores de lenguajes Java. Este procesador traduce HyperQuest a un lenguaje interno de representación del conocimiento (KRL, *Knowledge Representation Language*). El gestor de intercambio de datos (DEM, *Data Exchange Manager*) coordina los intercambios de datos entre herramientas.

- La base datos orienta a objetos O_2. Proporciona la persistencia de los objetos. Para ello se ha definido un esquema de datos en la base de datos orientada a objetos O_2 equivalente al KRL. La principal tarea del DEM es gestionar el intercambio de datos entre el KRL y el esquema de datos a través de JDBC.

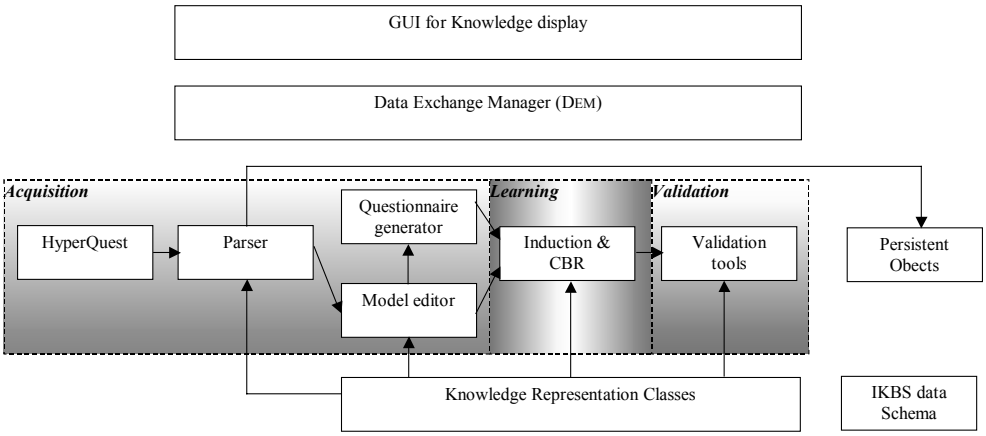


Figura 1-3. *Arquitectura de IKBS In:[Conruyt et al., 1997].*

3.2 *RIH.*

El sistema *RIH* [Grove & Hulse, 1999; Grove, 2000], es un sistema experto para identificar anfibios y reptiles de Pennsylvania (USA) que funciona en Internet. El sistema tiene una estructura cliente-servidor y está escrito en Java, JavaScripts y HTML. El cliente es un navegador web y proporciona la interfaz, el

servidor realiza las labores de sistema experto, gestiona las comunicaciones con el cliente y acceden a la base de imágenes cuando es necesario.

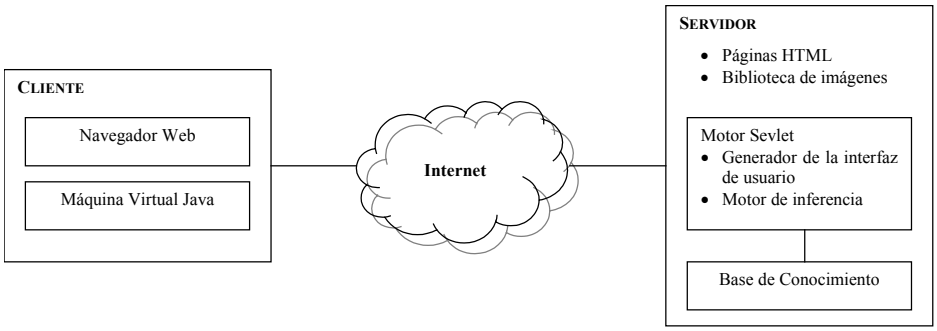


Figura 1-4. *Arquitectura de RIH [Grove, 2000]***.**

Este sistema utiliza un motor de inferencia de libre distribución escrito en Java que implementa el algoritmo RETE [Forgy, 1982]: Jess (*Java Expert System Shell*) [Friedman, 1998] y que toma los datos de una base de datos escrita en CLIPS. En su versión actual está implementado como un Servlet Java. Cuando comienza una sesión de identificación, el motor de Servlets crea un objeto *sesión* que corresponde a la petición del usuario y llama al Servlet requerido. El Servlet se ejecuta y crea diversos objetos, entre ellos un controlador, un motor de inferencia y un generador de salidas de datos.

Para el usuario, una sesión de identificación típica con *RIH* implica contestar a unas diez preguntas eligiendo de dos a cinco en cada paso. Todo el proceso de identificación está guiado mediante imágenes que pueden ser ampliadas para ver detalles adicionales. Tras la consulta, *RIH* presenta al usuario una lista de posibles especies ordenada en función de un factor de parecido que indica cuán seguro estaría el experto de dicha respuesta. Los factores de similitud se determinan por adelantado por el experto en el dominio y se codifican dentro de la base de conocimiento.

Durante la fase de prueba se validaron la interfaz de usuario y el porcentaje de identificaciones correctas realizadas por los usuarios. Estas pruebas mostraron que el sistema puede ser utilizado de forma efectiva con un pequeño adiestramiento sobre la identificación de especies biológicas y las mejoras que podrían hacerse.

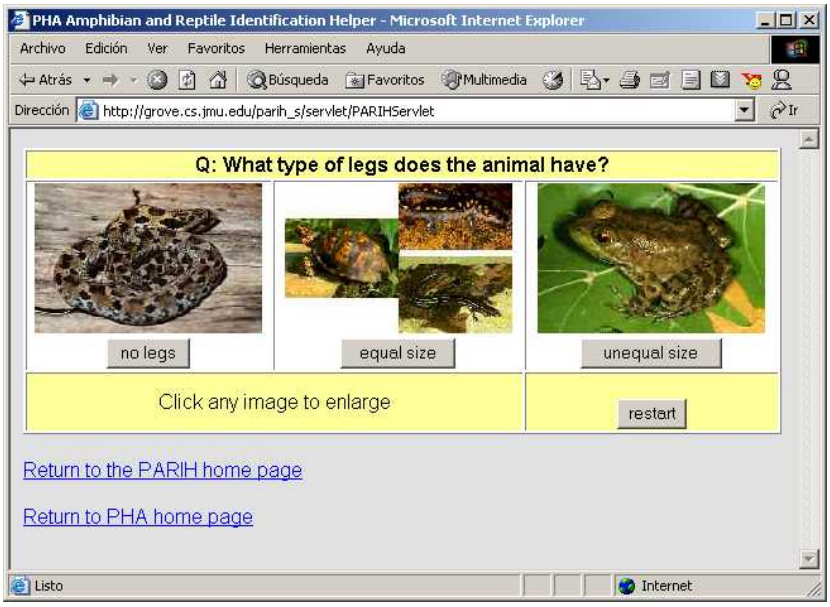


Figura 1-5. *Interfaz del Sistema RIH. Esta interfaz permite mostrar las imágenes en un tamaño mayor siempre a petición del cliente, de forma que se reduce el tráfico de información a través de la red.*

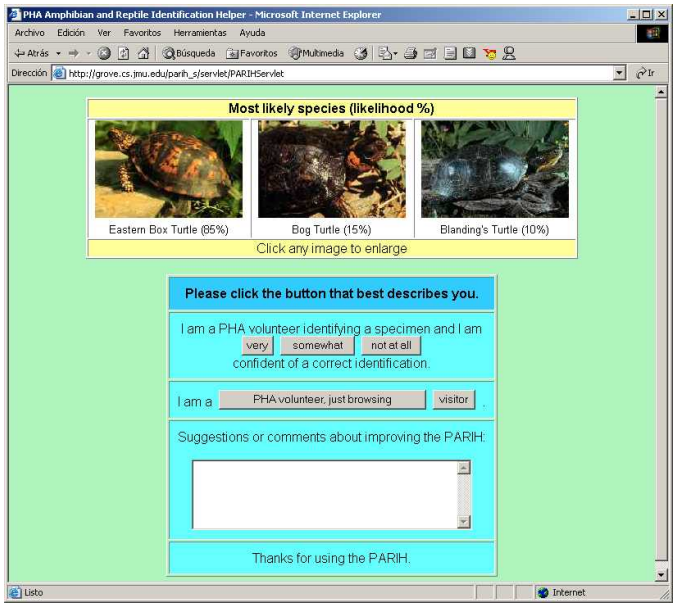


Figura 1-6. *Resultados devueltos por RIH.*

3.3 *Agassistant.*

Agassistant [Fermanian & Michalski, 1989; Fermanian & Michalski, 1992] es uno de los pocos sistemas expertos para Taxonomía vegetal descritos con exhaustividad en la literatura que utiliza algún tipo de inferencia probabilística (el sistema de inferencia fue desarrollado específicamente para tratar la incertidumbre propia del dominio de aplicación). Para ilustrar el uso de *Agassistant* sus autores describen el sistema WEEDER, que identifica especies de gramíneas y determina la extensión de su población.

CARACTERÍSTICAS GENERALES DE *Agassistant.*

A continuación citamos las características más importantes de *Agassistant*:

- Utiliza inferencia probabilística para el manejo de la incertidumbre de los datos y las reglas.
- El sistema puede adquirir conocimiento en forma de reglas de cuatro formas diferentes:
    - A partir de ejemplos. Para ello utiliza el programa NEWGEM [Reinke, 1984] que trabaja intentando encontrar una regla que cubra todos los sucesos positivos (aquellos que pertenecen a la clase bajo consideración) y ninguno de los negativos (el resto). NEWGEM asigna un peso o nivel de confianza (*confidence level*, CL) a cada selector de una regla según la expresión de la Fórmula 1-3. El peso representa la probabilidad de que se trate de una clase determinada dado que el selector se satisface:

$$peso = \frac{pe}{pe + ne}$$

Fórmula 1-3. *Cálculo del peso de los selectores de una regla en el sistema Newgem.*

Donde *pe* es el número de sucesos positivos cubiertos por el selector y *ne* el número de sucesos negativos cubiertos por el selector.

- Mejora de reglas a partir de ejemplos. *Agassistant* es capaz de mejorar su conocimiento a medida que se presentan nuevos ejemplos. Este modo de aprendizaje puede ser activado o desactivado según los deseos del usuario.

- Optimización de reglas. El sistema permite seleccionar un conjunto de reglas y optimizarlas en función de un determinado criterio, por ejemplo, la conversión de reglas de características a reglas discriminantes. Son reglas discriminantes aquellas que permiten distinguir una clase de otra (o de otras), el resto son reglas de características. Esta herramienta es útil para depurar las reglas proporcionadas por los expertos durante la adquisición del conocimiento.

- Edición de reglas. Este enfoque es muy apropiado en algunos dominios de aplicación y es además la única opción que ofrece el programa para adquirir una base de reglas jerárquica. Una vez editadas las reglas, se compila la base de reglas para confirmar que su sintaxis es correcta.

ARQUITECTURA DEL SISTEMA *Agassistant*.

A grandes rasgos, el sistema consta del siguiente conjunto de módulos (ver Figura 1-7):

- Compilador. Comprueba que las reglas tienen una sintaxis correcta y crea una versión compacta del sistema para que la ejecución sea más rápida. Las reglas se pueden crear a mano o pueden ser inducidas a partir de ejemplos.
- Asesor *(advisor)*. Toma como entrada una versión compilada del sistema experto que utiliza para hacer preguntas al usuario y dar un consejo en función de las respuestas a las preguntas.
- Motor de inferencia. Utiliza el programa NEWGEM [Reinke, 1984] para obtener reglas a partir de ejemplos.

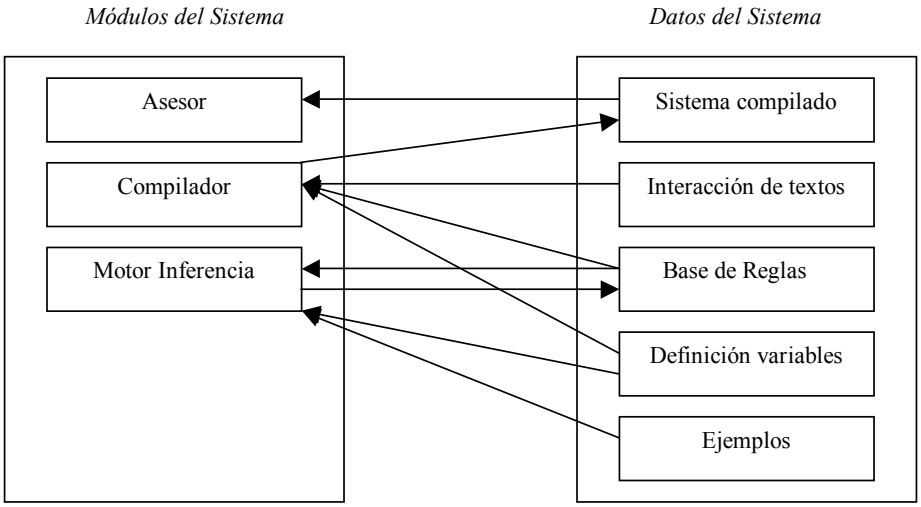


Figura 1-7. *Estructura de bloques del sistema Agassistant [Fermanian & Michalski, 1992].*

REPRESENTACIÓN DEL CONOCIMIENTO Y DE LA INCERTIDUMBRE.

El sistema representa el conocimiento mediante reglas que se pueden estructurar de forma jerárquica, de modo que la condición de una regla sea la conclusión para otra. Como hemos comentado, cada selector de una regla lleva asociado un peso o nivel de confianza (CL) con el que se indica la importancia relativa del selector como condición única para la toma de decisiones (ver Figura 1-8). Los selectores con disyunciones internas se satisfacen si se satisface alguno de los elementos de la disyunción; el grado con que un conjunto de condiciones debe satisfacerse es un umbral del sistema. El sistema trata con datos nominales, lineales, enteros o estructurados. Los selectores con variables lineales se satisfacen si la variable toma un valor dentro del rango definido por los puntos límite (incluidos).

Acción es X si	*Nivel de confianza*	
1. variable$_1$ es valor$_1$	*60*	
2. variable$_2$ es valor$_2$ o valor$_3$	40	Complejo$_1$
3. variable$_3$ es valor$_4$ o valor$_6$	20	
OR		
1. variable$_4$ es valor$_7$	50	Complejo$_2$
2. variable$_5$ es valor$_8$	60	

Figura 1-8. *Ejemplo de regla en Agassistant.*

La expresión general para evaluar un complejo viene dada por la Fórmula 1-4:

$$\frac{\sum(pesos\_selectores\_satisfechos)}{\sum(pesos\_complejo)}$$

Fórmula 1-4. *Expresión para la evaluación de un complejo en Agassistant.*

Supongamos que para el complejo$_1$ de la regla de ejemplo, la variable$_1$ toma el valor$_1$ y la variable$_2$ toma el valor$_3$, entonces tendríamos: $(60+40)/(60+40+20)=83\%$.

La evaluación de reglas con varios complejos se lleva a cabo según la expresión de la Fórmula 1-5, donde *V(x)* es el valor obtenido al evaluar el complejo *x*.

$$psum(V(n),V(n-1),...,V(1)) = V(n) + psum(V(n-1),...,V(1)) - V(n) * psum(V(n-1),...,V(1))$$

Fórmula 1-5. *Expresión para la evaluación de reglas con varios complejos en Agassistant.*

MECANISMO DE CONTROL.

El mecanismo de control de este sistema utiliza dos métodos para determinar qué preguntas se realizan.

- El esquema del control de utilidad. Selecciona las variables por las que preguntar, en función de la utilidad de las mismas. Este valor se calcula al compilar el sistema y refleja el grado con que la variable afectaría al nivel de confianza de todas las reglas. Aquellas variables que aparecen en la mayoría de las reglas tendrán mayor utilidad. Cuando una regla alcanza un nivel de confirmación mayor que un determinado umbral (experimentalmente sobre un 15%), el sistema se centra en dicha regla y continua preguntando por valores de variables de dicha regla hasta que la regla se rechaza o se agotan las variables de dicha regla.
- El esquema de control de vuelta atrás (*backtracking).* Se invoca de forma automática si la base de reglas es jerárquica. El sistema comienza preguntando

por variables de mayor utilidad hasta que el usuario responde que no conoce el valor de una variable y dicha variable aparece como consecuente de otra regla. En este punto el sistema intenta inferir dicho valor a partir de la información contenida en la base de conocimiento.

EL SISTEMA WEEDER.

Para preparar el conocimiento recogido por el sistema se elaboró una matriz de datos a partir de libros de texto, manuales de identificación y experiencia de los autores. Cada selector de una regla tiene asociado un nivel de confianza que fue definido por el experto en el dominio. Las reglas se obtuvieron utilizando el módulo NEWGEM de *Agassistant*. El sistema incluye reglas especiales que desactivan determinados caracteres considerándolos no aplicables cuando se dan determinadas condiciones. Un ejemplo puede ser:

IF seed heads are not present

THEN Florests, Flower, Awns, Disart, and Glumes do not apply

EVALUACIÓN DE WEEDER.

Para la evaluación de Weeder se seleccionaron 41 voluntarios y se separaron en dos grupos. Uno de los grupos estaba formado por voluntarios con experiencia previa y el otro grupo estaba formado por voluntarios sin experiencia. Se seleccionaron también 4 especies de gramíneas a identificar (de un conjunto de 15). Para hacer esta prueba se dotó a cada voluntario del equipo necesario para desarrollar la identificación (lupa, dibujos diagnóstico, etc.). El tiempo máximo para realizar la prueba era de 30 minutos por ejemplar. Este estudio revela que, mientras que el conocimiento es la piedra angular en los sistemas de ayuda a la decisión, cuando hablamos de identificación se añade como aspecto primordial: la habilidad para reconocer el valor de los atributos por los que el sistema pregunta. Se evidencia que el desarrollo de técnicas para mejorar estas habilidades en el usuario es fundamental para el incremento de la efectividad del sistema.

3.4 Sinopsis.

La información taxonómica es difícil de representar en un ordenador puesto que nos encontramos con [Conruyt *et al.,* 1997]:

- Polimorfismo. Por ejemplo, cuando hablamos de variabilidad intra-específica.
- Relaciones de dependencia entre atributos, objetos y valores.
- Dificultad de disponer de toda la información requerida cuando se consulta la base de conocimiento (las muestras a identificar no suelen estar completas).
- Atributos estructurados. Por ejemplo, para la hoja podríamos hablar de la forma, el color, la pilosidad, etc.

Las preguntas e ilustraciones utilizadas por los expertos al observar objetos biológicos son otra fuente de dificultad en las descripciones pues es común la aparición de errores de interpretación debido en parte a [Conruyt *et al.*, 1997]:

- La falta de explicaciones del experto. *¿Dónde observar?*
- La mala definición de los valores. *¿Son mutuamente excluyentes?*
- La falta de un marco de referencia. *¿Qué es grueso y qué es delgado para el diámetro de un tallo?*

Son muy pocos los sistemas expertos para identificación taxonómica desarrollados y bien documentados. *IKBS* es uno de los pocos sistemas expertos *sensu-estricto*, correctamente documentados que intentan solucionar los problemas intrínsecos de la identificación taxonómica. No obstante, obvia dos aspectos importantes:

- No utiliza un modelo de datos estándar que le facilite compatibilidad con otras aplicaciones.
- No contempla la reutilización de la información de su modelo taxonómico ni la integración con otro tipo de información (por ejemplo, referencias bibliográficas) y aplicaciones (generación de descripciones).

	IKBS	***RIH***	***Agassistant***
PUBLICACIÓN	Conruyt N., Grosser D, Ralambondrainy H.. 1997. IKBS: An Iterative Knowledge Base System for improving description, classification and identification of biological objects. In *Proceedings of the Indo-French Workshop on Symbolic Data Analysis and its Applications*, París (IFWSDAA'97).	Grove R. F. 1999. An Internet-Based Expert System for Reptile Identification. The First International Conference on the Practical Application of Java, London, UK. 165-173.	Fermanian T., Michalski R.S. 1992. Agassistant: A new generation tool for agricultural advisory systems. *Expert Systems in Developing Countries, practice and promise*. Chapter 5. Westview Press.
AÑO	1997.	2000.	1992.
CENTRO	Université de la Reunión.	Indiana University of Pennsylvania.	Departamento de Ciencias de la Computación de la universidad de Illinois.
CAMPO DE APLICACIÓN	Identificación de corales.	Anfibios y reptiles de Pennsylvania.	Identificación de pastos.
REPRESENTACIÓN DEL CONOCIMIENTO	Modelo en forma de árbol, base de casos y reglas.	Reglas (1200).	Reglas.
CONSTRUCCIÓN BC	A partir del modelo en árbol, el programa genera de forma automática un cuestionario que facilita al experto crear una base de casos a partir de la cual se generan reglas mediante la aplicación de un C4.5	No se documenta correctamente, pero la descripción del sistema apunta hacia la construcción manual de la base de conocimiento.	• Aprendizaje a partir de ejemplos. • Mejora de reglas a partir de ejemplos. • Optimización de reglas. • Edición de reglas.
TRATAMIENTO DE INCERTIDUMBRE	Tratamiento de atributos multievaluados.	Cada resultado se etiqueta con un factor de parecido que indica como de seguro se siente el experto ante la identificación	Peso asociado (de forma automática o manual) a cada uno de los selectores de una regla
MODELO DE INFERENCIA	No se documenta correctamente, pero la descripción del sistema apunta hacia el modelo hacia delante.	Jess, que permite razonamiento hacia delante y hacia atrás.	Hacia delante

Tabla 1-5. *Características de los sistemas IKBS, RIH y Agassistant.*

4. Sistemas basados en matrices.

Este apartado está dedicado al estudio de las herramientas desarrolladas por la comunidad de biólogos para resolver el problema de la identificación taxonómica. Estas herramientas tienen un factor común, todas organizan el conocimiento en forma matricial. En concreto, nos detendremos en el análisis de las características de los sistemas enumerados en la Tabla 1-6.

LucId	*Linnaeus II.*
XID	*Meka*
Navikey	*Pollyclave*
Nexus	*Pankey*
Pankey	*X:ID*
Delta	*DeltaAccess*

Tabla 1-6. *Algunos sistemas para identificación taxonómica basados en matrices.*

4.1 *LucID.*

LucID [CBIT, 1994] es un sistema comercial de utilización sencilla, diseñado específicamente para la identificación de especímenes y la edición rápida de claves dicotómicas. Se han desarrollado conjuntos de datos para los órdenes de insectos [University of Queensland, 1999], y libélulas del mundo [Silsby & Trueman, 2002]. *LucID* también ha sido utilizado para resolver otros problemas de diagnóstico:

- Medicina: identificación y diagnóstico de úlceras orales [Forrest & Walsh, 2000].
- Agricultura: insectos del arroz [Josie Lynn Catindig, 2001].
- Geología: clave de minerales, [Golding, 2000].

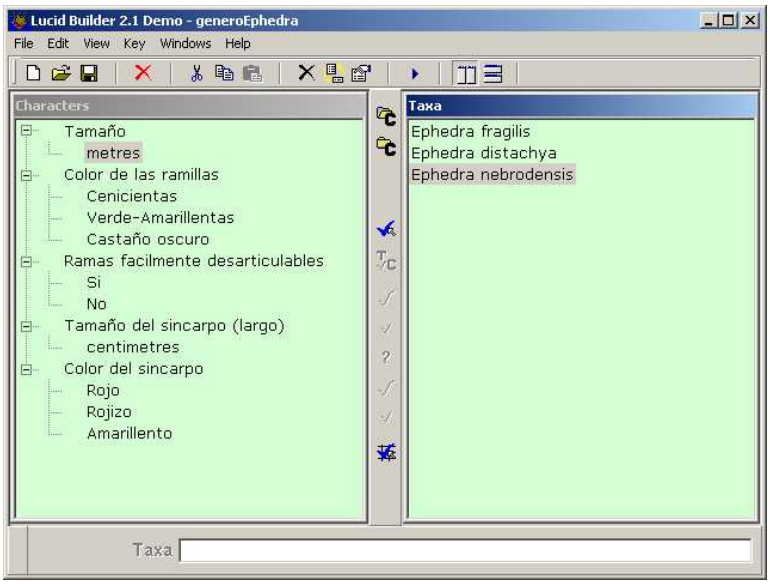


Figura 1-9. *Interfaz de LucID Builder.*

Consta de dos módulos (Figura 1-9, Figura 1-10):

- LUCID BUILDER. Para construir o modificar claves de identificación.
- LUCID PLAYER. Es un cliente que permite a los usuarios la visualización de claves *LucID*. Para ayudar al usuario, estas claves pueden incorporar texto, imágenes, vídeo y sonido. Durante el proceso de identificación el usuario selecciona estados de caracteres y se van eliminando aquellos *taxa* a los que no les pueden ser aplicados.

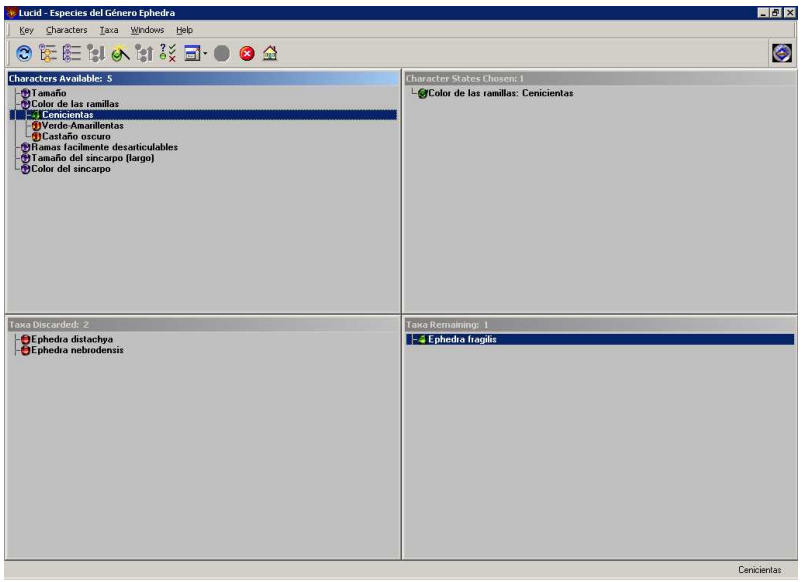


Figura 1-10. *Interfaz LucID Player.*

Entre sus características más importantes destaca su capacidad de:

- Seleccionar los atributos a utilizar durante la identificación en cualquier orden. Recordemos que el orden fijo en la selección de atributos es uno de los mayores problemas de las claves dicotómicas.
- Presentar al usuario los caracteres más prometedores, permitiendo que en caso de disponer de dicha información, la identificación sea más rápida.
- Modelar dependencias entre atributos.
- Representar caracteres numéricos. Cuando un atributo puede tomar valores numéricos, se representa mediante cuatro valores: Mínimo y máximo a lo sumo (es extraño que el atributo tome estos valores, pero es posible) y mínimo y máximo normal (el rango de valores habitual para el carácter).
- Seleccionar varios estados para un mismo taxon y un mismo carácter.
- Dar valores especiales a los estados de los caracteres diferenciando entre: "*presente de forma habitual*", "*raramente presente*", "*desconocido o incierto*", *"raramente presente y susceptible de errores de interpretación*", "*presente de forma habitual pero susceptible a errores de interpretación*".
- Asignar a un determinado taxon una subclave (por ejemplo, una clave de subespecies). De esta forma permite la representación de claves tanto en formato plano (el formato más habitual cuando hablamos de sistemas basados en matrices de datos) como en formato jerárquico.
- Crear conjuntos de caracteres para facilitar al usuario la selección de los mismos. Un mismo carácter puede pertenecer a más de un conjunto.
- Asociar texto, video o imágenes a los *taxa* y a los estados de los caracteres.
- Procesar y preparar la clave para su visualización desde un CD-ROM o en Internet.
- Traducir ficheros de datos en formato *Delta* a ficheros en formato *LucID* y viceversa. Esta traducción tiene sus limitaciones y solo implica a la matriz de datos y la lista de *taxa* y no a los conjuntos de caracteres, dependencias e información adjunta.

4.2 *Linnaeus II.*

Linnaeus II [Schalk & Troost, 1999; Schalk & Heijman, 1996] ha sido desarrollado por la ETI (*Expert Center for Taxonomic Identification*), una fundación sin ánimo de lucro asociada a la UNESCO dedicada a la mejora de la cantidad, calidad y accesibilidad de la información taxonómica. Para esto desarrolla software científico y educativo. El sistema *Linnaeus II* es uno de sus productos y ha sido construido por y para taxónomos y expertos en biodiversidad (se consultó a cientos de especialistas para construir esta herramienta).

Linnaeus II permite crear bases de datos taxonómicas, optimiza la construcción de claves de identificación, acelera la muestra y comparación de patrones de distribución y promueve el uso de información taxonómica para estudios de biodiversidad. Toda esta información sobre biodiversidad se distribuye en CD-ROMS para científicos, estudiantes, taxónomos, gestores medioambientales, etc. Ya se han publicado más de 70 que incluyen información sobre diferentes grupos taxonómicos y áreas geográficas, por ejemplo, sobre hongos micorrizas [Dodd & Rosendahl, 1996].

El sistema *Linnaeus II* consta de 4 módulos (Figura 1-11):

- Bases de datos taxonómicas. Se divide en tres módulos:
    - El *módulo de especies* es el más importante. Contiene texto e información multimedia sobre especies de un determinado grupo taxonómico: descripciones, sinónimos, nombres vulgares, información taxonómica, referencias literarias, fotografías, dibujos, audio y vídeo, etc.
    - El módulo de *mayor nivel taxonómico* almacena información de *taxa* a un nivel taxonómico mayor que el de especie mientras que el módulo de *menor nivel taxonómico* almacena información de subespecies y otros *taxa* de menor nivel taxonómico que especie.
- Bases de datos de apoyo. Almacenan información adicional como una introducción general sobre el grupo taxonómico, glosario, referencias bibliográficas, lista y jerarquía de *taxa*.

- Módulos de identificación. *Linnaeus II* incorpora tres módulos diferentes para la identificación.
    - *Text Key*™. Es una versión electrónica de claves dicotómicas en el sentido clásico.
    - *Picture Key*™. Similar a *Text Key*™ salvo que la identificación está basada en imágenes y no en texto.
    - *IdentifyIt*™. Es una clave multientrada basada en una matriz de *taxa*, caracteres y estados de caracteres. La versión actual es capaz de calcular el mejor carácter a añadir al patrón de búsqueda, permite utilizar caracteres numéricos e importa / exporta ficheros en formato *Nexus* [Maddison *et al.,* 1997]. También devuelve una medida del grado de emparejamiento del patrón de búsqueda con los posibles resultados.
- Sistema de información biogeográfica. *MapIt*™ es un sistema que permite introducir información geográfica como distribuciones y tipos de localidades sobre especies y otros *taxa*. Esta información se puede utilizar, por ejemplo, para buscar qué especies están presentes en una región determinada, comparar la distribución de unas especies con otras o mostrar la riqueza específica (*species-richness*).

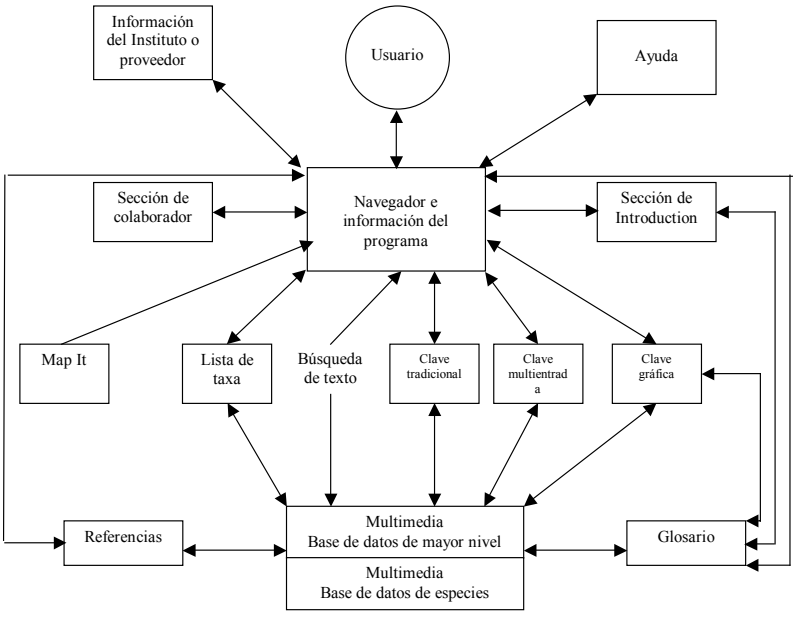


Figura 1-11. *Estructura modular del programa LinnaeusII [Schalk & Heijman, 1996]*.

A pesar de su fina apariencia, el proceso de identificación de ETI se basa en un enfoque dicotómico, con todas las limitaciones que esto implica [Diederich *et al.*, 2000].

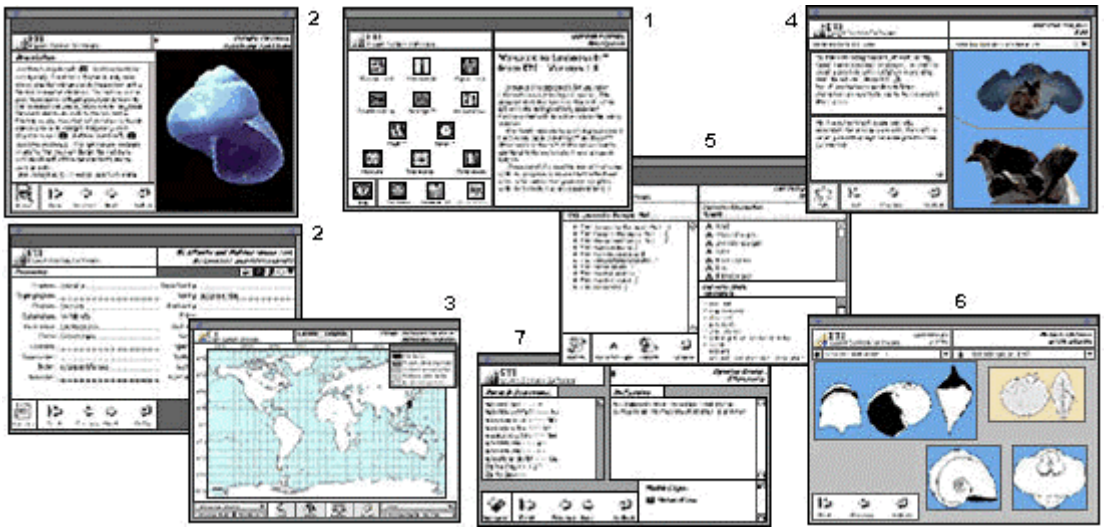


Figura 1-12. *Interfaz de Linnaeus II: Navegador (1), parte de la base de datos taxonómica (2), sistema de información geográfica (3), glosario y claves de identificación tradicional (4), IdentifyIt (5), clave de identificación basada en imágenes (6).*

4.3 *XID.*

XID [Intelsys Inc., 2000-2001] es un sistema comercial para problemas de identificación en general. Entre sus características destacamos que:

- Permite la identificación sin necesidad de seguir un orden determinado en la selección de caracteres.
- Ordena los caracteres a escoger en función de su utilidad.
- Permite visualizar simultáneamente las ilustraciones de todos los *taxa* candidatos y de esta forma se pueden comparar fácilmente. El usuario puede determinar el número y el tamaño de las imágenes a visualizar.
- Muestra las características seleccionadas durante el proceso de identificación de forma que permite seguir una traza del proceso seguido.
- Presenta informes sobre la frecuencia relativa de los atributos.

- Incluye gráfico y texto informativo para cada especie y para cada atributo.
- Este sistema no permite la identificación a través de Internet.

La Figura 1-13 ilustra la interfaz de *XID*. En el panel izquierdo superior se encuentran todos los atributos que pueden ser utilizados para la identificación. Para cada atributo, el usuario puede seleccionar uno o varios valores indicando si dicho valor está presente o ausente en la observación o si ese valor no se presenta nunca. En el panel izquierdo inferior el sistema muestra los taxones descartados por la información introducida y aquellos que todavía permanecen como posibles candidatos. En el panel derecho se muestran los atributos más prometedores durante el proceso de identificación o bien los resultados alcanzados si este proceso ha concluido satisfactoriamente.

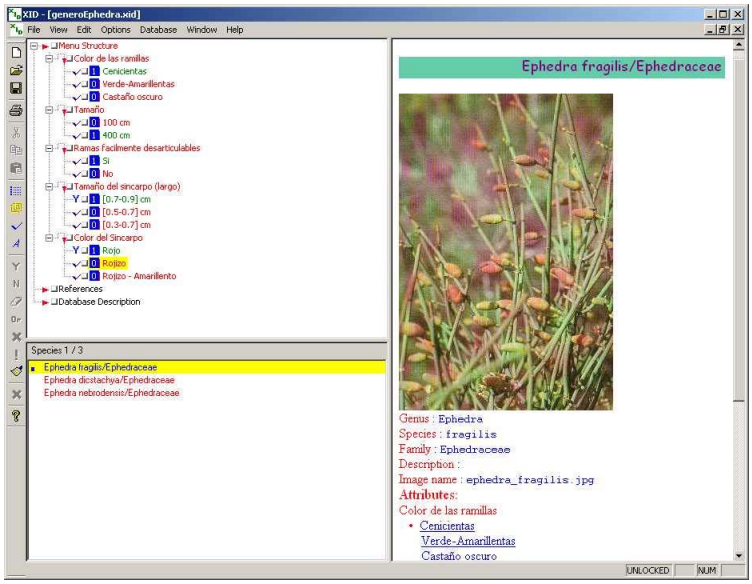


Figura 1-13. *Interfaz de* XID *tras haber realizado un proceso de identificación en el que el sistema ha concluido que la observación realizada corresponde a un individuo de la especie Ephedra fragilis.*

4.4 *Meka.*

Meka [Duncan & Meacham, 1986; Meacham, 1986-1996] es un sistema de claves multientrada de libre distribución. Se puede descargar desde "*http://ucjeps.berkeley.edu/meacham/Meka/*".

Su funcionamiento es similar al de los sistemas basados en matrices analizados en apartados anteriores. El usuario selecciona, de una lista de posibilidades, los estados de caracteres presentes en el especimen. Al seleccionar dichos caracteres *Meka* elimina los *taxa* que no concuerdan con los datos introducidos y los estados que no sirven para distinguir entre los *taxa* que quedan. Al tratarse de una clave multientrada, se puede llevar a cabo la identificación sin necesidad de introducir la información en un orden predeterminado.

La Figura 1-14 ilustra la interfaz de este sistema. Observamos, de izquierda a derecha, una lista de caracteres (*Score Character States*), una lista con los caracteres que son todavía útiles para la identificación (*Differentiating Character States*), una lista de *taxa* candidatos (*Matching Taxa*) y una lista con los *taxa* descartados (*Mismatched Character States*).

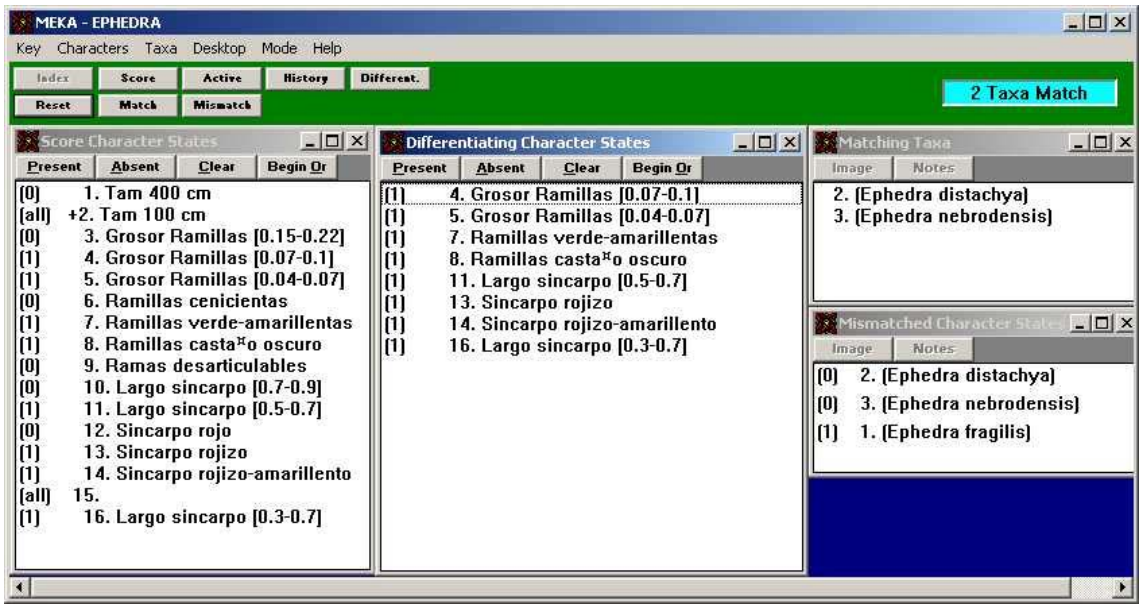


Figura 1-14. *Interfaz del sistema Meka.*

A menudo, al hacer identificaciones sobre un mismo taxon o grupo de *taxa* se hace necesaria la selección de los mismos caracteres una y otra vez. *Meka* conserva un histórico de los caracteres que han sido seleccionados para facilitar esta tarea.

Cada estado correspondiente a un carácter puede tomar cuatro valores: presente ("+"), ausente ("-"), desconocido ("*?*") o lo pueden presentar algunos miembros del taxon pero no todos ("*"). Esto quiere decir que *Meka* no distingue entre carácter y estados o valores que toma dicho carácter. Este aspecto es una fuente de ineficiencia e impide disponer de cualquier tipo de atributo diferente de los de tipo presencia / ausencia.

Otro problema es el derivado del editor de claves *Meka*edit, única forma de generar claves para *Meka*etit que solo está disponible para MS-DOS.

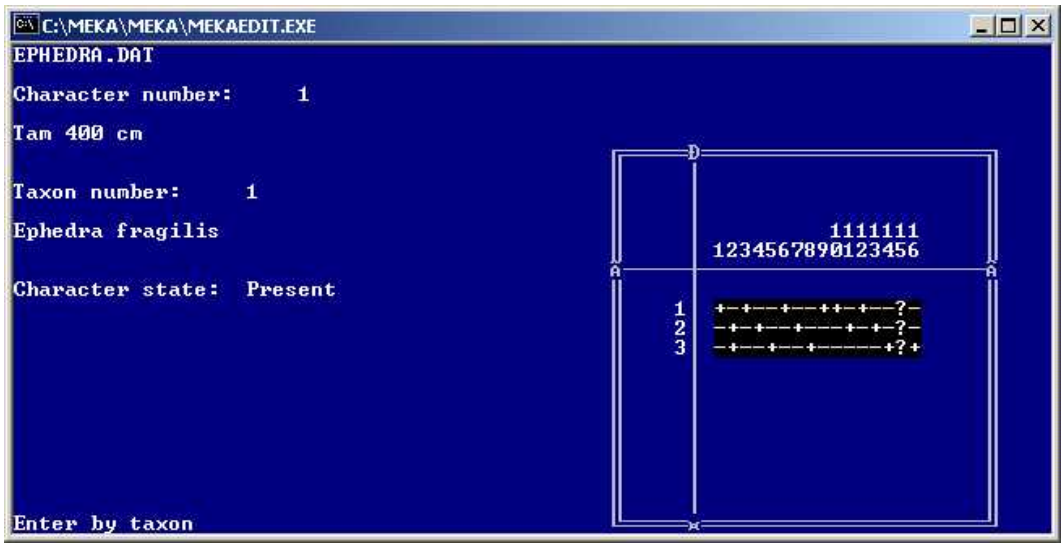


Figura 1-15. *Interfaz de Mekaedit.*

4.5 *Navikey.*

NaviKey [Bartley & Cross, 2000] es un Applet Java que actúa como interfaz web para ficheros *Delta*. La interfaz gráfica contiene cuatro paneles:

- Panel de caracteres, izquierda arriba.
- Selecciones que se han realizado, izquierda abajo.
- Estados del carácter seleccionado, derecha arriba. *NaviKey* solo muestra aquellos caracteres y estados que pueden reducir el número de candidatos. Los caracteres o estados no aplicables a los *taxa* candidatos tampoco se muestran. Esta característica hace que *NaviKey* funcione más lentamente, por lo que puede ser activada / desactivada a voluntad del usuario.
- *Taxa* candidatos, derecha abajo.

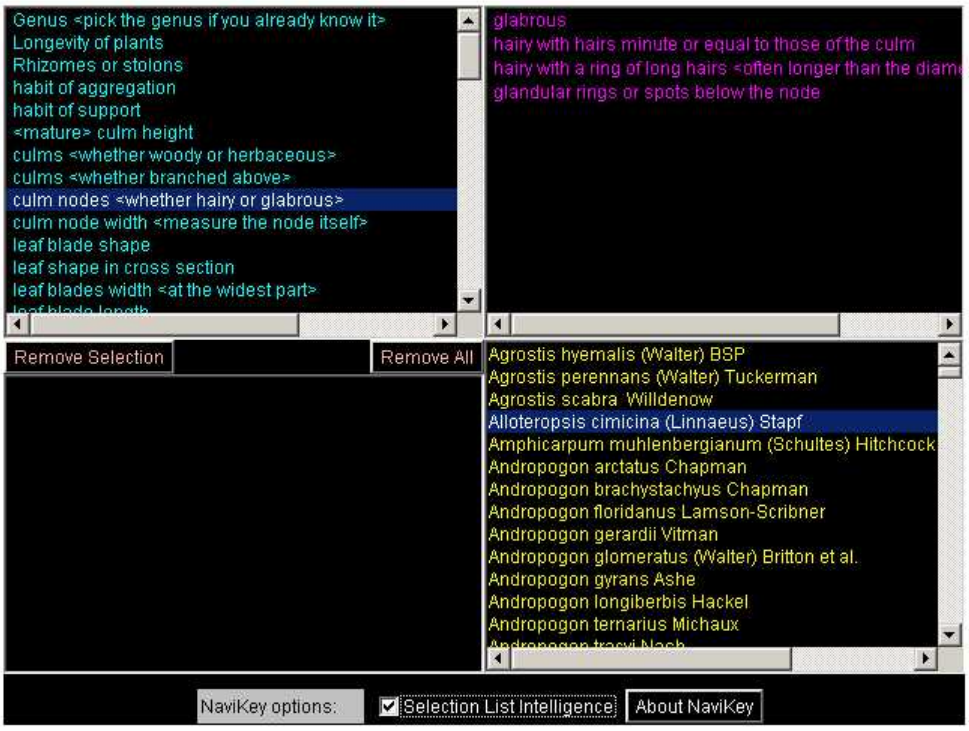


Figura 1-16. *Interfaz de Navikey[Guala, 1999].*

NaviKey es capaz de tratar con valores numéricos permitiendo la introducción de un valor o de un rango de valores. Funciona a partir de ficheros de texto escritos en formato *Delta*. Una vez leída la descripción en *Delta*, los datos son almacenados en un servidor SQL 7.0 con una estructura basada en la de *DeltaAccess* [Hagedorn, 1995-2003]. Las localizaciones de las imágenes se almacenan en una base de datos independiente, unida a la base de datos de caracteres a través de una base de datos de tesauros taxonómicos. Al seleccionar

cualquier taxon podemos acceder a una completa descripción del mismo, así como a enlaces a imágenes y otros recursos sobre el taxon disponibles.

4.6 *Pollyclave.*

Pollyclave [University of Toronto. Department of Botany *et al.*, 1996 a] es una clave de identificación multientrada que interactúa directamente con ficheros de datos escritos en formato *Delta*. Es una herramienta de libre distribución para investigación y uso académico. Se trata de un programa CGI, escrito en ANSI C, desarrollado para facilitar que los usuarios de *Delta* hicieran accesibles sus datos en Internet. Actualmente se encuentra en su versión 2.0 desarrollada en el año 2002.

El proceso de identificación comienza presentando al usuario una lista de caracteres, que tras ser rellenada con la información que se conoce es enviada al servidor (Figura 1-17).

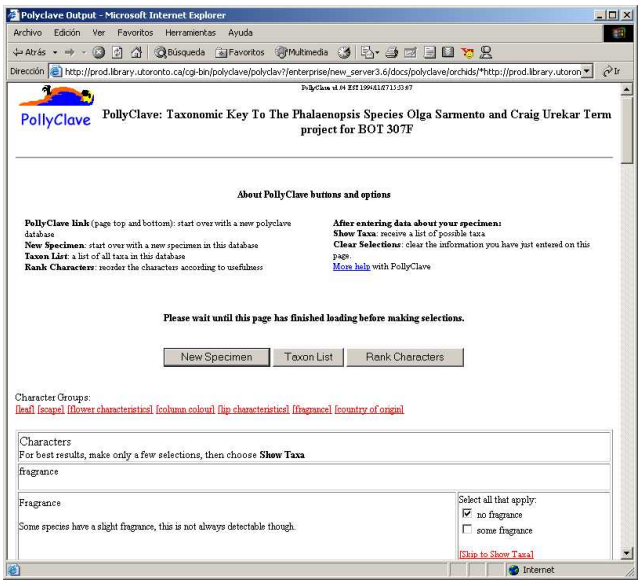


Figura 1-17. *Introducción de información en PollyClave [University of Toronto. Department of Botany et al., 1996 b].*

El servidor devuelve el conjunto de *taxa* que presentan las características que el usuario había seleccionado. El usuario puede seguir añadiendo información al patrón de búsqueda hasta que concluya la identificación. *Pollyclave* presenta al usuario los atributos ordenados en función de su utilidad (Figura 1-18), esta utilidad es un valor real calculado según el criterio descrito por Dallwitz en [Dallwitz *et al.*, 2000].

Cuando el sistema devuelve varios *taxa* hace una comparación de los mismos. Para esto comprueba si algún estado seleccionado por el usuario es diferente para alguno de los *taxa* seleccionados y lo presenta en negrita (Figura 1-19).

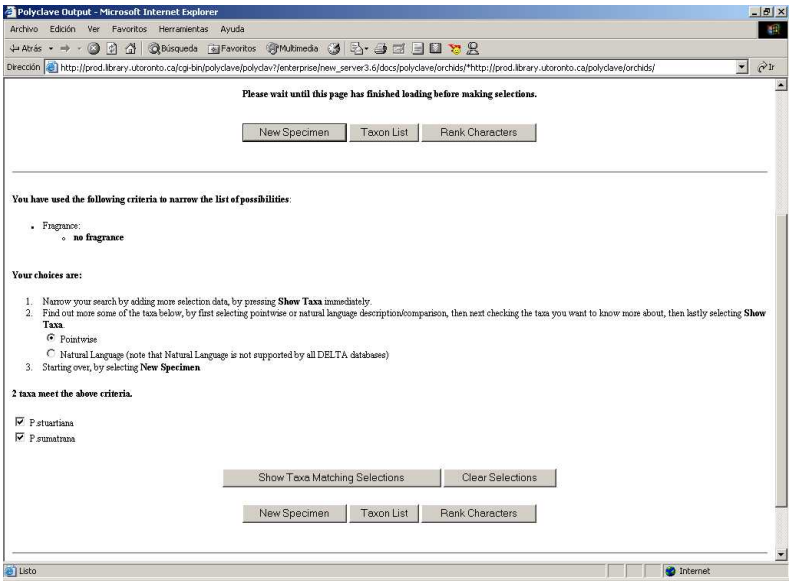


Figura 1-18. *Resultados devueltos por PollyClave [University of Toronto. Department of Botany et al., 1996 b].*

P.stuartiana
[pstuart.gif]

- Leaf pattern [Illustration: [char1a.gif]] :
  - **mottled**
- Scape orientation [Illustration: [char2a.gif]] :
  - upright
- Scape, relative position of bracts [Illustration: [char3a.gif]] :
  - small, far apart from one another
- Scape, number of flowers open at one time [Illustration: [char4a.gif]] :
  - **)5**
- Flower, colour of background :
  - white
- Flower, presence of markings [Illustration: [char6a.gif]] :
  - **spotted on the interior half of the lateral sepals**
  - **colour on the interior half of the lateral sepals**
- Flower, colour of the markings :
  - red to brown
- Flower, sepal to petal ratio [Illustration: [char8a.gif]] :
  - **petals wider than the sepals**
- Column colour [Illustration: [char9a.gif]] :
  - white
- Lip appendages [Illustration: [char10a.gif]] :
  - **horns**
- Lip pubescence [Illustration: [char11a.gif]] :
  - **not present**
- Lip, location of pubescence [Illustration: [char12a.gif]] : **Inapplicable**
- Lip, colour of background :
  - white
- Lip markings [Illustration: [char14a.gif]] :
- Lip, colour of markings :
- Fragrance :
  - no fragrance
- Country of origin :
  - **Philippines**

P.sumatrana

- Leaf pattern [Illustration: [char1a.gif]] :
  - **green**
- Scape orientation [Illustration: [char2a.gif]] :
  - upright
- Scape, relative position of bracts [Illustration: [char3a.gif]] :
  - small, far apart from one another
- Scape, number of flowers open at one time [Illustration: [char4a.gif]] :
  - **1-5**
- Flower, colour of background :
  - white
  - **yellow**
  - **green**
- Flower, presence of markings [Illustration: [char6a.gif]] :
  - **barred**
- Flower, colour of the markings :
  - red to brown
- Flower, sepal to petal ratio [Illustration: [char8a.gif]] :
  - **petals equal in width to the sepals**
- Column colour [Illustration: [char9a.gif]] :
  - white
- Lip appendages [Illustration: [char10a.gif]] :
  - **no horns**
- Lip pubescence [Illustration: [char11a.gif]] :
  - **dense**
- Lip, location of pubescence [Illustration: [char12a.gif]] :
  - **both sides and middle**
- Lip, colour of background :
  - white
- Lip markings [Illustration: [char14a.gif]] :
- Lip, colour of markings :
- Fragrance :
  - no fragrance
- Country of origin :
  - **Brunei Darussalam**
  - **Indonesia**
  - **Malaysia**
  - **Myanmar**
  - **Thailand**
  - **Vietnam**

Figura 1-19. *Comparación de taxa con PollyClave [University of Toronto. Department of Botany et al., 1996 b].*

4.7 *Pankey.*

Pankey [Pankhurst, 1991; Pankhurst, 1994] es un sistema comercial, disponible para Ms-DOS, que consta de un conjunto de programas basados en el estándar *Delta* v3 para problemas de identificación y diagnóstico:

- *Key3m3*. Es un programa dedicado a la generación automática de claves, de forma que la única interacción con el programa es el fichero de datos.
- *Kconi*. Es un programa dedicado a la generación interactiva de claves. Muestra la clave parcial resultado de la elección de caracteres realizada por el usuario y le sugiere la mejor opción en cada paso.
- *Onlin7*. Esta herramienta permite llevar a cabo una identificación interactiva en la que en cada paso se selecciona un carácter y se introduce el correspondiente valor. Este programa permite además configurar el número de caracteres erróneos que se permiten para dar por buena una identificación, conocer los caracteres necesarios para reconocer a un taxon concreto y es capaz de asesorar al usuario sobre el carácter más prometedor en cada paso.

- *Match*. Este programa no lleva a cabo un proceso de eliminación por pasos sino que calcula un coeficiente de similitud entre la descripción completa de un especimen y cada taxon almacenado por el sistema.
- *Spd1*. Herramienta que genera todas las posibles combinaciones de longitud mínima de caracteres diagnóstico que permiten diferenciar a un taxon dado.
- *Dedit-Delta editor*. Es un programa editor de *Delta* (ver Figura 1-20).
- *Chanal*. Este programa proporciona un medio para calcular correlaciones entre caracteres. Esta prueba es aconsejable porque un carácter *Delta* puede contener redundancia. Al no contar con variables estrictamente cuantitativas, esta correlación no se calcula en sentido estadístico, sino que utiliza un enfoque basado en la cantidad de información. Si este valor se aproxima a la unidad los dos caracteres pueden ser dos formas de decir lo mismo (redundantes) o un signo de agrupamiento taxonómico significativo.

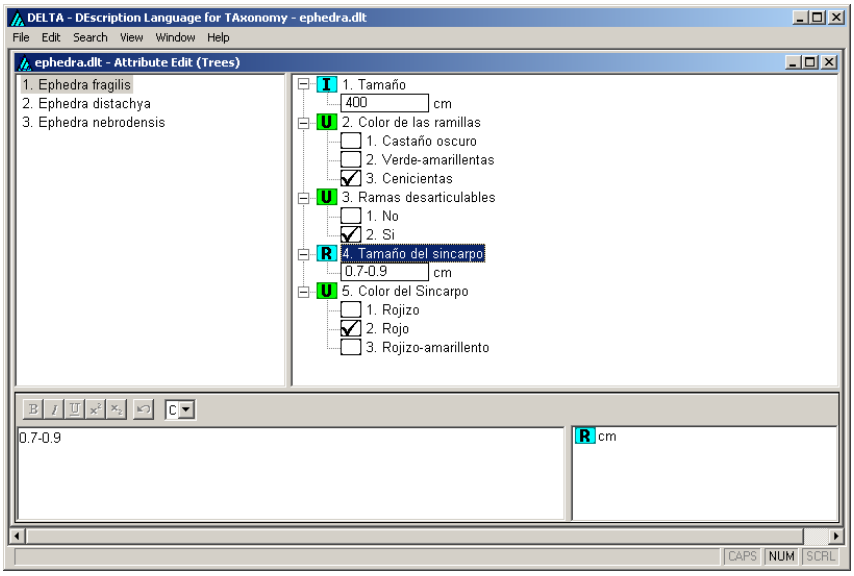


Figura 1-20. *Editor de descripciones en formato Delta del sistema Delta para Windows.*

4.8 El sistema *Delta*.

Es un sistema general para procesar descripciones taxonómicas basado en el modelo *Delta*, que describiremos en el Capítulo 2.1.1. Está formado por un conjunto de programas que implementan utilidades dentro del campo de la

Sistemática para: edición y generación de descripciones taxonómicas y claves, conversión de datos, identificación interactiva y recuperación de información taxonómica (ver Figura 1-21).

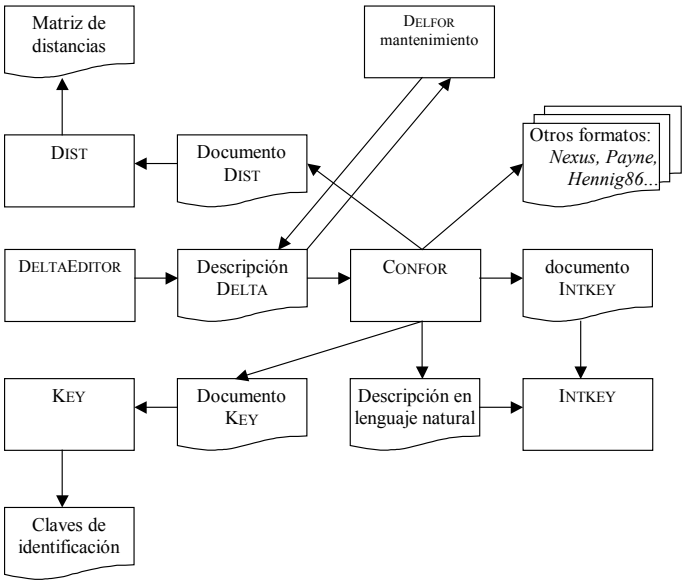


Figura 1-21. *Esquema del sistema Delta. A partir de la descripción en formato Delta, Confor elabora las distintas entradas de datos para el resto de los módulos del sistema.*

El programa Confor.

Traduce descripciones taxonómicas escritas en formato *Delta* a otros formatos como son:

- Lenguaje natural. Permite pasar de *Delta* a descripciones en lenguaje natural en formato RTF (Figura 1-22).
- *Key*: El formato del programa de generación de claves *Key*.
- *Intkey*: El formato para el programa de identificación taxonómica *Intkey*.
- *Dist*: El formato para el cálculo de distancias entre taxones.
- *Nexus*: Un modelo para representación de conocimiento taxonómico, utilizado por otros programas como *MacClade* [Maddison & Maddison, 1992] y *Paup* [Swofford, 1991] (Figura 1-23).
- *Hennig86*. Un modelo para representación de conocimiento filogenético utilizado por el programa de análisis filogenético *Hennig86* [Farris, 1988].

- *Payne*. El formato utilizado por el programa de generación de claves *Genkey* [Payne, 1975].

Descriptions

Ephedra fragilis

Tamaño 400 cm. Color de las ramillas Cenicientas. Ramas desarticulables Si. Tamaño del sincarpo 0.7–0.9 cm. Color del Sincarpo Rojo.

Ephedra distachya

Tamaño 100 cm. Color de las ramillas Verde-amarillentas. Ramas desarticulables No. Tamaño del sincarpo 0.5–0.7 cm. Color del Sincarpo Rojizo.

Ephedra nebrodensis

Tamaño 100 cm. Color de las ramillas Castaño oscuro. Ramas desarticulables No. Tamaño del sincarpo 0.3–0.7 cm. Color del Sincarpo Rojizo-amarillento.

Figura 1-22. *Descripción en lenguaje natural realizada por el programa Confor a partir de una descripción en formato Delta.*

```
#NEXUS

BEGIN DATA;
DIMENSIONS NTAX=3 NCHAR=3;
 [! ]
FORMAT MISSING=? GAP=- SYMBOLS="123";

CHARLABELS
[1(2)] 'Color de las ramillas'
[2(3)] 'Ramas desarticulables'
[3(5)] 'Color del Sincarpo'
;

STATELABELS
1 'Castaño oscuro' 'Verde-amarillentas' 'Cenicientas',
2 'No' 'Si',
3 'Rojizo' 'Rojo' 'Rojizo-amarillento',
;

MATRIX
'Ephedra fragilis'              322
'Ephedra distachya'              211
'Ephedra nebrodensis'            113
;

END;

BEGIN ASSUMPTIONS;
OPTIONS DEFTYPE=unord PolyTCount=MINSTEPS;
TYPESET * untitled = unord: 1-3;

WTSET * untitled = 1: 1-3;

END;
```

Figura 1-23. *Descripción en formato Nexus realizada por el programa Confor a partir de una descripción en formato Delta.*

El programa Delfor.

Permite reestructurar datos y directivas de los documentos *Delta* y cambiar el orden en el que aparecen los caracteres en la lista de caracteres (y por tanto en las descripciones de lenguaje natural). Estas operaciones se deben realizar rutinariamente tras realizar cambios considerables en los datos.

```
*SHOW Tidy the main data files (using Delfor - cannot be used with Confor).

*LISTING FILE tidy.lst

*COMMENT This directives file cannot be used with CONFOR.

*INPUT FILE specs

*OUTPUT WIDTH 80

*INPUT DELTA FILE confor.cph

*REFORMAT specs
*REFORMAT chars
*REFORMAT items
```

Figura 1-24. *Fichero de configuración para el programa Delfor. Este fichero contiene un conjunto de directivas que indican a Delfor como debe operar.*

El programa Key.

Es un programa para construir claves de identificación. Para ello, construye un árbol de decisión e incorpora aspectos adicionales para que las claves estén más adaptadas a la realidad. Entre estos aspectos destacan:

- **Control del efecto de la utilidad de los caracteres**. El parámetro *RBASE* controla este aspecto y toma valores en el intervalo [1, 5]. El coste *c* de un carácter, está relacionado con la utilidad (*reliability*) mediante la expresión $c = \mathrm{Rbase}^{5-r}$, donde *r* es el valor asignado a la utilidad del carácter. Así si *RBASE* =1, todos los caracteres tendrán la misma influencia en la formación de la clave.

- **Control del efecto de la variabilidad intra-taxon**. El valor del parámetro *VARYWT* controla el efecto de la variabilidad intra-taxon en la selección de caracteres para la clave. Si *VARYWT* tiene valor 0, se excluyen de la clave aquellos caracteres que tengan cualquier variabilidad intra-taxon, si por el contrario el valor de este parámetro es 1 no se penaliza este aspecto.
- **Control del efecto de la frecuencia de un ítem**. La frecuencia f de un ítem, está relacionada con la abundancia, *a*, (*abundance*) del mismo mediante la expresión $f = \text{Abase}^{a-5}$. Si *ABASE* =1, todos los ítem tendrán las mismas frecuencias y la abundancia no tendrá influencia en la información de la clave.
- **Caracteres de confirmación**. El programa es capaz de añadir a la clave caracteres de confirmación, esto es, caracteres cuyos valores tienen la misma distribución que el carácter principal seleccionado en ese punto de la clave. El carácter principal y los caracteres de confirmación son equivalentes, cualquiera de ellos podría haberse utilizado en la identificación.
- **Selección no automática de caracteres**. *Key* selecciona los caracteres de forma automática, no obstante es posible indicarle que seleccione determinados caracteres en determinados pasos de la clave haciendo uso de la directiva *PRESETS CHARACTERS*.

Con los aspectos arriba citados, y bajo la suposición de que el costo es aditivo (el coste de utilizar dos caracteres es la suma del coste de utilizar cada uno de dichos caracteres), el sistema pretende la construcción de un clave en la que el costo medio de la identificación se minimice.

El número mínimo de preguntas a responder para identificar un carácter bajo la suposición de que todos los caracteres tienen dos estados viene dado por la expresión de la Fórmula 1-6:

$$L_{\min} = Log_2 n$$

Fórmula 1-6. *Número mínimo de preguntas en una identificación.*

Supongamos que se utiliza un carácter de costo *c* para dividir un grupo de *n taxa* en *s* subgrupos. La clave para todo el grupo se obtiene mediante la

generación de una subclave para cada subgrupo. El costo medio, *C*, de una identificación viene dado por la Fórmula 1-7:

$$C = c + (\sum_{j=1}^{s} f_j c_j L_j) \Big/ (\sum_{j=1}^{s} f_j)$$

Fórmula 1-7. *Costo medio de una identificación.*

Donde:

- c_j: Es el costo medio de los caracteres utilizados en la subclave *j*-ésima y viene dado por la expresión $c = R\text{base}^{5-r}$.
- f_j: Es la frecuencia total de los ítem del grupo *j*-ésimo. La frecuencia f de un ítem y viene dada por la expresión $f = A base^{a-5}$.
- L_j: Es la longitud de la subclave *j*-ésima.

Si asumimos que las subclaves tienen una longitud próxima a la longitud media y que la mayoría de los caracteres tienen dos estados, entonces podemos aproximar la Fórmula 1-7 mediante la expresión de la Fórmula 1-8:

$$C = c + (\sum_{j=1}^{s} f_j c_j \log_2 n_j) \Big/ (\sum_{j=1}^{s} f_j)$$

Fórmula 1-8. *Primera aproximación al costo de una identificación.*

Donde n_j es el número de *taxa* en el subgrupo *j*-ésimo.

La estimación de c_j antes de la construcción de las subclaves es difícil. En el programa todas las c_j se equiparan a c_{min}, el menor costo para los caracteres bajo consideración, quedando la expresión de la Fórmula 1-9.

$$C = c + c_{\min} (\sum_{j=1}^{s} f_j \log_2 n_j) \Big/ (\sum_{j=1}^{s} f_j)$$

Fórmula 1-9. *Segunda aproximación al costo de una identificación.*

Esta expresión se completa con un término *V*, que permite al controlar la cantidad de variabilidad intra-taxon en la clave. La expresión final para el cálculo del costo de un carácter es la descrita en la Fórmula 1-10:

$$C = \left[c + c_{\min} \left(\sum_{j=1}^{s} f_j \log_2 n_j \right) \Big/ \left(\sum_{j=1}^{s} f_j \right) \right] + V$$

$$V = \left(\frac{1 - \text{Varywt}}{\text{Varywt}} \right) \left(\frac{n+8}{n \log_2 n} \right) \left(\sum_{j=1}^{s} n_j - n \right)$$

Fórmula 1-10. *Expresión final para el cálculo del costo de un carácter en el programa Key.*

KEY version 2.12 Windows

M.J. Dallwitz and T.A. Paine

CSIRO Division of Entomology, GPO Box 1700, Canberra, ACT 2601, Australia
Phone +61 2 6246 4075. Fax +61 2 6246 4000. Email delta@ento.csiro.au

Run at 10:09 on 10-JUN-03.

Characters - 5 in data, 3 included, 2 in key.
Items - 3 in data, 3 included, 3 in key.

RBASE = 1.40 ABASE = 2.00 REUSE = 1.01 VARYWT = .80
Number of confirmatory characters = 3

Average length of key = 1.0 Average cost of key = 1.0
Maximum length of key = 1 Maximum cost of key = 1.0

Preset characters (character,column:group) 2,1:1

Characters included 2-3 5

1(0). Color de las ramillas Castaño oscuro; Color del Sincarpo Rojizo-amarillento..... Ephedra nebrodensis
Color de las ramillas Verde-amarillentas; Color del Sincarpo Rojizo..... Ephedra distachya
Color de las ramillas Cenicientas; Color del Sincarpo Rojo..... Ephedra fragilis

Figura 1-25. *Ejemplo de clave dicotómica generada por Key***.**

El programa Dist.

Es un programa que genera matrices de distancia a partir de una versión modificada del coeficiente de similitud de Gower [*Gower, 1971*].

La directiva *MATCH OVERLAP* especifica que, para caracteres multiestado sin orden, la contribución de un carácter a la distancia entre dos ítem es 0 si dichos ítem tienen alguno de los valores del carácter en común. Para caracteres multiestado sin orden, la contribución D_{ikj} de un carácter k a la distancia entre el ítem i y j es:

- Si se utiliza la directiva *MATCH OVERLAP*. Si i y j tienen algún valor de estado en común, entonces D_{ijk} es 0 y 1 en otro caso.
- Si no se utiliza la directiva *MATCH OVERLAP*:

$$D_{ijk} = 0.5 * (| P_{ilk} \text{-^} P_{jlk} + || P_{isk} \text{-^} P_{jsk} | + || P_{ink} \text{-^} P_{jnk} |)$$

Fórmula 1-11. *Cálculo de la distancia para caracteres multiestado sin ordenar en el programa Dist.*

Donde:

- P_{isk} es la probabilidad de que el ítem i tenga el estado s para el carácter k.
- P_{jsk} es la probabilidad de que el ítem j tenga el estado s para el carácter k.
- N es el número de estados del carácter k.
- Si el carácter tiene más de un valor en un ítem, la probabilidad se divide entre los valores presentes.

Para caracteres multiestado con orden o valores numéricos la contribución de cada carácter a la distancia será:

$$D_{ijk} = | X_{ik} - X_{jk} | / R_k$$

Fórmula 1-12. *Cálculo de la distancia para caracteres numéricos y multiestado ordenados en el programa Dist.*

Donde R_k es el rango de X_{ik} para todos los ítem *i* incluidos y X_{ij} se calcula para cada ítem *i* incluido y para cada carácter *j* durante el paso de formato *Delta* a formato *Dist*.

La contribución a la distancia de cada carácter se multiplica por el peso del carácter y la suma de esas contribuciones ponderadas se divide por la suma de los pesos. Si un carácter no está codificado o es inaplicable para algún ítem, no contribuirá a la distancia total o a la suma de los pesos.

El programa Intimate.

Permite al autor de conjuntos de datos *IntKey* asociar imágenes con caracteres o *taxa* determinados y añadir anotaciones a dichas imágenes.

El programa Intkey.

Es un programa para la identificación interactiva de especies mediante la comparación de sus atributos con descripciones taxonómicas almacenadas en formato *Delta*. Sus principales características son las siguientes:

- Permite añadir y borrar estados de caracteres en cualquier momento y en cualquier orden durante la identificación.
- Calcula los caracteres más prometedores durante la identificación.
- Permite errores (bien en los datos, bien cometidos por el usuario) durante la identificación.
- Es capaz de expresar variabilidad o incertidumbre en los atributos.
- Muestra información adicional sobre los caracteres, *taxa*, etc.
- Es capaz de manejar valores numéricos, incluidos rangos de valores.
- Trata de forma diferente los valores desconocidos e inaplicables en función del tipo de aplicación.
- Permite el tratamiento de dependencia entre caracteres.

Intkey5, la versión para Windows 95/98/NT de *IntKey* puede acceder a datos e imágenes a través de Internet. Para ello, se auto instala como una *helper application*. En la página web habrá un enlace a un fichero *startup*, que indica a *Intkey* dónde está el conjunto de datos y las imágenes. Al activar el enlace, el navegador activa *IntKey* que descarga los datos de la red, extrae su contenido y comienza la identificación.

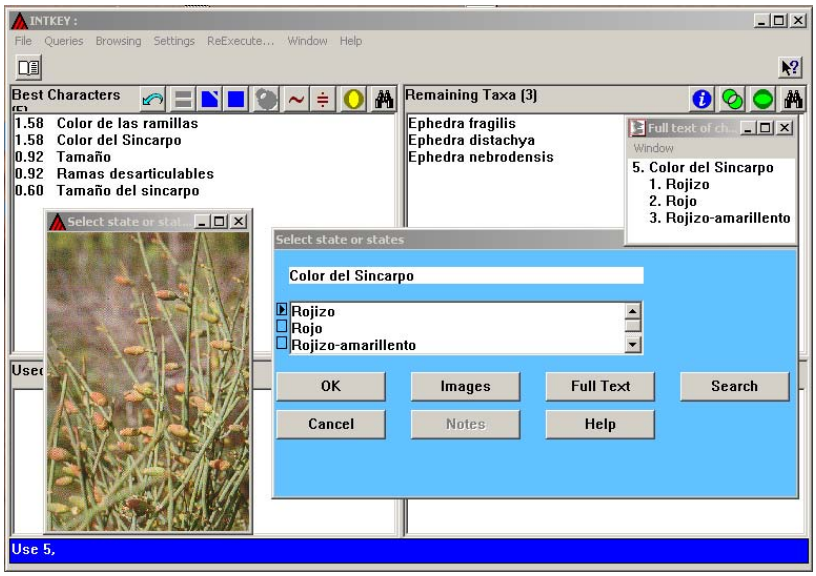


Figura 1-26. *Interfaz del sistema de identificación Intkey.*

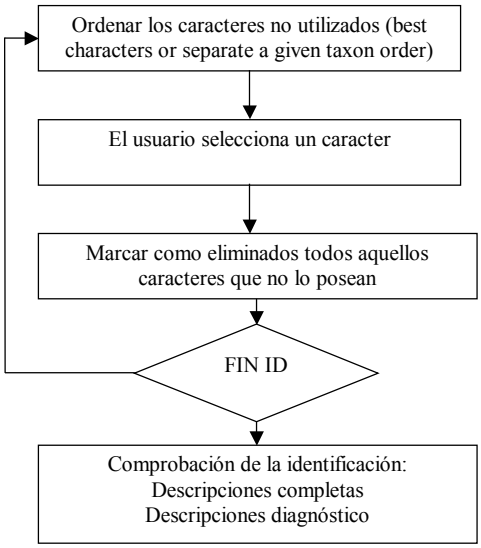


Figura 1-27. *Proceso de identificación de Intkey.*

4.9 *X:ID*.

X:ID [UBio, 2003] es un sistema para la generación de claves diagnóstico y la identificación taxonómica basado en XML. Permite a los usuarios crear sus propias claves a través de la web y ejecutarlas de forma local o remota. El formato XML combinado con XSLT permite a los desarrolladores personalizar el aspecto de las claves.

El programa consta de dos módulos: *X:ID Builder*, que crea y edita claves y *X:ID player* que utiliza dichas claves. Los principales módulos están escritos en PHP con la librería gráfica GD y el procesador XSLT Sablotron. *X:ID* almacena y lee ficheros ASCII que el *player* convierte al DTD de *X:ID*. También ejecuta claves *LucID* salvadas con el formato *LIF* (*LucID Interchange Format*, ver Figura 1-30).

X:ID BUILDER.

Permite al autor crear y editar claves diagnóstico. La clave consiste en un fichero, almacenado en formato *LIF*, con la posibilidad de incorporar recursos multimedia como imágenes o vídeos.

Este sistema no incluye ninguna funcionalidad para tratar atributos numéricos, que tienen que ser discretizados e introducidos como atributos simbólicos por el desarrollador (Figura 1-28). Utiliza la tecnología XML como herramienta para la visualización más que para facilitar la interoperatividad y el intercambio de información entre sistemas. Los valores posibles para un determinado atributo son fijos (Figura 1-29) y permite al usuario seleccionar entre cuatro modos de identificación diferente:

- Estricta.
- Permitiendo valores desconocidos.
- Permitiendo valores malinterpretados.
- Permitiendo tanto valores desconocidos como malinterpretados.

Figura 1-28. *Tratamiento de atributos numéricos con X:ID.*

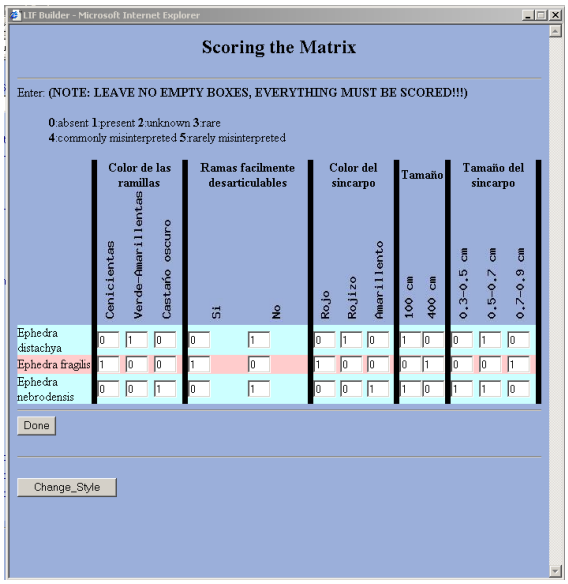


Figura 1-29. *Introducción de la matriz de datos en X:ID.*

```
#Lucid Interchange Format File v. 2.1

[..Character List..]
Tamaño
 metres
Color de las ramillas
 Cenicientas
 Verde-Amarillentas
 Castaño oscuro
Ramas facilmente desarticulables
 Si
 No
Tamaño del sincarpo (largo)
 centimetres
Color del sincarpo
 Rojo
 Rojizo
 Amarillento

[..Taxon List..]
Ephedra fragilis
Ephedra distachya
Ephedra nebrodensis

[..Main Data (txs)..]
6100106100
6010016011
6001016001

[..Metric Data..]
1       1       0.000   0.000   4.000   4.000
1       7       0.700   0.700   0.900   0.900
2       1       0.000   0.000   1.000   1.000
2       7       0.500   0.500   0.700   0.700
3       1       0.000   0.000   1.000   1.000
3       7       0.300   0.300   0.700   0.700

[..Dependency Data..]

[..Character Sets..]

[..Multimedia Directories..]
Sounds "C:\Archivos de
programa\Lucid\generoEphedra\Sounds\"
Images "C:\Archivos de
programa\Lucid\generoEphedra\Images\"
Macros "C:\Archivos de
programa\Lucid\generoEphedra\Macros\"
Text "C:\Archivos de
programa\Lucid\generoEphedra\HTML\"
Video "C:\Archivos de
programa\Lucid\generoEphedra\Video\"

[..State Text Data..]

[..State Media Data..]
2 "" "Ephedra_con_ramillas_cenicientas.bmp" ""
3 "" "ramillas_verdes.bmp" ""
5 "" "Tallos Articulados.bmp" ""
7 "" "Tamaño_Sincarpo.bmp" ""
9 "" "color_rojizo.bmp" ""
10 "" "color_rojizo_amarillento.bmp" ""

[..Taxon Media Data..]
1 "" "ephedra_fragilis.jpg" ""
1 "" "ephedra_fragilisDistri.gif" ""
2 "" "ephedra_distachya.jpg" ""
2 "" "ephedra_distachyaDistri.gif" ""
3 "" "ephedra_nebrodensis.jpg" ""
3 "" "ephedra_nebrodensisDistri.gif" ""

[..Taxon Text Data..]
```

Figura 1-30. *Ejemplo de descripción taxonómica en formato LIF.*

X:ID PLAYER.

Permite identificar un objeto de una lista de objetos descritos por un conjunto de criterios proporcionados por un autor. A medida que el usuario selecciona criterios, algunos *taxa* se irán descartando de las posibles soluciones, hasta que finalmente sólo quede un taxon. La toma de decisiones por parte del usuario puede estar apoyada por el uso de recursos multimedia como videos, imágenes, etc.

Además de la identificación, realiza otras funciones adicionales:

- Similitud. Permite seleccionar un taxon de forma que el resto de *taxa* se compararán con este y se devuelve una medida que representa el porcentaje de estados similares que un determinado taxon tiene con el taxon seleccionado.
- Comparación. Permite seleccionar varios *taxa* y devuelve los estados comunes y los no comunes.
- Poda de estados redundantes. Elimina aquellos estados que no afectan en la identificación de los *taxa* que quedan en un momento determinado.
- Búsqueda de estados determinantes. Analiza los estados que no se han utilizado todavía para determinar si la elección de alguno de ellos permite resolver el proceso de identificación.
- Búsqueda de estados prometedores. Devuelve para cada estado el porcentaje de *taxa* que se eliminarían en el proceso de decisión si se selecciona dicho estado.
- Claves enlazadas. Las claves se pueden enlazar de forma que, al terminar con una clave, pasemos a una nueva clave de identificación.
- Enlace a URLs. *X:ID* puede también enlazar con páginas web. De esta forma los desarrolladores pueden crear su propias páginas de *taxa* con objetos multimedia. Las características XML / XSLT del sistema permiten adaptar su aspecto con el de dicha página.

Capítulo 2. Modelos para la representación del conocimiento taxonómico.

En Taxonomía, las descripciones de *taxa* son uno de los principales medios para almacenar información tanto en bruto como altamente procesada. De hecho, uno de los requisitos del *International Codes of Botanical and Zoological Nomenclature* es la obligatoriedad de incluir una descripción diagnóstica en cualquier publicación válida de un nuevo taxon. Las descripciones taxonómicas forman además el núcleo de las monografías biológicas, floras y guías de campo.

Estas descripciones pueden tomar varias apariencias, La más conocida es la descripción en lenguaje natural, una descripción semiestructurada y semiformalizada de un organismo, o de un taxon. Las descripciones en lenguaje natural pueden ser simples, breves y escritas en un lenguaje claro (como las de las guías de campo) o extensas, extremadamente formales y de terminología especializada (este es el caso, por ejemplo, de las monografías taxonómicas).

Las descripciones basadas totalmente en información altamente estructurada son escasas. Este es el caso de documentos *LucID* y descripciones

Delta y *Nexus*. La inmensa mayoría de las descripciones taxonómicas están desprovistas de anotaciones (*data markup*) y son difíciles de tratar mediante motores analíticos y rutinas de *data-mining*. A continuación, describimos los modelos de representación de información taxonómica más utilizados.

1.1 El modelo *Delta*.

Cuando la forma de codificar descripciones taxonómicas digitales viene impuesta por los requerimientos de un programa, hay que reestructurar los datos cada vez que pretendemos realizar una operación diferente. Esto conlleva una disminución de la capacidad de automatización e interconexión con otros grupos de trabajo. El sistema *Delta*, *Description Language for TAxonomy*, [Dallwitz, 1974; Dallwitz *et al.,* 2000], propone un modelo de almacenamiento de descripciones para superar estos problemas. Fue adoptado como estándar del TDWG en 1991.

Los elementos esenciales del modelo *Delta* son:

- Listas de caracteres.
- Descripciones de *taxa*.
- Valores implícitos.
- Dependencias entre caracteres.

LISTAS DE CARACTERES.

Delta describe los *taxa* mediante listas de caracteres, cada uno de los cuales consiste en un nombre o característica (*feature*) y un conjunto de estados. El modelo reconoce cinco tipos de caracteres agrupados en tres categorías:

- Caracteres multiestado, categóricos o cualitativos. Aquellos con un número de estados fijo. Pueden ser de dos tipos:
  - Ordenados (*ordered multistate*, *OM*). Este tipo de caracteres representa una escala nominal formada por un conjunto de valores con los que no tiene sentido establecer una secuencia.

  - Sin orden (*unordered multistate*, *UM*). Este tipo de caracteres representa una escala ordinal formada por un conjunto de valores con los que tiene sentido establecer una secuencia.
- Caracteres numéricos o cuantitativos. Pueden ser enteros (*integer*, *IN*) o reales (*real*, *RN*).
- Caracteres textuales (*text*, *TE*). Almacenan cualquier tipo de información textual. Se utilizan, por ejemplo, para incluir referencias bibliográficas.

Los símbolos especiales "*V*", "*U*" y "–" representan los valores variable (*variable*), desconocido (*unknown*) y no aplicable (*not applicable*) respectivamente y son denominados seudo-valores (*pseudo-values*).

#1. striated area on maxillary palp <presence>/
    1. present/
    2. absent/
#2. pronotum <colour>/
    1. red/
    2. black/
    3. yellow/
#3. eyes <size>/
    1. of normal size <i.e. less than 0.5mm in diameter>/
    2. very large <i.e. more than 0.5mm in diameter>/
#4. frons <setae>/
    1. with setae on anterior middle and above eyes/
    2. with setae above eyes only/
    3. without setae/
#5. number of lamellae in antennal club/
#6. length/ mm/
#7. <comments>/

Ejemplo 2-1. *Ejemplo de lista de caracteres. Los caracteres 1, 2 y 3 son caracteres UM, el 4 es de tipo OM. Los caracteres 5 y 6 son de tipo IN y RN respectivamente. El carácter 7 es de tipo texto. Los comentarios están delimitados por ángulos "<>".*

La descripción de un carácter (ver Ejemplo 2-1) incluye el nombre (*feature*) y un número que lo identifica. Los números de caracteres son enteros consecutivos en orden ascendente comenzando por 1.

- En el caso de caracteres multiestado, el nombre del carácter está seguido por la descripción de los estados válidos para ese carácter. Cada estado tiene asociado un nombre y un número de estado. Los números de estado también son enteros consecutivos comenzando por 1 dentro de cada carácter.
- En el caso de caracteres numéricos, el nombre del carácter puede estar seguido de las unidades en que se mide dicho carácter.

DESCRIPCIONES DE *TAXA*.

La descripción de un taxon consiste en una o más descripciones de ítem (*item description*), cada una de las cuales define una forma o variedad (*variant*) del taxon. Por lo general, una descripción por taxon es suficiente, pero también es posible representar dos o más subespecies como ítem diferentes dentro de una especie, o bien representar un taxon variable (*variable*) mediante varias descripciones de ítem.

Una descripción consiste en un nombre de ítem seguido de un conjunto de atributos. La forma más sencilla de un atributo es un par *c,v* donde *c* es un número de carácter y *v* es un valor del carácter o un seudo-valor. La no aparición de un atributo equivale a incluir dicho atributo con el seudo-valor *U* (con la excepción de las variedades de un taxon multi-ítem, o del caso de dependencias entre caracteres o valores implícitos). Por ejemplo, el ítem:

Species A/ 1,1 3,2 5,2 6,9 4,1

es equivalente a:

Species A/ 1,1 2,U 3,2 4,1 5,2 6,9

Las descripciones de un taxon multi-ítem se agrupan e identifican como pertenecientes al mismo taxon añadiendo el símbolo “+”. La primera descripción se denomina ítem principal (*main item*) y el resto variedades (*variants*). Los atributos que no aparecen en el ítem principal se consideran desconocidos (o implícitos o dependientes), mientras que los atributos que no aparecen en las variedades, tienen el mismo valor en el ítem principal.

Por ejemplo, el taxon multi-ítem:

Species B (Australia)/ 1,1 2,1/2<rare> 3,1 5,3 6,5–6
#+ Species B (New Guinea)/ 3,2 5,U

es equivalente a:

Species B (Australia)/ 1,1 2,1/2<rare> 3,1 5,3 6,5–6
Species B (New Guinea)/ 1,1 2,1/2<rare> 3,2 5,U 6,5–6

OTROS ELEMENTOS DE *DELTA*.

Delta añade a esta estructura general de listas de caracteres y *taxa* la capacidad de representar:

- Caracteres implícitos. Algunos caracteres tienen un estado que se presenta en la mayoría de *taxa*. Es posible especificar que un valor se interpretará como implícito a menos que se indique lo contrario. El objetivo es mejorar las descripciones en lenguaje natural omitiendo los valores de los atributos que se dan generalmente.
- Dependencia de caracteres. Algunas veces determinados valores de un carácter, denominado carácter de control, implican que otros caracteres no pueden ser aplicados. Un ejemplo muy habitual es un carácter que especifica la presencia o ausencia de alguna estructura. Si la estructura está ausente, es evidente que todos los caracteres relacionados con ella son inaplicables.
- Utilidad de un carácter. *Delta* permite controlar el efecto de la utilidad (*reliability*) de los caracteres. Este parámetro se utiliza para seleccionar un atributo durante la generación de claves dicotómicas: los caracteres con mayores valores de utilidad se utilizarán en pasos más tempranos de una clave.
- Peso de un ítem. Es posible asignar a los ítem un peso cuya interpretación dependerá del programa que los esté utilizando. Los ítem con mayor peso se enfatizan de alguna manera, por ejemplo, aparecen antes en las claves.

- Caracteres obligatorios. *Delta* permite especificar que un carácter debe ser codificado de forma obligatoria para todo ítem. Si esto no sucede, el sistema dará un mensaje de aviso.
- Precisión de los límites de error. El modelo *Delta* permite especificar los límites de error que serán aplicados a caracteres de tipo real al convertir al modelo de generación de claves dicotómicas y de identificación. Así, un error absoluto *r* aplicado a un valor *v* produce un rango [*v-r, v+r*].

1.2 El modelo *DeltaAccess*.

La realización del trabajo relacionado con la Sistemática dentro de un entorno integrado en un sistema gestor de bases de datos es de suma utilidad. Permite combinar datos de observaciones y experimentos con información más elaborada y resumida, relacionada con la nomenclatura, referencias bibliográficas, etc. Por otro lado, la disponibilidad de reglas de integridad abre nuevos horizontes para el análisis de datos y la verificación de la integridad de los datos. El propósito fundamental de *DeltaAccess* [Hagedorn, 1995-2003] es acomodar información *Delta* en un entorno de base de datos relacional estándar.

DeltaAccess puede ser utilizado de tres formas diferentes:

1. Como módulo para importar información en formato *Delta* a una base de datos.
2. Como herramienta de análisis de datos (por ejemplo, análisis estadísticos de los caracteres).
3. Como almacén central de información en torno al que se organiza el trabajo de un determinado grupo.

CARACTERÍSTICAS GENERALES DEL MODELO *DELTAACCESS*.

DeltaAccess reconoce los mismos tipos de datos que *Delta*: categóricos o cualitativos (*UM* y *OM*), numéricos o cuantitativos (*IN* y *RN*) y textuales (*TE*).

Inspirado en *Delta*, incluye muchas de sus capacidades. Por ejemplo, permite especificar la obligatoriedad de la presencia de un carácter (*mandatory characters*) y marcar determinados estados de caracteres como implícitos. También define límites de error para caracteres numéricos y multiestado ordenados (*OM*).

Al igual que *Delta*, incorpora una medida de la utilidad de un carácter, pero en este caso se define en función de dos factores:

- La utilidad del carácter (*reliability*). Representa su variabilidad y lo bien que ha sido valorado o medido por el autor del conjunto de datos.
- La disponibilidad (*availability*). Mide la facilidad de valoración del carácter por alguien que utiliza el conjunto de datos para la identificación.

La abundancia (*abundance*) de un taxon determina la frecuencia con que será identificado. Este valor depende mucho de la localidad, por lo que la tendencia es dejar de utilizar este factor.

ELEMENTOS DEL MODELO *DELTAACCESS*.

DeltaAccess organiza el conocimiento taxonómico en un conjunto de tablas relacionadas (ver Figura 2-1) que describimos de forma breve a continuación:

Tabla de caracteres: X_CHAR.

Almacena la descripción de los caracteres válidos dentro de una descripción. Los atributos más destacados son:

- *CID (Character Definition Attribute)*. Es un código que forma la llave primaria de la tabla.
- *Character Name*. El Nombre del carácter, no puede ser nulo y ni estar repetido en la base de datos.
- *Type*. Tipo del atributo.

Tabla de estados: X_CS.

Describe el conjunto de estados válidos para cada carácter. Los estados especiales (*U*, *V*,-) se definen de forma implícita, sin necesidad de que aparezcan como estados en la lista de estados de un carácter. El estado especial *TE* permite introducir información textual en un estado de un carácter. Los elementos más importantes de esta tabla son:

- *CID*. Es la llave externa con la que cada estado se enlaza con un carácter de la tabla X_CHAR.
- *CS*. Código para el estado del carácter. Su valor más frecuente es un entero positivo, aunque también puede almacenar valores especiales *V*, *U* y estadísticos.
- *Character state name*. Nombre del estado del carácter.
- *Implicit*. Para establecer dicho estado por defecto.

Tabla de ítem*: X_ITEM.*

Contiene el nombre del taxon e información adicional, como referencias bibliográficas.

- *IID*. La clave primaria (IID) enlaza el ítem con su descripción X_DESCR.
- *ItemName*. Nombre del ítem, puede ser un taxon o cualquier otra frase que lo identifique. No tiene por qué ser único, pues un mismo ítem puede ser descrito varias veces (por ejemplo, para diferenciar las descripciones de fuentes bibliográficas diferentes).
- *LitRef*[6]. Enlace a un sistema de referencias literarias en formato legible para el usuario.

[6] La información recogida dentro de la base de datos puede venir, bien de información bibliográfica, bien de estudios de especies existentes en colecciones. Se hace necesario almacenar cuáles son las fuentes de información para poder comprobar en cualquier momento quién fue el autor, validar dicha información y determinar si siempre es cierta o, por ejemplo, depende de la localidad.

- *Abundance*. Representa la frecuencia con que se identificará el ítem. Se utiliza para asignarle caminos más cortos al generar claves dicotómicas y en la identificación interactiva.

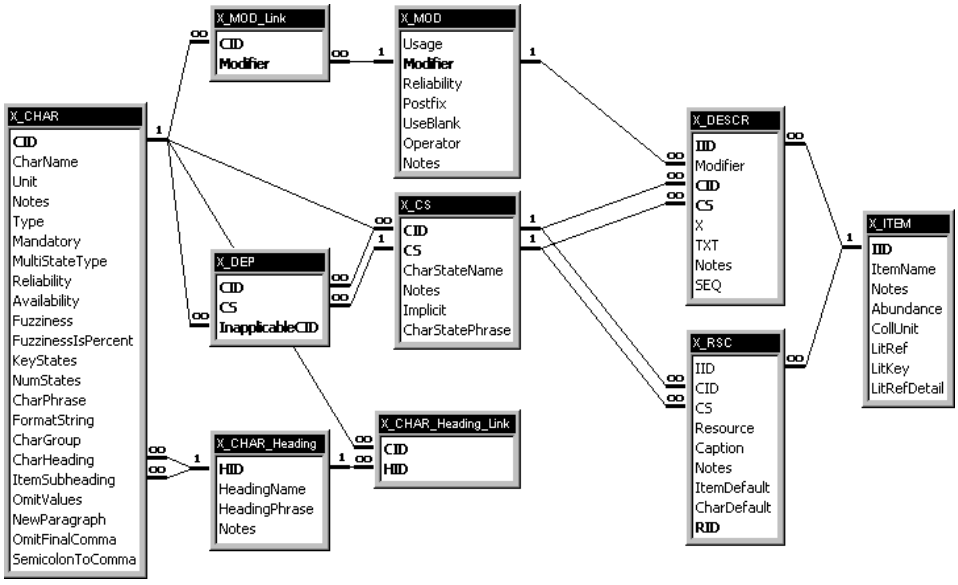


Figura 2-1. *Modelo de Base de datos de DeltaAccess completo.*

Tabla de dependencias: X_DEP.

Modela la dependencia que se produce cuando un carácter es inaplicable ante la presencia de un carácter de control.

- *CID.* Atributo de control en la relación de dependencia.
- *CS.* Estado del atributo de control en la relación de dependencia.
- *InaplicableCID*. Llave externa al carácter dependiente.

Tabla de descripciones: X_DESCR.

Almacena las descripciones de los ítem almacenados en la base de datos.

- *IID*. Llave externa a la tabla X_ITEM que identifica al ítem.
- *CID*. Llave externa a un carácter de la tabla X_CHAR.
- *CS*. Llave externa a un estado de la tabla X_CS, que junto con el atributo CID forma una referencia a la tabla X_CS.

- *X*. Valor real que representa el valor numérico de los atributos numéricos y estadísticos. Se utiliza en atributos de tipo real y entero.
- *TXT*. Sólo se utiliza este atributo en el caso de estados especiales *TE*. El uso de TXT y X es exclusivo.

Otras tablas del modelo.

El modelo se completa con otras tablas como:

- X_CHAR_HEADING. Facilita la descripción de grupos de caracteres.
- X_PROPERTY. Contiene propiedades de cada proyecto.
- X_RSC. Almacena enlaces a recursos externos como imágenes, ilustraciones, hipervínculos, etc.

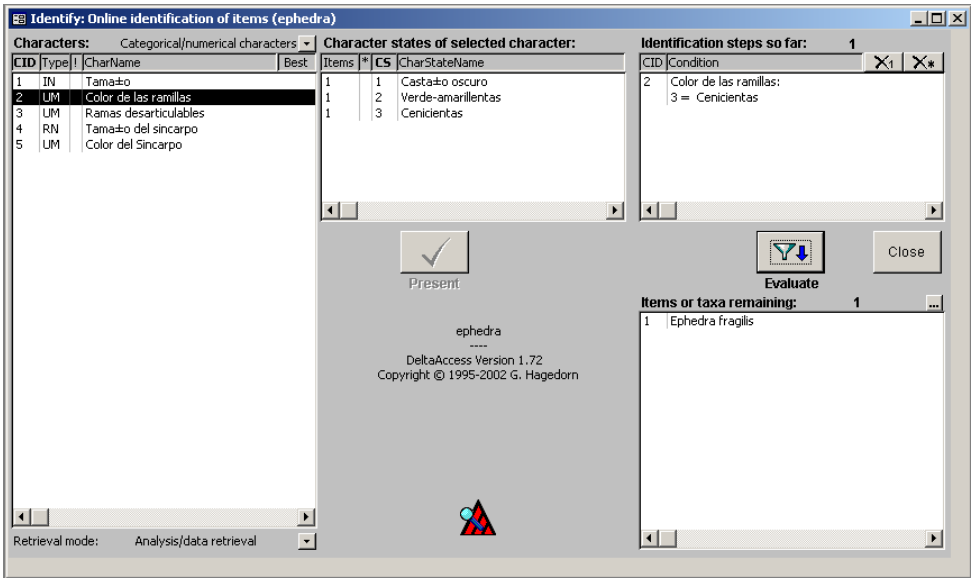


Figura 2-2. *Modulo Identify de DeltaAcces.*

EL MÓDULO DE IDENTIFICACIÓN TAXONÓMICA *IDENTIFY.*

DeltaAccess, incluye la utilidad *Identify*, un módulo para identificación taxonómica en fase de desarrollo, por lo que no tiene totalmente implementadas características como la ordenación de caracteres y la utilización de imágenes. La Figura 2-2 ilustra la apariencia de dicho módulo. De izquierda a derecha el

usuario selecciona uno o varios caracteres y uno o varios estados de forma que sus elecciones se visualizan en el panel superior derecho. Tras activar el botón de evaluación (*Evaluate*), se presenta, ordenado por nombre, el conjunto de *taxa* que cumplen con las condiciones especificadas. Durante la identificación se pueden reconsiderar las condiciones impuestas en cualquier momento.

1.3 El modelo *Nexus*.

Es un modelo que nace en 1987 con la pretensión de unificar los formatos utilizados por el programa *MacClade3*[7] [Maddison & Maddison, 1992] y el programa *Paup3*[8] [Swofford, 1991]. Posteriormente Roderic Page adoptó *Nexus* en el programa *Component2*[9] [Page, 1993].

Además de los proyectos enumerados, existen otros proyectos, dedicados en su mayoría a la realización de estudios filogenéticos, que han adoptado *Nexus* como base de trabajo. Este es el caso de *SplitTrees* [Huson & Wetzel, 1994] y *TreeBASE* [Donoghue *et al.,* 1996].

Como comprobamos, *Nexus* se ha centrado en el software para estudios filogenéticos. Una de las últimas novedades es el sistema *Mesquite* [Maddison & Maddison, 2003], un conjunto de pequeños programas dedicados a análisis específicos para la comparación de información sobre organismos, que hace hincapié en los estudios filogenéticos, aunque también es capaz de realizar estudios para biología evolutiva.

El modelo *Nexus* pretende evitar el problema de transformar los ficheros de datos cada vez que el usuario utiliza un programa diferente y ser independiente

[7] Dedicado a la edición de información sistemática y el análisis filogenético de la evolución de caracteres.

[8] Diseñando para realizar inferencia filogenética.

[9] Dedicado a comparar (*comparting)* árboles y conciliar filogenias asociadas, por ejemplo, árboles de parásitos y huéspedes.

del sistema operativo. Su diseño procura facilitar la capacidad de añadir nuevos elementos y que los programas no preparados para la lectura de estos nuevos elementos puedan seguir funcionando con normalidad.

EL MODELO DE DATOS DE *NEXUS*.

El modelo *Nexus* es modular, consiste en diversos bloques bien diferenciados, cada uno de los cuales contiene un tipo especial de información. Los bloques son series de directivas (*commands*) que comienzan con la palabra BEGIN y terminan con la palabra END (ver Figura 2-3).

```
BEGIN block-name;
   command-name token
   command-name token
END;
```

Figura 2-3. *Estructura general de un bloque Nexus.*

Existen dos tipos básicos de bloques:

- Bloques privados, que contienen información relevante para un programa específico.
- Bloques públicos, que contienen información de carácter general utilizada por varios programas. Se distinguen nueve tipos básicos de bloques públicos:

1. TAXA. Define a los *taxa* y les da nombre (ver Figura 2-4).
2. CHARACTERS. Contiene información sobre datos discretos y continuos, e incluye caracteres para la definición de estructuras morfológicas y secuencias moleculares. El modelo permite además representar polimorfismo e información sobre frecuencia.
3. UNALIGNED. Es similar al bloque CHARACTERS pero contiene información sobre secuencias moleculares sin alinear.
4. DISTANCES. Este bloque contiene matrices de distancia entre *taxa*.
5. SETS. Contiene descripciones sobre colecciones de objetos. Estos objetos pueden ser caracteres, *taxa* ,árboles y estados.

6. ASSUMPTIONS. Contiene suposiciones sobre los datos, por ejemplo, una asignación de pesos a determinados caracteres.
7. CODONS. Designan los lugares en la secuencia de nucleótidos que codifican proteínas y la posición dentro de un codón de cada lugar y asigna un código genético a la secuencia molecular.
8. TREES. Codifican relaciones entre *taxa*. Si los *taxa* son especies, los árboles representan filogenia. Por otro lado, si los *taxa* son secuencias de ADN los árboles representan una genealogía genética.
9. NOTES. Recoge información adicional como texto e imágenes

```
BEGIN TAXA;
   DIMENSIONS NTAX=number-of-taxa;
   TAXLABELS taxon-name [taxon-name...];
END;
```

Figura 2-4. *Estructura del bloque taxa. DIMENSIONS indica el número de taxa que se describe y TAXLABELS especifica su nombre.*

Debido a su importancia y a su estrecha relación con la identificación taxonómica y la generación de claves de identificación, describimos brevemente el bloque CHARACTERS (ver Figura 2-5).

EL BLOQUE *CHARACTERS.*

Describe los caracteres válidos dentro del documento *Nexus* y la matriz de datos describe, para cada taxon, el valor que toman los atributos. A grandes rasgos incluye:

- Información de carácter general. Por ejemplo, DIMENSIONS especifica con NCHAR el número de caracteres en la matriz de datos.
- Información sobre el tipo de los datos de la matriz de datos. En función del valor de la directiva DATATYPE, el contenido de la matriz de datos podrá ser: STANDARD (caracteres discretos utilizados generalmente para información morfológica), CONTINUOUS, DNA, RNA, NUCLEOTIDE y PROTEIN.

```
BEGIN CHARACTERS;
  DIMENSIONS [NEWTAXA NTAX=number-of-taxa] NCHAR=number-of-characters;
  [FORMAT
    [DATATYPE={ STANDARD | DNA | RNA | NUCLEOTIDE | PROTEIN | CONTINUOUS } ]
    [RESPECT CASE]
    [MISSING=symbol]
    [GAP=symbol]
    [SYMBOLS="symbol [symbol...]"]
    [EQUATE="symbol=entry [symbol=entry...]"]
    [MATCHAR=symbol]
    [[NO] LABELS]
    [TRANSPOSE]
    [INTERLEAVE]
    [ITEMS=[MIN] [MAX] [MEDIAN] [AVERAGE] [VARIANCE] [STDERROR] [SAMPLESIZE]
    [STATES] )]
    [STATESFORMAT={STATESPRESENT|INDIVIDUALS|COUNT|FREQUENCY}]
    [[NO]TOKENS]
  ;]
  [ELIMINATE character-set;]
  [TAXLABELS taxon-name [taxon-name...];]
  [CHARSTATELABELS
    character-number [character-name][/state-name [state-name...]]
    [,character-number[character-name][/state-name[state-name...]]
    ...]
  ;]
  [CHARLABELS character-name] [character-name...];]
  [STATELABELS
    character-number [state-name [state-name...]]
    [, character-number [state-name [state-name...]]]
  ;]
  MATRIX data-matrix;
END;
```

Figura 2-5. *Estructura general del bloque CHARACTERS.*

- Información sobre el formato de los datos. Por ejemplo, si se hará distinción entre mayúsculas o minúsculas (RESPECTCASE), los símbolos para designar información que debería aparecer, pero que por algún motivo no está disponible (MISSING), los símbolos que pueden ser utilizados en la matriz de datos (SYMBOL), etc.
- La matriz de datos. MATRIX contiene una secuencia de nombres de *taxa* y el valor de los estados para cada taxon.

```
BEGIN DATA
   DIMENSION NCHAR=7;
   FORMAT DATATYPE=DNA MATCHAR=.;
   MATRIX
      taxon_1  GACCTTA
      taxon_2  ...T..C
      taxon_3  ..T.C..;
END;
```

Figura 2-6. *Matriz de datos con información genética.*

- Información sobre el formato de la matriz de datos. Por ejemplo, TRANSPOSE indica que la matriz de datos se presenta en formato transpuesto (ver Figura 2-7) e INTERLEAVE que la matriz está dividida en secciones.

```
MATRIX
   character_1   1 0 1 1 0 1
   character_2   0 1 1 1 0 0
   character_3   0 1 1 1 1 0;
```

Figura 2-7. *Ejemplo de matriz traspuesta.*

1.4 El modelo *SDD*.

En septiembre de 1998, el IUBIS-TDWG fundó el subgrupo *SDD*. Este subgrupo trabaja en el desarrollo de un estándar basado en XML para representar y manejar información descriptiva de organismos. La necesidad del mismo surge tras comprobar que su predecesor, *Delta*, no era capaz de adaptarse a los nuevos retos demandados por la comunidad científica. Por otro lado, la ausencia de un estándar internacional e independiente para descripciones taxonómicas impide una mayor utilización de información descriptiva digitalizada, y acarrea una ineficiencia sustancial a la Taxonomía como un todo.

SDD pretende proporcionar una plataforma flexible e independiente para representar y almacenar descripciones taxonómicas, que facilite el intercambio sin pérdida de información de conjuntos de datos entre aplicaciones y la utilización de una misma descripción para diferentes propósitos. Su diseño se inspira en los modelos de representación más populares y aspira a constituirse como un superconjunto de requerimientos de información para los programas ya existentes. La finalidad es desarrollar un modelo de representación del conocimiento suficientemente estandarizado y ampliamente aceptado por toda la comunidad de biólogos, sea cual sea su disciplina de origen.

Entre sus características debe estar la facilidad de ampliación para reflejar futuros requerimientos de datos, ser legible para un humano y facilitar la anotación (*markup*) progresiva de las descripciones existentes, particularmente de

las descripciones en lenguaje natural. Al estar basado en XML proporcionará un esquema para la validación de documentos.

VERSIONES DE *SDD*.

Tras la reunión anual del TDWG en Brasil en marzo de 2002, se presentó la primera versión de *SDD*, *SDD strawman* 0.5. Esta versión fue sometida a análisis en la reunión que el TDWG celebró en Lisboa en octubre de 2003 evolucionando hasta su versión actual, la versión *SDD* 0.9[10] [TDWG: SDD, 2003], que este año ha sido nuevamente modificada en la reunión anual que TDWG ha celebrado en octubre en Nueva Zelanda. Las nuevas versiones pretenden solucionar problemas de diseño de versiones anteriores y ampliar los conceptos que se pueden representar.

En la actualidad, los esfuerzos se centran en llegar a un consenso sobre la estructura más adecuada para la representación de información descriptiva, por lo que todavía no se ha desarrollado ninguna herramienta completamente funcional basada en el mismo. Más bien, se han implementado traductores de otros formatos como *LIF* [UBio, 2003] o *EFG* [Morris & Stevenson, 2003] a *SDD*. El objetivo de estos programas es conservar la compatibilidad de los sistemas existentes y detectar fallos en el diseño del estándar.

Para el desarrollo de esta tesis, hemos seguido la versión 0.5 del estándar, debido a los motivos que se exponen a continuación.

- *SDD* evolucionará nuevamente en el 2004. Cualquier herramienta desarrollada sobre la base de la versión 0.9 debe evolucionar hacia una versión que, aunque todavía no está definida, se sabe que aparecerá a principios de 2005.
- De esta forma, y puesto que el trabajo se inició con la versión 0.5, no parece conveniente una revisión de versión hasta que el estándar se encuentre mejor cerrado.

[10] Las versiones comprendidas entre la 0.5 y la 0.9 son versiones intermedias que no se han publicado.

- Puesto que estos conceptos que pretendemos manejar (taxon, carácter, etc.) son muy estables, la adecuación a la versión que aparecerá en los próximos meses consiste básicamente en adaptar las primitivas de lectura / escritura de documentos XML.

ESTRUCTURA BÁSICA DE UN DOCUMENTO *SDD*.

Los documentos *SDD* se estructuran utilizando seis niveles de elementos XML. Cuatro de estos son obligatorios y el resto son opcionales (Figura 2-8):

- Elementos obligatorios: <*Document*>, <*Generator*>, <*ProjectDefinition*>, <*Terminology*>.
- Elementos opcionales: <*Descriptions*>, <*ResourceDefinitions*>, <*ApplicationData*>.

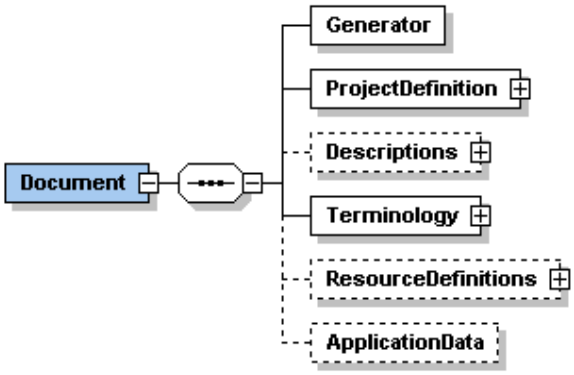


Figura 2-8. *Estructura general de un documento SDD.*

Elemento <Document>.

Es la raíz de un documento *SDD*, y contiene al resto de los elementos.

Elemento <Generator>.

En la mayoría de los casos un documento es generado por un programa informático. El elemento <*Generator*> especifica información sobre dicho

programa. Notar que este elemento sólo proporciona información sobre el generador más reciente del documento y no sobre otros generadores que pudieran haber sido utilizados con anterioridad. La Tabla 2-1 describe los atributos más importantes de este elemento.

ATRIBUTO	DEFINICIÓN
Application	Hace referencia al nombre del programa que creó el documento *SDD*.
Version	Es la versión del programa que creó el documento *SDD*.
Routine	En aquellos casos en que el programa dispone de varias rutinas de exportación, especifica el nombre de la rutina de exportación utilizada para crear el documento.
Authors	Autor o autores del programa generador del documento.
Institution	Institución que desarrolló el programa generador del documento.
Copyright	Copyright del programa.
LastUpdateDate	Fecha de la última actualización del programa generador.

Tabla 2-1. *Atributos del elemento <Generator>.*

Elemento <Project Definition>.

Describe información sobre el proyecto de investigación que desarrolla el documento (autores, colaboradores, fechas de publicación y revisión, etc.). La Figura 2-9 muestra la estructura de este elemento.

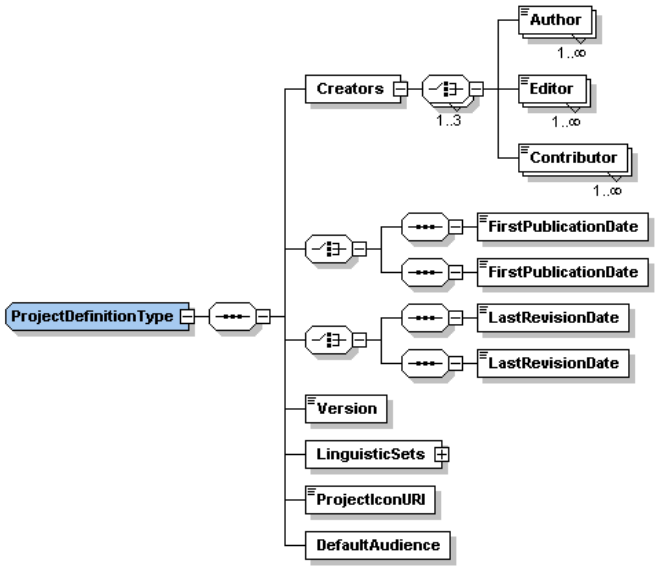


Figura 2-9. *Esquema del elemento <ProjectDefinition>.*

La Tabla 2-1 describe de forma breve los subelementos incluidos dentro de este.

ELEMENTO	DEFINICIÓN
<Creators>	Información referente a los autores, editores y colaboradores del proyecto.
<FirstPublicationDate>	Fecha de publicación del documento.
<LastRevisionDate>	Fecha de la última revisión del documento.
<Version>	Versión del proyecto.
<LinguisticSet>	Información textual sobre el proyecto que puede aparecer en uno o varios idiomas.
<DefaultAudience>	Tipo de público al que por defecto está destinado el documento. El tipo de público se caracteriza por dos aspectos, el idioma y nivel de conocimientos.

Tabla 2-2. *Elementos de <ProjectDefinition>.*

Elemento <Descriptions>.

Define una lista de entidades (por ejemplo, *taxa* o especimenes) y sus descripciones codificadas, o en lenguaje natural. Para permitir documentos que solo contengan terminología, las descripciones de *taxa* son opcionales.

El elemento *<Descriptions>* actúa como contenedor de elementos. En realidad, el elemento central de una descripción es el *<Item>*. Un *<Item>* tiene la estructura de la Figura 2-10. A continuación describimos de forma breve la semántica de sus elementos.

- *<ItemDefinition>*. Almacena el nombre y el rango taxonómico.
- *<NaturalLanguageDescription>*. Contiene una descripción en lenguaje natural. Esta descripción es opcional y puede presentarse en varios idiomas.
- *<Resources>*. Almacena referencias a recursos como imágenes, etc. relacionados con el *<Item>* que se describe.
- *<CodedDescription>*. Recoge la descripción del taxon en función de los caracteres y atributos que lo caracterizan.

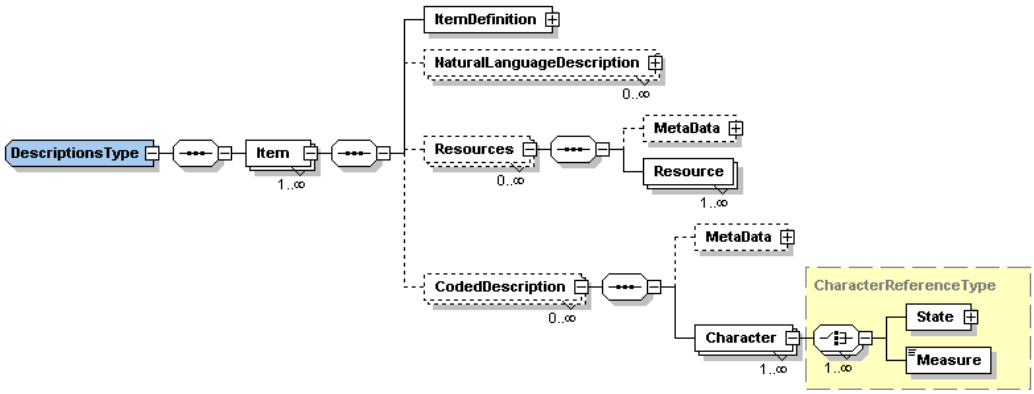


Figura 2-10. *Estructura del elemento <Descriptions>.*

Elemento <Terminology>.

Es el elemento más complejo de un documento *SDD* y define los caracteres y atributos que se consideran válidos dentro del documento. Además de caracteres y atributos, la terminología define otros elementos como tipos de público, estados de carácter global, etc. Su estructura se presenta en la Figura 2-11.

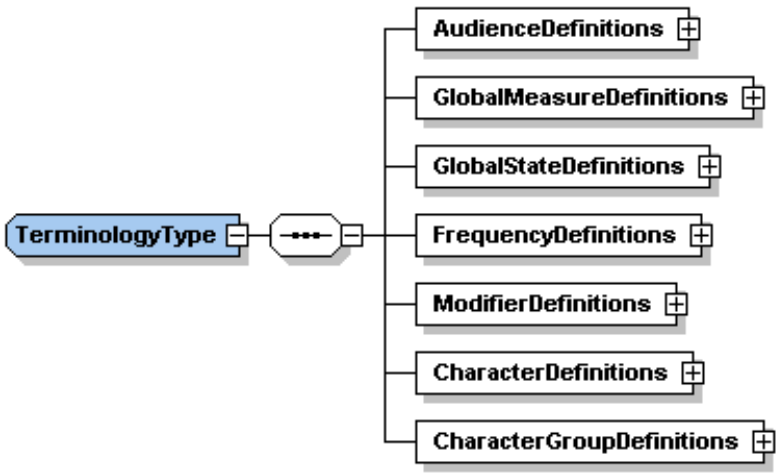


Figura 2-11. *Elemento <Terminology>.*

A continuación detallamos el contenido de los elementos que contiene.

- *<AudienceDefinitions>*. Define los tipos de público para el documento.

- <*GlobalMeasureDefinitions*>. Define la semántica de medidas numéricas como el mínimo, el máximo y la media. Estos elementos los fija el estándar *SDD* y no el investigador que diseña la terminología del documento.
- <*GlobalStateDefinitions*>. Define estados de carácter global que pueden ser utilizados por diferentes caracteres.
- <*FrequencyDefinitions*>. Define modificadores sobre la frecuencia de aparición de un determinado estado en un carácter (por ejemplo, raramente, frecuentemente, etc.). Los modificadores de frecuencia son globales y pueden ser utilizados por cualquier estado.
- <*ModifierDefinitions*>. Definen modificadores sobre la expresión de un estado no relacionados con la frecuencia de aparición de los mismos (fuertemente, etc.). Se definen de forma global y pueden ser utilizados por cualquier estado.
- <*CharacterDefinitions*>. Este elemento recoge las definiciones locales de los caracteres que serán utilizados en las descripciones. Un carácter queda definido mediante la especificación de su nombre y los estados que puede presentar.
- <*CharacterGroupDefinitions*>. Define grupos planos (por ejemplo, grupo caracteres para generar claves) y estructuras jerárquicas de caracteres (por ejemplo, el grupo caracteres de la planta contiene los subgrupos hojas, aspecto general y fruto, que a su vez pueden subdividirse).

Elemento <ResourceDefinitions>.

Define recursos relacionados con los elementos descritos en el documento. Estos recursos pueden ser referencias a la URI donde se localiza o el propio recurso embebido dentro del documento (por ejemplo, una foto en modelo *gift* codificada).

Elemento <Application Data>.

Este elemento almacena información específica para algún programa. Cada programa deberá leer sólo aquella información referida a si mismo. El

objetivo de este elemento es facilitar la traducción entre el modelo *SDD* y los modelos utilizados por otros programas.

1.5 Sinopsis.

Hemos visto como hasta la actualidad, cada grupo de investigación ha desarrollado su propio estándar para la representación de información así como las aplicaciones que lo utilizan.

La tendencia actual es la de favorecer la interoperatividad entre las aplicaciones y hacer que la información sea accesible desde cualquier lugar del mundo a través de Internet. Por esto cobra vital importancia el diseño y utilización de un estándar XML consensuado por la comunidad científica. Concretamente, el TDWG trabaja en la actualidad en *SDD*, un estándar para descripción de taxonomía en formato XML. Este estándar, además de almacenar información sobre descripciones taxonómicas, es capaz de almacenar dicha información en varios idiomas y para diferentes grupos de usuarios en función de su nivel de conocimientos. Al estar basado en XML, ofrece además otras posibilidades muy interesantes como la personalización de los interfaces por parte de los desarrolladores haciendo uso de XSLT (*eXtensible Style Sheet Transformation*).

SDD va a ser el sucesor de *Delta* como modelo de representación de información taxonómica. Dado lo reciente del estándar, que todavía se encuentra en fase de desarrollo y pruebas, todavía no se han desarrollado herramientas basadas en el mismo. A continuación presentamos ***DAtoSDD*** (*DeltaAccess to SDD*), una herramienta para trasladar conjuntos de datos desde *Delta* y *DeltaAccess* a *SDD*.

1.6 La herramienta *DAtoSDD*.

DAToSDD exporta conjuntos de datos de *DeltaAccess* a *SDD*. (ver Figura 2-12). Es una herramienta de sencilla utilización que hemos desarrollado durante una estancia en el *Federal Biological Research Centre for Agriculture and Forestry (BBA)* de Berlín. A continuación describimos cómo hemos realizado la correspondencia entre estos dos modelos de representación de conocimiento taxonómico.

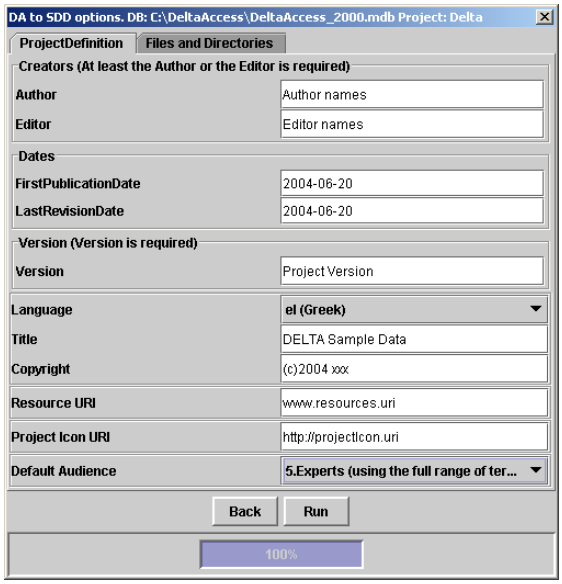


Figura 2-12. *DAtoSDD: Interfaz de petición de valores al usuario.*

ELEMENTO <GENERATOR>.

Este elemento almacena la información relacionada con el programa que genera el documento XML. En este caso es el programa *DAtoSDD*. Hemos incorporado en la traducción todos los atributos obligatorios de *SDD* y los opcionales que están presentes en *DeltaAccess* (Tabla 2-3).

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
Application	Sí	Valor por defecto: "DAtoSDD"
Version	Sí	Valor por defecto: "0.5"
LastUpdateDate	Sí	Fecha de la última actualización del programa *DAtoSDD*.
Authors	No	Valor por defecto: "Eva Lucrecia Gibaja Galindo"
Institution	No	Valor por defecto: "University of Granada"

Tabla 2-3. *DAtoSDD: Traducción del elemento <Generator>.*

ELEMENTO <PROYECTDEFINITION>.

El elemento *<ProyectDefinition>* almacena información de carácter general sobre el conjunto de datos que se describe en el documento. En *DeltaAccess*, la mayor parte de esta información se obtiene de la tabla *_PROPERTY*. *<ProjectDefinintion>* contiene siete elementos, todos ellos de aparición obligada: *<Creators>*, *<FirstPublicationDate>*, *<LastRevisionDate>*, *<Version>*, *<LinguisticSets>*, *<ProjectIconUri>* y *<DefaultAudience>*. A continuación se hace una descripción detallada de su correspondencia con *DeltaAccess*.

Elemento <Creators>.

El estándar *SDD* establece que al menos debe aparecer un *<Autor>* o un *<Editor>* en la descripción del proyecto. En *DeltaAccess* esta información no es de carácter obligatorio. Por este motivo se presentan dos casos:

1. Si la información no está presente en *DeltaAccess*, el usuario debe suministrarla a través de la interfaz.
2. Si la información está presente en el modelo *DeltaAccess*, se obtiene de la siguiente forma:

- **<***Author***>.** Se considera *<Autor>* del proyecto el contenido del campo "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" tiene el valor "*ProjectAuthors*".
- **<***Editor***>**. Se considera *<Editor>* del proyecto el contenido del campo *"_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" tiene el valor "*ProjectEditors*".

HIJOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
<*Author*>	Sí	_PROPERTY.TextValue donde _PROPERTY.PropertyName = "ProjectAuthors"
<*Editor*>	Sí	_PROPERTY.TextValue donde _PROPERTY.PropertyName = "ProjectEditors"

Tabla 2-4. *DAtoSDD: Traducción del elemento <Creator>.*

Elementos <FirstPublicationDate>, <LastRevisionDate>.

DeltaAccess no incluye esta información. Dada su obligatoriedad, el usuario debe introducirla a través de la interfaz. Por defecto, hemos considerado la fecha del sistema. El formato de esta fecha puede ser:

- Un año representado por un entero en el rango [1970-2100].
- Una fecha en la forma AAAA-mm-dd.

Elementos <Version>, <Title>, <Rights>, <ProjectIconURI>.

Estos elementos obligatorios en *SDD* son opcionales en *DeltaAccess*. En caso de no ser suministrados por el usuario o no estar presentes en *DeltaAccess*, se toman por defecto los valores indicados en la Tabla 2-5.

HIJOS	TRADUCCIÓN	VALOR POR DEFECTO
<*Version*>	Se obtiene de la tabla "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName"* tiene el valor *"ProjectVersion*"	"*Project Version*".
<*Title*>	Se obtiene del campo "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" tiene el valor "*ProjectTitle*"	"*ProjectTitle*"
<*Rights*>	Se obtiene del campo "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" toma el valor "*ProjectCopyRight*"	*"© año -xxx*"
<*ProjectIconUri*>	Se obtiene del campo "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" tiene el valor "*ProjectIconURL"*	*"http://projectIcon.uri*"

Tabla 2-5. *DAtoSDD: Traducción y valores por defecto de los elementos <Version>, <Title>, <Rights> y <ProjectIconUri>.*

Elemento <Description>.

El modelo *SDD* admite, de forma opcional, una breve descripción del proyecto. Esta información también se almacena, de forma opcional, en *DeltaAccess*. En el caso de estar presente, el valor de este elemento se obtiene del campo "*_PROPERTY.TextValue*" donde "*_PROPERTY.PropertyName*" tiene el valor "*ProjectDescription*".

Elemento <DefaultAudience>.

DeltaAccess no representa esta información mientras que *SDD* obliga a definir un tipo de público por defecto. Si el usuario no introduce un valor mediante la interfaz de la aplicación, *DAtoSDD* asigna por defecto el valor "*en5*" (idioma inglés y nivel de conocimientos 5).

ELEMENTO <DESCRIPTIONS>.

El elemento <*Descriptions*> contiene al menos un elemento **<***ItemDefinition***>.** La correspondencia entre este elemento y *DeltaAccess* se detalla en la Tabla 2-6 y la Tabla 2-7.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
Identification	Sí	_ITEM.ItemName. En caso de ser nulo se toma _ITEM.IID.
Abundance	No	_ITEM.Abundance
InternalNotes	No	_ITEM.Notes
CollectionSpecimenID	No	_ITEM.CollUnit

Tabla 2-6. *DAtoSDD: Correspondencia entre los atributos del elemento <ItemDefinition> y DeltaAccess.*

Elemento <Resources>.

Dentro de la definición de un <*Item*>, el elemento <*Resources*> es opcional. En caso de aparecer, consta al menos de un elemento <*Resource*> que

tiene un único atributo obligatorio *keyref*. El valor de este atributo se obtiene del valor del campo "*_RSC.RID*".

HIJOS	OBLIGATORIO	TRADUCCIÓN
<*Identification*>	Sí	_ITEM.ItemName. En caso de ser nulo se toma _ITEM.IID.
<*Abundance*>	No	_ITEM.Abundance
<*InternalNotes*>	No	_ITEM.Notes
<*CollectionSpecimenID*>	No	_ITEM.CollUnit
<*ReferenceID*>	No	_ITEM.LitKey
<*ReferenceFreeDescription*>	No	_ITEM.LitRef
<*Wording*>	No	_ITEM.ItemWording

Tabla 2-7. *DAtoSDD: Correspondencia entre los subelementos del elemento <ItemDefinition> y DeltaAccess.*

Elemento <CodedDescription>.

Un elemento <*ItemDefinition*> puede contener de forma opcional una descripción formulada sobre la base de los caracteres (y sus correspondientes estados) definidos en el elemento <*Terminology*>. En caso de aparecer, <*CodedDescription*> contiene al menos un elemento <*Character*>.

<*Character*> tiene un atributo obligatorio, *keyref*, que es una referencia a un carácter descrito en la sección <*Terminology*>. El valor del atributo *keyref* se obtiene del valor del campo "*_DESCR.CID*".

- En el caso de caracteres de tipo numérico, el conjunto de valores que puede tomar el carácter se define utilizando el elemento <*Measure*> cuya correspondencia con *DeltaAccess* se detalla en la Tabla 2-8.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
keyref	Sí	_DESCR.CS
value	Sí	_DESCR.X

Tabla 2-8. *DAtoSDD: Correspondencia entre los atributos de <Measure> y DeltaAccess.*

- En el caso de caracteres de tipo categórico, el conjunto de valores que puede tomar el carácter se define utilizando el elemento <*State*> cuya traducción se detalla en la Tabla 2-9 y Tabla 2-10.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
keyref	Sí	_DESCR.CS _DESCR.TE (en el caso de caracteres de tipo texto)

Tabla 2-9. *DAtoSDD: Correspondencia entre los atributos de <State> y DeltaAccess.*

HIJOS	OBLIGATORIO	TRADUCCIÓN
<*Frequency*>	No	_DESCR.Modifier donde _MOD.Modifier = _DESCR.Modifier y _MOD.usage=”Frequency”
<*Modifier*>	No	_DESCR.Modifier donde _MOD.Modifier = _DESCR.Modifier y _MOD.usage!=”Frequency
<*ReportedNotes*>	No	_DESCR.Notes
<*InternalNotes*>	No	_DESCR.TXT

Tabla 2-10. *DAtoSDD: Correspondencia entre los subelementos de <State> y DeltaAccess.*

ELEMENTO <TERMINOLOGY>.

Este elemento es el más complejo definido por el estándar. Contiene, de forma obligatoria, siete hijos: <*AudienceDefinitions*>, <*GlobalMeasureDefinitions*>, <*GlobalStateDefinitions*>, <*FrequencyDefinitions*>, <*ModifierDefinitions*>, <*CharacterDefinitions*> y <*CharacterGroupDefinitions*> cuya correspondencia don *DeltaAccess* se detalla a continuación.

Elemento <AudienceDefinitions>.

El elemento <*AudienceDefinitions*> no tiene atributos y consta, al menos, de un elemento <*AudienceDefinition*> que describe el tipo de público por defecto. *DeltaAccess* no representa este tipo de información. Por ese motivo hemos incluido de forma automática el conjunto de <*AudienceDefinition*> más habitual, donde también se encuentra la seleccionada como <*DefaultAudience*>.

Elemento <GlobalMeasureDefinitions>.

Este elemento representa medidas de carácter global. Son predefinidas por el estándar y se incluyen de forma automática en la traducción. Al traducir de *DeltaAccess* a *SDD*, se establece la correspondencia de la Tabla 2-11.

SDD	DELTAACCESS	SDD	DELTAACCESS
Mean	Mean	MeanMinus1DS	-SD
UndefinedLowerRangeLimit	-Low	ConfInterval025	-CI95
UndefinedUpperRangeLimit	+High	ConfInterval975	+CI95
Min	Min	ConfInterval05	-CI90
Max	Max	ConfInterval95	+CI90
Median	Median	Percentile95	+Q90
Mode	Mode	Percentile05	-Q90
StdDeviation	SD	Percentile90	+Q80
StdErrorMean	SE	Percentile10	-Q80
Value	Val	Percentile75	+Q50
SampleSize	N	Percentile25	-Q50
MeanPlus1SD	+SD		

Tabla 2-11. *DAtoSDD: Correspondencia entre las medidas globales de DeltaAccess y SDD.*

Elemento <GlobalStateDefinitions>.

El elemento <*GlobalStateDefinitions*> define estados globales. *SDD* distingue tres tipos de estados globales:

1. Estados globales especiales (*SpecialStates*). Indican por qué el valor de un carácter es desconocido. Son establecidos por el estándar, por lo que se incluyen automáticamente en la traducción. Al hacer la traducción de *DeltaAccess* a *SDD* se establece la siguiente correspondencia:
   - Los valores "*null*" en *DeltaAccess* tomarán valor "*Empty*" en *SDD*.
   - Los valores "*U*" en *DeltaAccess* tomarán valor "*Unknown*" en *SDD*.
   - Los valores "–" en *DeltaAccess* tomarán valor "*NotApplicable*" en *SDD*.
2. Estados globales de cálculo (*ComputedSpecialStates*). Indican por qué no ha podido ser calculado un determinado valor. Son predefinidos por el estándar y también se incluyen de forma automática en la traducción.

3. Estados globales definidos por el usuario. *DeltaAccess* considera locales todos los estados definidos por el usuario. No obstante, *SDD* obliga a definir un conjunto de estados globales del usuario. Se trata de un problema en el diseño del estándar detectado durante la realización de este proyecto. Se ha incluido un conjunto definido por defecto.

Elemento<FrecuencyDefinitions>.

SDD distingue dos tipos de modificadores, los modificadores de frecuencia y los que no lo son. *DeltaAccess* permite definir diferentes grupos de modificadores de frecuencia. Cada uno de estos grupos estará representado en *SDD* por un elemento <*FrequencyDefinitionSet*>.

Nombre del modificador	Valores
Frequency 1	abundant
	almost always
	almost never
	always
	at least one
	commonly
	frequently
	never
	very rarely

Tabla 2-12. *DAtoSDD: Ejemplo de modificador de frecuencia.*

<*FrequencyDefinitionSet*> tiene un atributo obligatorio, *name*, cuyo valor se corresponde con el valor del campo "*_MOD.Usage*" si dicho valor contiene la palabra "*Frequency*". La etiqueta <*Label*> de este conjunto de modificadores de frecuencia también será el valor "*_MOD.Usage*".

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
name	Sí	_MOD.Usage ⇔contiene la palabra "Frequency"
SUBELEMENTOS	**OBLIGATORIO EN *SDD***	**TRADUCCIÓN**
<*Label*>	Sí	_MOD.Usage ⇔contiene la palabra "Frequency"

Tabla 2-13. *DAtoSDD: Correspondencia entre los atributos y subelementos de <FrequencyDefinitionSet> y DeltaAccess.*

El elemento <*FrequencyDefinitionSet*> esta formado, al menos, por una definición de modificadores de frecuencias, <*FrequencyDefinition*>. Cada elemento <*FrequencyDefinition*> consta de un conjunto de atributos cuya correspondencia con *DeltaAccess* se detalla en la Tabla 2-14.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
key	Sí	_MOD.Modifier
PropagateOnCollation	Sí	Valor por defecto: false
LowerLimit	Sí	_MOD.LowerFreq
UpperLimit	Sí	_MOD.UpperFreq

Tabla 2-14. *DAtoSDD: Correspondencia entre los atributos y subelementos de <FrequencyDefinition> y DeltaAccess.*

Elemento <ModifierDefinitions> .

<*ModifierDefinitios*> define modificadores que no son de frecuencia. Puede haber distintos tipos de modificadores, cada uno de ellos definido por un elemento <*ModifierDefinitionSet*>.

<*ModifierDefinitionSet*> tiene un atributo obligatorio, *name*, cuyo valor se toma del campo "*_MOD.Usage*", si y solo si este valor no contiene la palabra "*Frequency*". La etiqueta (<*Label*>) del elemento también tiene este valor.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
name	Sí	_MOD.Usage <=> no contiene la palabra Frequency
SUBELEMENTOS	**OBLIGATORIO EN *SDD***	**TRADUCCIÓN**
<*Label*>	Sí	_MOD.Usage <=> no contiene la palabra Frequency

Tabla 2-15. *DAtoSDD: Correspondencia entre los atributos y subelementos de <ModifierDefinitionSet> y DeltaAccess.*

El elemento <*ModifierDefinitionSet*> tiene, al menos, un hijo <*ModifierDefinition*>. que consta de un atributo obligatorio, *key*, cuyo valor se corresponde con el valor del campo "*_MOD.Modifier*".

NOMBRE DEL MODIFICADOR	VALORES	NOMBRE DEL MODIFICADOR	VALORES
Time	briefly	Morphology	asymmetrical
	earlier		attenuately
	early		densely
	late		irregularly
	soon		rounded
	when old		

Tabla 2-16. *DAtoSDD: Dos ejemplos de modificadores.*

Elemento <CharacterDefinitions>.

Contiene la definición de todos los caracteres locales que pueden aparecer en la descripción de un taxon o individuo. Consta de uno o más elementos <*CharacterDefinition*>. Cada uno de estos elementos tiene un atributo obligatorio, *key*, cuyo valor se corresponde con el valor del campo "*_CHAR.CID*". Tiene además los elementos hijo citados en la Tabla 2-17:

SUBELEMENTOS	OBLIGATORIO EN *SDD*
<*LinguisticSets*>	Sí
<*DescriptorDefinitions*>	Sí
<*FrequencyModifierSelections*>	No
<*ModifierSelections*>	No
<*Resources*>	No

Tabla 2-17. *DAtoSDD: Subelementos de <CharacterDefinition>.*

- **<***LinguisticSets***>**. Tiene un atributo obligatorio, *keyref*, que es una referencia al valor seleccionado como <*DefaultAudience*>. La Tabla 2-18 resume la correspondencia entre sus subelementos y *SDD*.

SUBELEMENTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
<*Label*>	Sí	_CHAR.charname
<*Wording*>	No	_CHAR.charWording

Tabla 2-18. *DAtoSDD: Subelementos de <LinguisticSets>.*

- **<***DescriptorDefinitions***>** Especifica los estados permitidos para un carácter. Un estado válido puede ser, una referencia a un estado global,

una referencia a un estado local o bien una referencia una medida. (ver Tabla 2-19).

SUBELEMENTOS	OBLIGATORIO EN *SDD*
<*GlobalStateSetReferernce*>	No
<*LocalStateDefinitions*>	No
<*MeasureDefinitions*>	No

Tabla 2-19. *DAtoSDD: Subelementos de <DescriptorDefinitions>.*

- **<***GlobalStateSetReference***>.** Referencia a los estados especiales. En caso de aparecer tiene un atributo obligatorio, *keyref*, que se corresponde con el valor "*SpecialStates*". Se incluye un elemento de este tipo cuando el campo "*_CS.CS*" tiene valores "*U*", "-". Para cada uno de estos valores se añade un <*DescriptorSelection*> que tomará los valores "*Unknown*" o "*NotApplicable*" respectivamente.

- **<***LocalStateDefinitions***>**. Define estados locales para el carácter. Cada uno de estos estados será representado por uno o varios elementos <*StateDefinition*>. <*StateDefinition*> tiene un atributo, *key*, obligatorio cuyo valor es el del campo "*_CS.CS*". Puede contener los subelementos descritos en la Tabla 2-20.

SUBELEMENTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
<*Label*>	Sí	_CS.CharStateName
<*Wording*>	No	_CS.StateWording
<*InternalNotes*>	No	_CS.Notes
<*Resources*>	No	_RSC.RID" donde "_RSC.CS" coincide con la clave "_CS.CS

Tabla 2-20. *DAtoSDD: Subelementos de <LocalStateDefinitions>.*

- **<***MeasureDefinitions***>** Define, en el caso de atributos continuos, el rango de valores posibles. Consta de un elemento **<***MeasureDefinition***>.**
  - **<***MeasureUnit***>.** Opcional y con un atributo:

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
key	Sí	Concatenación entre el atributo *key* del carácter y la referencia a la <*GlolbalMeasureDefinition*> correspondiente.
keyref	Sí	Referencia a la <*GlobalMeasureDefinition*> correspondiente

Tabla 2-21. *DAtoSDD: Subelementos de <MeasureDefinitions>.*

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
Postfix	No	_CHAR. UnitIsPrefix

Tabla 2-22. *DATOSDD. Traducción de los atributos del elemento <MeasureUnit>.*

Elemento <CharacterGroupDefinitions>.

Este elemento define grupos de caracteres. Cada grupo se define utilizando un elemento <*CharacterGroupDefinition*>. Este elemento tiene dos atributos obligatorios cuya correspondencia con *DeltaAccess* se resume en la Tabla 2-23. La traducción de sus subelementos se detalla en la Tabla 2-24.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
key	Sí	"IdentificationGroup" en el caso de grupos de caracteres para la identificación interactiva. "NaturalLanguageGroup" en el caso de grupos de caracteres destinados al lenguaje natural. "CharDefReportGroup" en otros casos.
type	Sí	Por defecto "UserDefinedHierarchy"

Tabla 2-23. *DAtoSDD: Atributos de <CharacterGroupDefinition>.*

SUBELEMENTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
<*Label*>	Sí	"Identification Group" en el caso de grupos de caracteres para la identificación interactiva. "Natural Language Group" en el caso de grupos de caracteres destinados al lenguaje natural. "Character Definition Report Group" en otros casos.
<*Purposes*>	No	"InteractiveIdentification" en el caso de grupos de caracteres para la identificación interactiva. "NaturalLanguageReporting" en el caso de grupos de caracteres destinados al lenguaje natural. "DefaultGeneral" en otros casos

Tabla 2-24. *DAtoSDD: Subelementos de <CharacterGroupDefinition>.*

Un grupo de caracteres está formado por uno o varios <*CharacterGroupItem*> que representa cada uno de los elementos del grupo. Este elemento tiene un atributo, *key*, obligatorio que toma el valor "*_CHAR.CID*" concatenado con "*_CHAR_Heading.HID*"donde "*_CHAR.CID*" se corresponde con "*_CHAR_Heading_Link.CID*". Puede contener:

- <*InternalNotes*> (opcional). *_CHAR_Heading.Notes.*
- <*Character*>. En este caso, el atributo obligatorio *keyref* hace referencia a un carácter definido en la sección <*CharacterDefinitions*>.
- <*CharacterGroupIte*m>. Es decir, el contenido de un elemento de un grupo de caracteres puede ser a la vez, otro grupo de caracteres.

DeltaAccess distingue, tres categorías de grupos de caracteres en función de su utilidad:

- Grupos de caracteres para la identificación taxonómica.
- Grupos de caracteres para realizar descripciones en lenguaje natural.
- Grupos de caracteres para generar informes.

Al realizar la traducción hemos incluido cada una de estas categorías como un grupo que contendrá a su vez diferentes subgrupos de caracteres. Por ejemplo, dentro del grupo de identificación interactiva puede haber subgrupos de caracteres relacionados con las hojas, el tronco, etc. (ver Figura 2-13).

ELEMENTO <RESOURCEDEFINITIONS>.

El elemento <*ResourceDefinitions*> describe recursos como imágenes, descripciones, etc. que pueden estar ligados a los caracteres e ítem descritos en el documento. Un recurso concreto queda definido por un elemento <*ResourceDefinition*>. Dicho elemento tiene un atributo, *type*, obligatorio que puede tomar un valor dentro de un conjunto predeterminado por el estándar. Su valor por defecto será "*Semantics*". El atributo *key* también es obligatorio y su valor será "*_RSC.RID*". El atributo *uri* es opcional y permite ligar el recurso a su localización en una red. Por defecto se toma el valor "*projectDefinition.resourceDefaultURL*" concatenado con "*_RSC.Resource*".

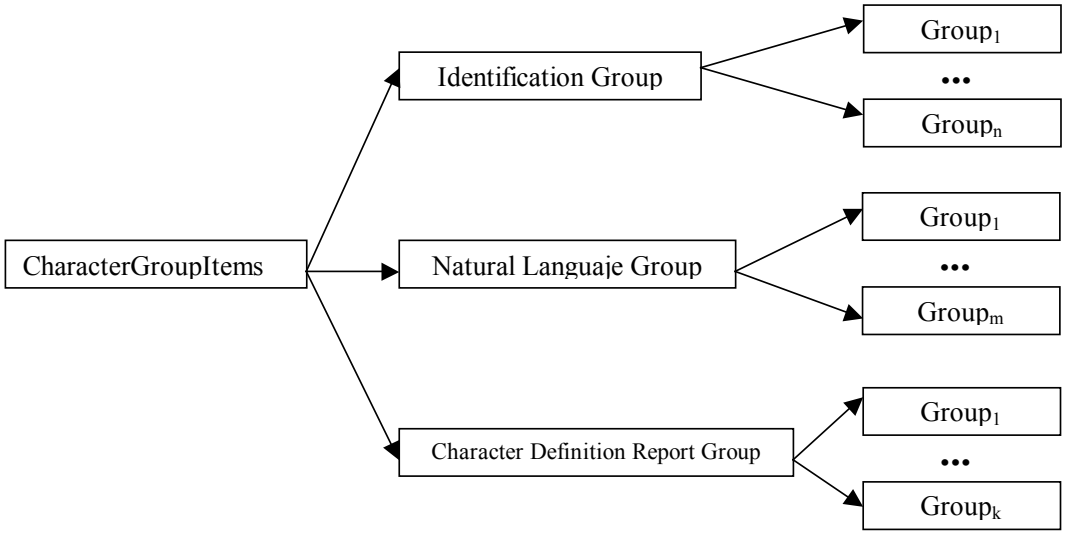


Figura 2-13. *DAtoSDD: Grupos de caracteres.*

Capítulo 3. Una aportación a la identificación taxonómica automatizada: el sistema *GREEN*.

El estudio sobre el estado de la utilización de las nuevas tecnologías en el área de la Biodiversidad y la Taxonomía pone de manifiesto la escasez de trabajo multidisciplinar entre expertos en informática y expertos en ciencias biológicas. Esta dicotomía conlleva que:

- La mayor parte de las publicaciones se realizan dentro del ámbito de la Biología y se centran, habitualmente, en obtener funcionalidad y resultados finales más que una descripción detallada sobre el diseño y técnicas aplicadas.
- Es habitual la falta de formación informática dentro del área de la Biología. El resultado son diseños poco adecuados y herramientas de escasa utilización que se quedan obsoletas.
- En el ámbito de las Ciencias de la Computación no abunda este tipo de publicaciones. Las herramientas desarrolladas en esta área no se adaptan adecuadamente a las necesidades de la comunidad de biólogos debido, fundamentalmente, a que solo actúan como proveedores de conjuntos de

ejemplos, y de forma poco frecuente se pretende implantar en los sistemas sus métodos de trabajo.

- Los trabajos realizados hasta la actualidad han sido desarrollados por grupos de investigación aislados, con el objeto de solucionar problemas muy concretos y sin atender a la existencia de otros sistemas y estándares para la representación de información. Esta falta de estandarización conduce a la imposibilidad de que varias aplicaciones compartan información.

Tras detectar la comunidad científica estos problemas, la tendencia está cambiando. Así, los últimos trabajos realizados evidencian que se ha captado la importancia del desarrollo de un trabajo global y reutilizable. A la cabeza de esta corriente encontramos a TDWG y GBIF que centran sus investigaciones en la estandarización de las descripciones taxonómicas y el tratamiento de las colecciones biológicas.

A continuación presentamos el sistema *GREEN*, un sistema experto en la identificación de especímenes biológicos que hemos desarrollado para abordar, desde un equipo multidisciplinar, el reto de conciliar los puntos de vista de los expertos botánicos e informáticos.

1. El sistema *GREEN*.

El sistema *GREEN* (*Gymnosperms Remote Expert Executed Over Networks*) es el fruto de un proyecto desarrollado por miembros del grupo de investigación ARAI (*Aproximate Reasoning and Artificial Intelligence*) y del Herbario de la Universidad de Granada, que estaba interesado en el desarrollo de un sistema para la identificación de especies vegetales. El grupo elegido para el desarrollo de nuestra investigación fueron las Gimnospermas presentes en la Península Ibérica, por ser un grupo bien representado tanto en número de especies como por formar masas forestales integrantes de nuestro paisaje.

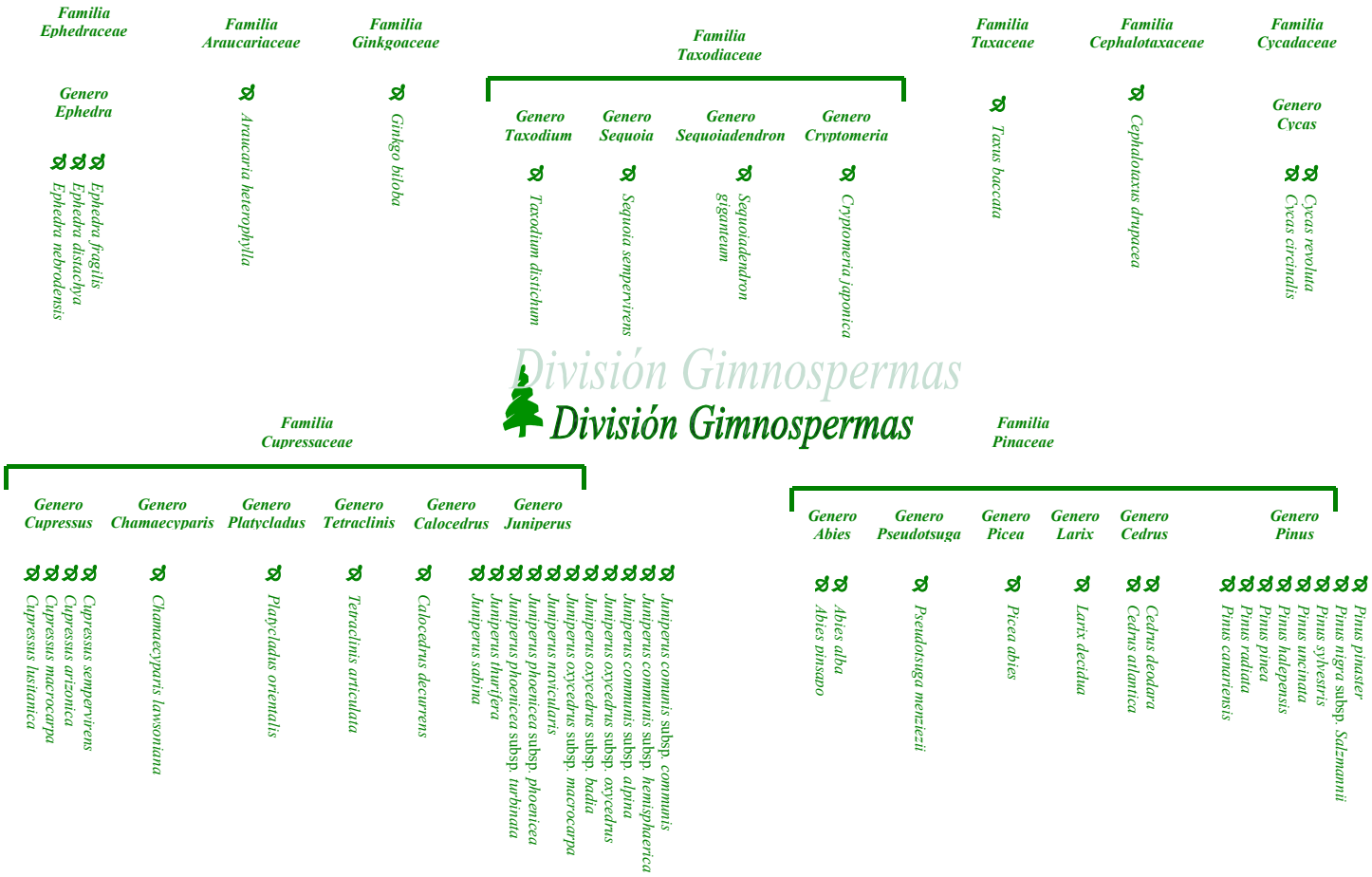


Figura 3-1. *Conocimiento representado en el sistema GREEN.*

A pesar de ello, hemos de señalar que se han tenido en cuenta no sólo los taxones silvestres o naturalizados; también se han incluido aquellos árboles utilizados con frecuencia para usos forestales, reforestaciones y jardinería (ver Figura 3-1). *GREEN* es un sistema de carácter docente y divulgativo, dirigido a alumnos que comienzan sus estudios de Botánica y Biología y al público general, en estado operativo y accesible en Internet.

1.1 Estructura General del sistema.

Al estudiar los métodos de trabajo de los botánicos se pone de manifiesto que, para identificar especies vegetales, utilizan claves dicotómicas que podemos equiparar con reglas IF-THEN donde cada regla conduce a otra regla o a una especie vegetal. Así, cuando un botánico desea identificar un especimen concreto, se distinguen:

- Una fuente de conocimiento constituida por la información disponible sobre el grupo taxonómico. Esta información se presenta habitualmente en forma de claves dicotómicas.
- Un proceso de utilización de dicho conocimiento para resolver el problema concreto.

Esta descripción es afín a la arquitectura básica de un sistema basado en el conocimiento, y más concretamente a la de un sistema experto. Además de esta separación entre el conocimiento y la forma de utilizarlo, observamos otras características, típicas de los dominios de aplicación en que es factible la implantación de sistemas expertos:

- Utilización de conocimiento muy específico sobre un dominio.
- Naturaleza más heurística que algorítmica del conocimiento utilizado.
- El problema no se puede resolver con métodos de computación tradicionales y sí con técnicas de razonamiento simbólico.
- Se pretenden resultados similares a los de un experto.
- El sistema ha de tratar con conocimiento impreciso.

Siguiendo esta estructura, *GREEN* está constituido por una base de conocimiento que almacena el conocimiento sobre el dominio del problema y un motor de inferencia que extrae conocimiento a partir de la base de conocimiento y de la información suministrada por el usuario durante la sesión de consulta.

El sistema codifica el conocimiento mediante reglas, que proporcionan una estructuración del mismo comprensible por el usuario y análoga a las claves dicotómicas utilizadas por los expertos botánicos.

Además de la base de conocimiento y del motor de inferencia, y reflejando la estructura ideal de un sistema basado en el conocimiento, *GREEN* consta de:

- Un módulo de tratamiento de la incertidumbre. Se distinguen dos fuentes de incertidumbre: la que es inherente a la naturaleza y la debida a la subjetividad de las observaciones humanas.
- Un módulo justificador que explica los resultados alcanzados por el sistema en un lenguaje próximo al lenguaje natural.

Hemos añadido también dos módulos de apoyo al usuario.

- Un catálogo multimedia para la consulta de especies conocidas.
- Un glosario de términos científicos para acercar la terminología a usuarios no expertos en Botánica.

Vista la estructura del sistema desde la perspectiva de su diseño modular y de las Ciencias de la Computación, continuamos con algunas observaciones desde el punto de vista operativo y funcional. Debido al auge de las telecomunicaciones el sistema se consulta en la web, lo que conlleva la ampliación del espectro de receptores y una difusión mayor y más rápida del conocimiento. El sistema se consulta a través de una conexión con un servidor que atiende las peticiones de los usuarios (o clientes) y envía los resultados a través de Internet. En la Figura 3-2 podemos ver un esquema de la arquitectura del sistema *GREEN*.

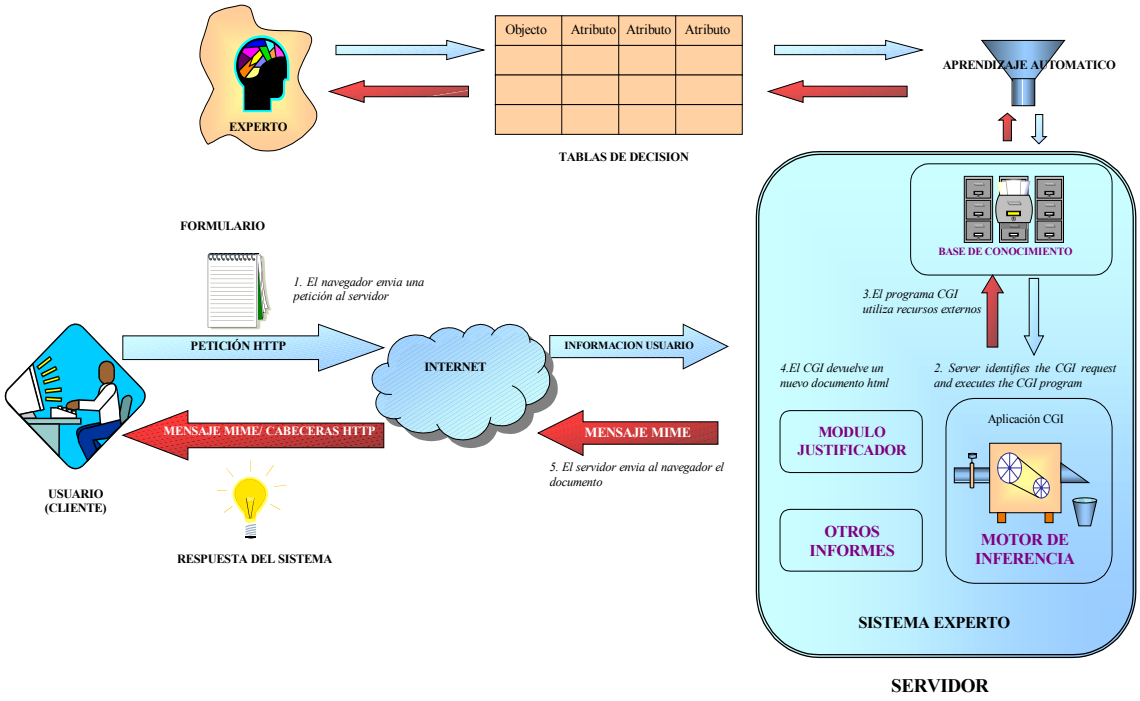


Figura 3-2. *Arquitectura de GREEN. El sistema está compuesto por dos módulos: el generador de reglas y el motor de inferencia.*

Descrita la estructura general del sistema, detallamos el proceso de diseño e implementación del mismo, indicando las técnicas de Inteligencia Artificial que hemos utilizado en cada fase.

1.2 Adquisición y elicitación del conocimiento.

Como hemos comentado, hemos seleccionado a las Gimnospermas ibéricas como grupo de trabajo. Este grupo es suficientemente grande y está muy estudiado, lo que lo hace adecuado para hacer una primera aproximación al problema y un análisis significativo de los resultados obtenidos.

La información disponible sobre el dominio de aplicación es dispersa, incompleta y poco estructurada. Esto hace necesario un proceso de adquisición y elicitación del conocimiento que permita trasladarla de forma adecuada a un

sistema informático. El proceso se inició a partir de varias claves dicotómicas cuya información recabamos y resumimos para elaborar un listado de los caracteres diagnóstico (descriptores o atributos) en tres niveles: familia, género y especie. Esta organización jerárquica del conocimiento ofrece la ventaja de respuestas a varios niveles: para alcanzar un objetivo en los niveles más altos y generales de la jerarquía (familia y género), se necesita (por lo general) la información que el usuario observa con mayor facilidad. Así, aún disponiendo de poca información es posible concluir un objetivo. A medida que conocemos más información la respuesta se refina hasta llegar al nivel más bajo y específico de la jerarquía (especie).

Con estos caracteres elaboramos un conjunto de tablas de decisión [Durkin, 1994] o ficheros de inducción [Gonzalez & Dankel, 1993], en los que cada fila recoge el valor de los caracteres descriptores de un taxon determinado. Toda esta información fue confrontada y refinada a partir de la consulta a los expertos y pliegos del Herbario. La Tabla 3-1 muestra un fragmento de una de las tablas de decisión obtenidas.

	DISPOSICIÓN DE LAS ARCÉSTIDAS	COLOR DE LA ARCÉSTIDA	ARCÉSTIDA PRUINOSA	TAMAÑO DE LA ARCÉSTIDA	Nº DE SEMILLAS DE LA ARCÉSTIDA
Juniperus communis subsp. *communis*	Axilares	Negro-azulado	Si	Entre 0.6 y 1 cm	3
Juniperus communis subsp. *hemisphaerica*	Axilares	Negro-azulado	Si	Entre 0.6 y 1 cm	3
Juniperus communis subsp. *alpina*	Axilares	Negro-azulado	Si	Entre 0.6 y 1 cm	3
Juniperus oxycedrus subsp. *oxycedrus*	Axilares	Castaño	No	Entre 0.6 y 1 cm	1-3
Juniperus oxycedrus subsp. *badia*	Axilares	Castaño	No	Más de 1 cm	1-3
(...)	(...)	(...)	(...)	(...)	(...)

Tabla 3-1. *Fragmento de una tabla de decisión.*

El proceso de clasificación en Biología involucra una gran cantidad de caracteres, en el caso particular de Gimnospermas ibéricas se describieron un total de 88. Para facilitar el acceso a los mismos, los dividimos en siete grupos: "*aspecto general del taxón*", "*características de la hoja*", "*características de las*

ramas", "*características de las ramillas*", "*características de la fructificación (piña y arcéstida)*", "*características de las semillas*" y "*ecología del taxón*".

1.3 Generación automática de la base de conocimiento.

La Botánica utiliza claves de identificación elaboradas manualmente por un experto en la materia. La Inteligencia Artificial describe técnicas que permiten obtenerlas de forma automática. En particular, la inducción de reglas identifica la información importante en grandes conjuntos de datos y la resume en forma de reglas que forman la base de conocimiento. Además, reduce el tiempo necesario para adquirir el conocimiento y es adecuada en casos como este en que no hay una base de casos disponible.

LA HERRAMIENTA *KEYMANAGER*.

La herramienta *KeyManager* procesa las tablas de decisión obtenidas durante la adquisición y elicitación del conocimiento con el algoritmo ID3 para obtener el conjunto de reglas que forman la base de conocimiento. *KeyManager* organiza el conocimiento en una base de datos de estructura análoga a las tablas de decisión en la que distinguimos:

- Tablas de datos. Durante la adquisición del conocimiento, se genera una tabla de datos por cada grupo taxonómico que se pretende representar (una por cada tabla de decisión).
- Tabla de información. Realiza la función de *fichero de configuración* y contiene información sobre las tablas de datos almacenadas en la base de datos.
- Tabla de traducción. Describe el conjunto de nombres válidos dentro de la base de datos.

EL ALGORITMO ID3.

El acrónimo TDIDT (*Top-Down Induction On Decision Trees*) hace referencia a todos los algoritmos *divide y vencerás* que construyen árboles de decisión desde la raíz hasta las hojas de forma recursiva. Hemos aplicado un algoritmo de este tipo para obtener, de forma automática y a partir de las tablas de decisión, el conjunto de reglas que forma la base de conocimiento.

Durante el proceso de identificación taxonómica es frecuente no disponer de toda la información necesaria que lleva desde la raíz del árbol de decisión hasta un objetivo. Por ejemplo, el fruto es un carácter determinante en muchos grupos taxonómicos, pero es posible no disponer del mismo por no ser la época del año adecuada. Por desgracia, no podemos construir todos los posibles árboles de decisión derivados de un conjunto de casos de entrenamiento para quedarnos con el más adecuado[11].

Bajo la suposición de que la información de entrada no siempre es completa, nuestro algoritmo genera varios modelos de clasificación para un mismo conjunto de ejemplos. Para ello hemos modificado el algoritmo ID3 propuesto por Quinlan [Quinlan, 1986] que genera un único modelo de clasificación de los datos de entrada. Así, para construir un modelo de clasificación M_2, supondremos que no se dispone del carácter seleccionado para ramificar por primera vez el árbol de clasificación M_1 (ver ejemplo de la Figura3-3).

[11] Este problema es *NP* completo [Wang *et al.,* 2000]). La construcción de árboles de decisión se suele realizar de forma descendente mediante algoritmos *greedy* de eficiencia de orden *O(nlogn)*.

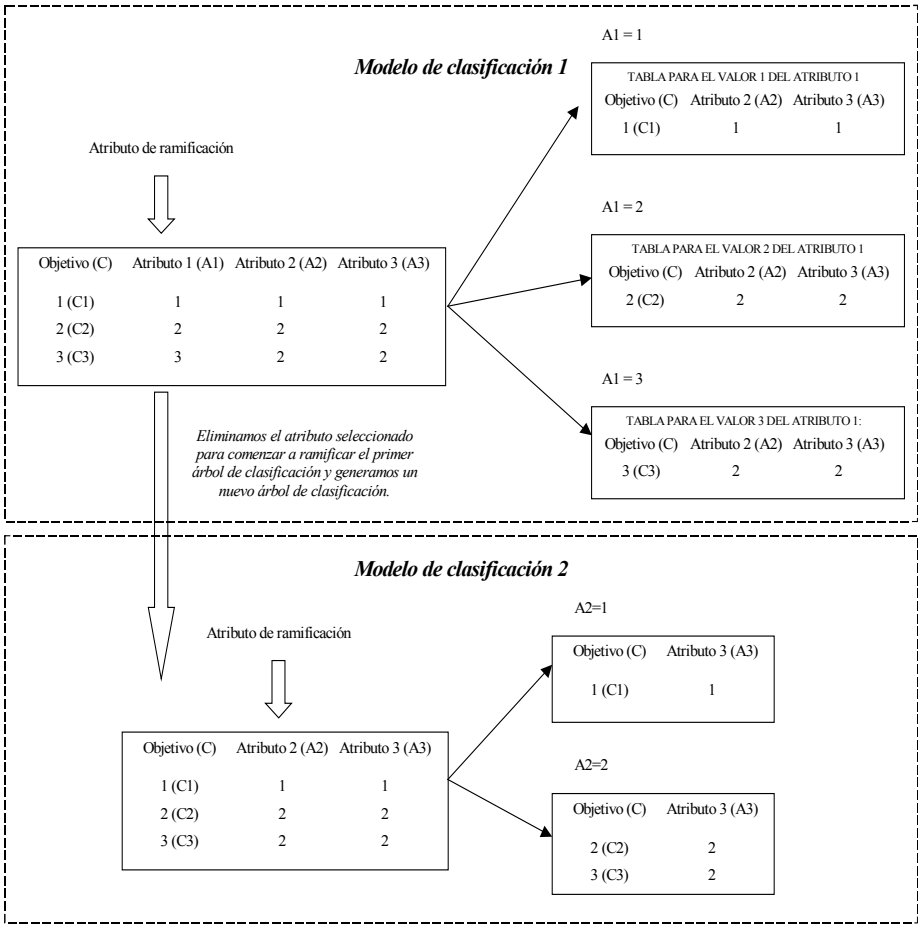


Figura3-3. *Generación de varios modelos de clasificación.*

Dadas las peculiaridades de la información taxonómica, un algoritmo para generar claves de identificaión debe incluir atributos que [Dallwitz *et al.,* 2000]:

- Dividan los *taxa* en subgrupos más uniformes.
- Tengan un valor alto de utilidad (*reliability*) [12] asociado.
- Discriminen en los primeros pasos de la clave aquellos individuos que van a ser identificados con más frecuencia. Esta frecuencia se denomina abundancia (*abundance*) de un taxon.
- Presenten poca variabilidad intra-taxon. Este caso se produce cuando el valor de un carácter apenas varía dentro de un mismo grupo taxonómico.

[12] El concepto de utilidad (*reliability*) [Dallwitz *et al.*, 2000] es equivalente al de peso de un carácter. Un carácter con un alto peso asociado será enfatizado de alguna forma por la aplicación que lo utiliza.

El algoritmo ID3 utiliza la entropía como regla de división. Esta regla satisface los requisitos que acabamos de especificar y pretende maximizar la ganancia de información conseguida al utilizar el atributo A_i para ramificar el árbol de decisión mediante la minimización de la función *I* de la Fórmula 3-1:

$$I(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) H(C \mid A_{ij}) = \sum_{j=1}^{M_i} p(A_{ij}) \left(- \sum_{k=1}^{J} p(C_k \mid A_{ij}) \log_2 p(C_k \mid A_{ij}) \right)$$

Fórmula 3-1. *Criterio de división del algoritmo ID3***.**

Donde:

- A_i es el atributo para ramificar el árbol.
- M_i es el número de valores diferentes del atributo A_i.
- $p(A_{ij})$ es la probabilidad de que el atributo *i* tome su *j-ésimo* valor.
- $H(C \mid A_{ij})$ es la entropía de clasificación del conjunto de ejemplos en los que el atributo A_i toma su *j-ésimo* valor.
- *J* es el número de clases del problema.
- $p(C_k \mid A_{ij})$es una estimación de la probabilidad de que un ejemplo pertenezca a la clase C_k cuando su atributo A_i toma su *j-ésimo* valor. Esta estimación no es más que la frecuencia relativa, $f(C_k \mid A_{ij})$, en el conjunto de entrenamiento utilizado.

La expresión $H(C \mid A_{ij})$ mide la impureza de un nodo del árbol y alcanza su máximo cuando la composición del nodo es menos uniforme y su mínimo cuando todos los elementos de un nodo son iguales. Esto hace que la entropía divida los *taxa* en grupos más uniformes y minimice la variabilidad intra-taxon. Su expresión también tiene en cuenta la abundancia de un individuo, de forma que aparecerán en los primeros pasos de la clave los individuos más frecuentes.

La aplicación de la entropía como regla de división, permite minimizar la cantidad de información necesaria para generar el árbol de clasificación, por lo que las reglas obtenidas serán de longitud mínima. El algoritmo para generar la

base de conocimiento obtiene un conjunto de reglas que permite identificar un determinado individuo con poca información y cuyo contenido es más completo que el de las claves dicotómicas, pues genera más de un árbol de clasificación.

TRATAMIENTO DE LOS CARACTERES DIFERENCIADORES.

Hay algunos valores de caracteres que sólo se manifiestan en un determinado grupo taxonómico. Por ejemplo sólo la especie "*Abies pinsapo*" presenta "*disposición inclusa de las escamas tectrices*". Estos caracteres se denominan caracteres diferenciadores (*differenciating attributes* [Dallwitz, 2000 b]). Estos descriptores son poco deseables para la clasificación porque solo diferencian a un individuo del resto, pero son muy útiles para la identificación porque permiten llegar a la determinación en un solo paso.

La generación de reglas es un proceso de clasificación en el que los caracteres diferenciadores aparecen en los últimos niveles de encadenamiento. Al generar reglas de forma jerárquica, estos caracteres aparecen en las tablas de decisión de especies. Por este motivo, el conjunto de reglas obtenido no permite inferir con un único carácter, por ejemplo, que un individuo es un "*Abies pinsapo*" a partir de la observación de que la disposición de las escamas de su piña es inclusa. Antes habría que inferir que se trata de la familia *Pinaceae* y del género *Abies*.

Por este motivo, hemos modificado la técnica de obtención de reglas de forma que, cada vez que se genere una regla para un objetivo, *O*, con un carácter diferenciador, se generará un conjunto de reglas complementario, que contiene una regla adicional por cada nivel taxonómico superior al del objetivo *O*. De esta forma, el encadenamiento de reglas permitirá concluir el objetivo *O* sin necesidad de conocer más información que la observación del carácter diagnóstico. En nuestro ejemplo del "*Abies pinsapo*" (ver Figura 3-4) se genera una regla para la especie y dos reglas más, una para el género y otra para la familia.

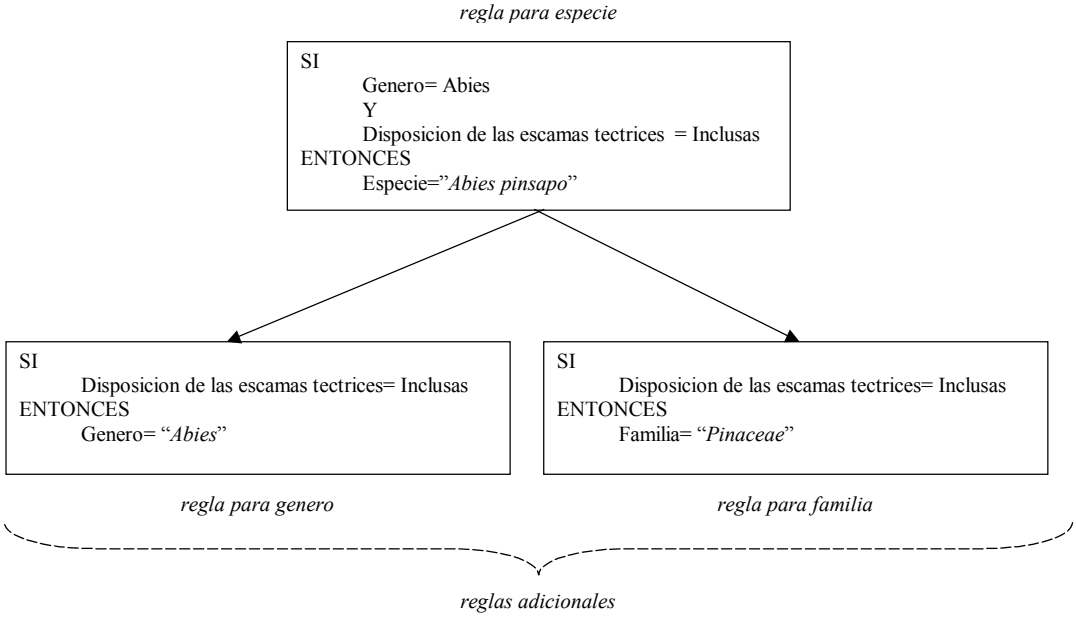


Figura 3-4. *Reglas generadas para especie, género y familia utilizando el atributo diferenciador "Disposición de las escamas tectrices = Inclusas".*

1.4 Tratamiento de la incertidumbre.

El conocimiento experto no siempre se define con total certeza, y durante la sesión de consulta algunos datos no son conocidos con total seguridad; a esto hemos de añadir que se pueden cometer errores de medida y que, aunque la información sobre el dominio del problema se basa en lo que sucede normalmente, la única regla válida en la naturaleza es la excepción. No en vano, algunos investigadores sostienen que "*la limitada capacidad de los sistemas basados en el conocimiento para tratar con incertidumbre es la causa de que disminuya su rendimiento*" [Mamdami & Efstathiou, 1985].

Por este motivo existen diversas teorías para el tratamiento de la incertidumbre como la teoría de la evidencia [Dempster, 1967; Shafer, 1976] , la teoría de la probabilidad o la teoría de conjuntos difusos [Zadeh, 1965]. El método utilizado depende, entre otros aspectos, del tipo de sistema, la naturaleza de la incertidumbre con que tratamos, la disponibilidad de información histórica y los requerimientos de eficiencia computacional. Dadas las características de la identificación taxonómica, así como la ausencia de datos históricos y las

dificultades para definir variables difusas y funciones de masa de probabilidad, hemos seleccionado la teoría de los factores de certeza [Shortlife & Buchanan, 1975]. Este modelo ha sido utilizado con éxito en muchos otros sistemas de reciente publicación [Mahaman *et al.*, 2002; Cabrero-Carnosa *et al.,* 2003].

UTILIZACIÓN PRÁCTICA DE LOS FACTORES DE CERTEZA.

La teoría de los factores de certeza es un modelo computacionalmente simple y más natural para un usuario no matemático que otros enfoques [Harrison & Kovalchic, 1998]. Además permite a los expertos la expresión de estimaciones subjetivas de certeza y establecer resultados intermedios que diferencian qué respuestas del sistema son más adecuadas. Este modelo también facilita la representación del conocimiento en forma de reglas y el encadenamiento sencillo del efecto de varias reglas que tienen como consecuente el mismo objetivo.

Cada regla de la base de conocimiento tiene un factor de certeza, CF (*certainty factor*), asociado que se determina durante la generación de la misma y no cambia, a no ser que un experto redefina este valor. Cuando aplicamos técnicas de aprendizaje automático se sigue el siguiente criterio para asignar un factor de certeza a cada regla:

- Si se genera un nodo hoja puro (todos los elementos a clasificar pertenecen a la misma clase), la regla asociada tiene un factor de certeza asociado cuyo valor es +1.
- En el caso en que el nodo hoja no es puro, el resultado es el de una regla con disyunción en el consecuente, que se fracciona en reglas con consecuentes simples. El factor de certeza de la regla original, también se fracciona atendiendo a la Fórmula 3-2.

$$CF = \begin{cases} \dfrac{\min\{p(x),\, p(x \mid a) - p(x)\}}{1 - p(x)} & si \quad p(x) \le p(x \mid a) \\ \dfrac{\max\{p(x),\, p(x \mid a) - p(x)\}}{1 - p(x)} & si \quad p(x) > p(x \mid a) \end{cases}$$

Fórmula 3-2. *Cálculo del factor de certeza de una regla con OR en el consecuente.*

Donde:

- *p(x)* es la probabilidad del objetivo en cuestión en la tabla inicial.
- *p(x/a)* es la probabilidad de que se dé el objetivo *x* sabiendo que se da el antecedente *a*, sea este simple o compuesto.

Durante la sesión de consulta al sistema distinguimos la certeza de los antecedentes, C(A), medida sobre el grado de creencia del usuario en sus observaciones y la certeza de la conclusión C(X). Cuando el antecedente de una regla es cierto, puede utilizarse para computar un nuevo valor de certeza para su conclusión según la expresión de la Fórmula 3-3.

$$C(X/A) = \begin{cases} = C(X) + CF * (1 - C(X)) & si \quad C(X) \quad y \quad CF \quad \geq 0 \\ = C(X) + CF * (1 + C(X)) & si \quad C(X) \quad y \quad CF \quad < 0 \\ = \dfrac{C(X) + CF}{1 - \min(|C(X)|, |CF|)} & si \quad C(X) \quad y \quad CF \quad de \quad dist\,into \quad signo \end{cases}$$

Fórmula 3-3. *Cálculo de la certeza de la conclusión.*

En el caso en que la certeza de los antecedentes no valga 1, calculamos un nuevo CF para la regla proporcional a la certeza del antecedente (Fórmula 3-4).

$$CF' = CF * C(A)$$

Fórmula 3-4. *Cálculo del CF cuando la certeza del antecedente no tiene el valor 1.*

En aquellos casos en que el antecedente tiene condiciones complejas (está formado por varias cláusulas conectadas por conjunciones o disyunciones), la certeza del mismo viene determinado por el conjunto de expresiones de la Fórmula 3-5.

$$C(A \wedge B) = \min\{C(A), C(B)\}$$
$$C(A \vee B) = \max\{C(A), C(B)\}$$
$$C(\neg A) = -C(A)$$

Fórmula 3-5. *Cálculo de la certeza de los antecedentes con condiciones complejas.*

1.5 Mantenimiento de la consistencia.

Durante el desarrollo de la base de conocimiento pueden aparecer inconsistencias debido a fallos durante la fase de adquisición y elicitación del conocimiento. Estas incoherencias influyen en los resultados alcanzados por el sistema, sin olvidarnos de que añaden un impacto adicional debido a la utilización de factores de certeza. Todo ello hace necesario que *GREEN* incorpore un reforzador de consistencia que analiza sistemáticamente cada una de las reglas de la base de conocimiento para detectar y corregir inconsistencias. De este modo, garantizamos que la base de conocimiento ha sido correctamente diseñada e implementada y que cumple con los requerimientos de completitud (*completeness*) y consistencia (*consistency*).

Sea *n* el número de reglas de la base de conocimiento
DESDE i=1 HASTA i=*n*-1
  DESDE j=i HASTA j= *n*
    Si $regla_i$ y $regla_j$ tienen el mismo consecuente
      Si $regla_i$ y $regla_j$ tienen el mismo número de antecedentes
        Si $CF(regla_i)=CF(regla_j)$
- Todos los antecedentes son iguales salvo uno que es contradictorio: CONDICION IF INNECESARIA ⇨ Eliminar la $regla_j$, y eliminar la condición innecesaria de la $regla_i$.
- Todos los antecedentes son iguales: REGLAS REDUNDANTES ⇨ Eliminar la $regla_j$.

        Si $CF(regla_i)\neq CF(regla_j)$
          Todos los antecedentes son iguales

            $CF(regla_i) = -CF(regla_j)$: REGLAS CONFLICTIVAS ⇨ Eliminar las dos reglas.
      Si $regla_i$ y $regla_j$ tienen distingo número de antecedentes
        CF de la regla más corta es ±1

          Los antecedentes de la regla más corta están contenidos en los de la más larga entonces REGLAS SUBSUMIDAS ⇨. Eliminar la regla más larga.

Tabla 3-2. *Algoritmo para analizar la consistencia de la base de conocimiento.*

Bajo la suposición de que la sintaxis de una regla es suficientemente restrictiva, es posible analizar cualesquiera dos reglas de la base de conocimiento para detectar problemas. Teniendo en cuenta esta premisa, hemos comprobado la presencia en la base de conocimiento de reglas redundantes, conflictivas, subsumidas y condiciones IF innecesarias. El algoritmo propuesto para realizar esta verificación pretende minimizar el número de veces que hay que recorrer la

base de conocimiento y el número de reglas que hay que revisar (ver Tabla 3-2). Para esto determina si una regla puede ser o no una fuente de inconsistencias en función de su sintaxis y antes de analizarla.

1.6 El motor de inferencia.

La inferencia es el proceso utilizado en un sistema experto para derivar nueva información mediante la combinación de los hechos de entrada con el contenido de la base de conocimiento. El motor de inferencia de *GREEN*, es un módulo bien diferenciado de la base de conocimiento, esta separación:

- Hace posible representar el conocimiento de forma más natural. Un modelo de conocimiento separado del proceso de inferencia refleja mejor el mecanismo de resolución seguido por un ser humano que un modelo que incrusta conocimiento dentro del proceso de inferencia.
- Permite captar y organizar el conocimiento independientemente del procesamiento que posteriormente se realice. Como veremos en el Capítulo 3.2, *GREEN* puede operar con conjuntos de datos desarrollados en el Herbario y con conjuntos de datos desarrollados por otros expertos con diferentes modelos para representación de conocimiento taxonómico.
- Permite cambiar el contenido de la base de conocimiento sin necesidad de cambiar el sistema de control y utilizar el mismo motor de inferencia para solucionar problemas diferentes.

PRIMERA APROXIMACIÓN.

Para diseñar el motor de inferencia de *GREEN* tomamos como referencia el modelo con encadenamiento hacia delante propuesto por Ignizio [Ignizio, 1991], que basa su funcionamiento en la utilización de tres estructuras de datos:

- MEMORIA DE TRABAJO O PIZARRA (*Working memory*): Almacena los hechos deducidos durante la consulta.

- COLA DE ATRIBUTOS (*Attribute queue*): Almacena, en orden, los antecedentes cuyo valor necesita conocer el motor de inferencia.

- TABLA DE REGLAS Y PREMISAS (*Rule/premise table*): Recuerda el estado de cada regla y el de cada premisa.

Este modelo no presenta tratamiento de la incertidumbre y pregunta al usuario cada vez por un único atributo, aspecto que resulta ineficiente en Internet. Por este motivo, hemos desarrollado un motor de inferencia más adaptado al dominio de la Taxonomía, que incluye el tratamiento de incertidumbre mediante factores de certeza y pregunta por el valor de varios atributos para reducir el trasiego de información entre el cliente y el servidor.

Nuestro diseño utiliza sólo dos de las estructuras de datos citadas con anterioridad: la tabla de reglas y premisas y la memoria de trabajo. Inicialmente todas las reglas están en estado activo, sus cláusulas tienen el valor *false* y la memoria de trabajo almacena las certezas de los atributos de entrada introducidos por el usuario (en caso de ser desconocidas, se asigna el valor 0). Durante su ejecución, comprueba regla por regla si se cumplen sus antecedentes es decir, si en la tabla de certezas y atributos el valor de la certeza de todos los antecedentes de la regla es distinto de cero. En este caso se dispara la regla y, utilizando la expresión de la Fórmula 3-3, se actualiza la certeza que hasta entonces tenía asociada el consecuente. El motor de inferencia termina su ejecución cuando se cumple una de estas dos condiciones:

1. No quedan reglas activas. Esto quiere decir que se han disparado todas las reglas, de modo que no tiene sentido continuar la inferencia.

2. No se ha conseguido disparar ninguna regla en una pasada. Si en una pasada el motor no infiere nada nuevo, en posteriores pasadas tampoco inferirá nada.

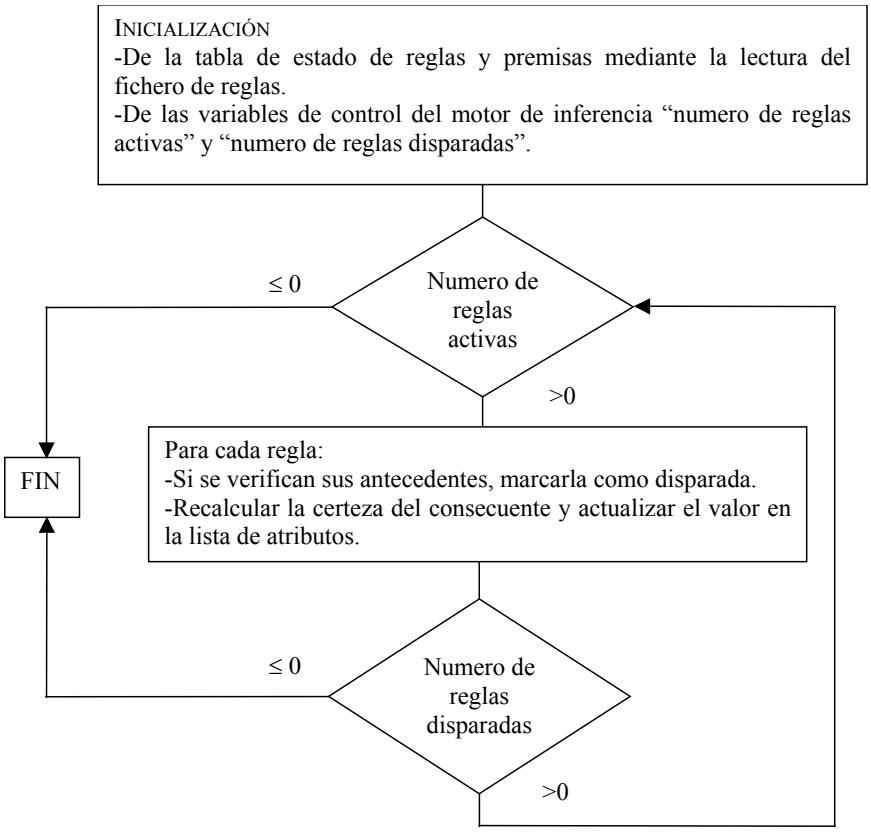


Figura 3-5. *Diagrama de flujo del motor de inferencia de GREEN. Primera aproximación.*

MOTOR DE INFERENCIA MEJORADO.

La primera aproximación del motor de inferencia utiliza una estrategia de razonamiento hacia delante (de los datos a las conclusiones). Esta estrategia es adecuada cuando se consulta al sistema sin ningún tipo de hipótesis previa sobre la conclusión. No obstante, puede ser habitual la consulta para probar una determinada suposición. En este caso es más adecuada una estrategia de razonamiento hacia atrás (de los objetivos a los datos), de forma que el proceso de búsqueda está más orientado y se pueden acotar las preguntas que el sistema realiza al usuario.

Por este motivo, hemos dotado al sistema, de la capacidad de razonamiento hacia atrás. Durante una sesión de consulta el usuario puede combinar estas dos estrategias. El diagrama de flujo del motor de inferencia definitivo es el de la (Figura 3-6).

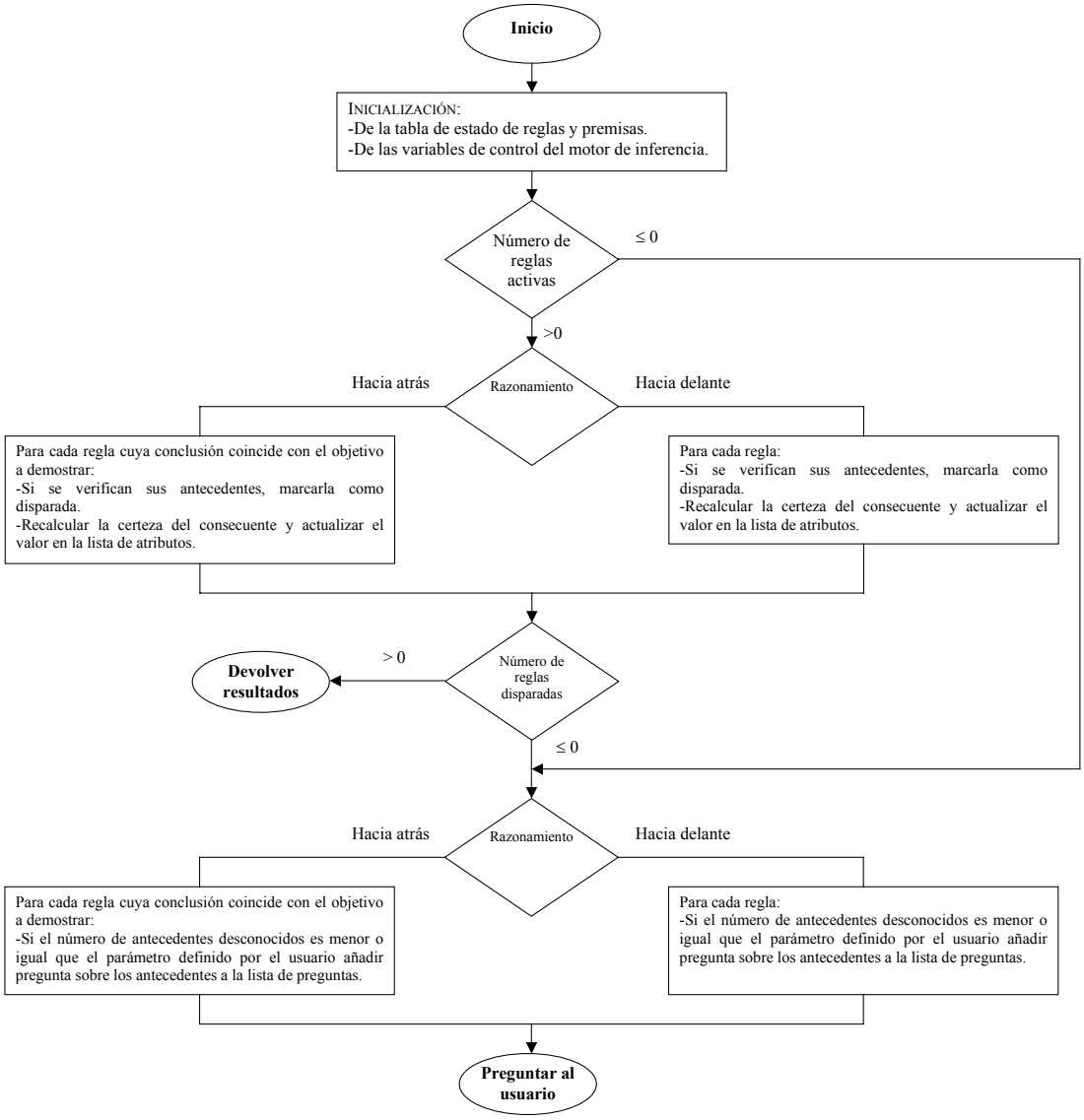


Figura 3-6. *Motor de inferencia mejorado.*

1.7 Otros aspectos relacionados con *GREEN*.

Aparte de lo descrito, destacamos otras características importantes de *GREEN*:

- Su interfaz es fácil de utilizar (ver Figura 3-7). Los descriptores se encuentran agrupados en categorías generales (aspecto general, hoja, rama, piña, etc.) cuya denominación es familiar a cualquier tipo de usuario.

- Tras introducir los datos, se ejecuta la inferencia y el justificador del sistema devuelve el conjunto de resultados ordenados en función de su adecuación a la consulta. También devuelve una traza ordenada del esquema de razonamiento seguido para alcanzar las conclusiones. La justificación de las conclusiones aumenta la confianza del usuario en el sistema y le permite familiarizarse con los caracteres a observar y con el proceso de razonamiento del experto humano que ha aportado su conocimiento al sistema. La justificación de resultados también facilita la depuración de la base de conocimiento.

- Es posible ampliar la información sobre los resultados obtenidos por el sistema accediendo al catálogo multimedia (ver Figura 3-12).

- *GREEN* está diseñado específicamente para trabajar en Internet por lo que la interacción con el usuario se realiza a través de formularios que envían los datos y las consultas a un servidor remoto. Hemos minimizado el trasiego de información a través de la red con el fin de no sobrecargar el servidor y obtener un tiempo de respuesta satisfactorio.

- El diseño modular de *GREEN* es independiente del grupo taxonómico, de modo que puede adaptarse con facilidad a la identificación de especies distintas de las Gimnospermas.

1.8 Ejemplo de funcionamiento.

Supongamos que el usuario ha realizado una observación y que está seguro de que "*la hoja es de tipo escamoso*" y que "*se trata de un individuo resinoso*" (ver Figura 3-8).

Figura 3-7. *Aspecto general de la interfaz del sistema GREEN.*

Figura 3-8. *Introducción del carácter "Características de la hoja".*

Solo con esta información, el sistema puede concluir que se trata de un individuo de la familia *Cupressaceae* con un factor de certeza igual a 1 (ver Figura 3-9).

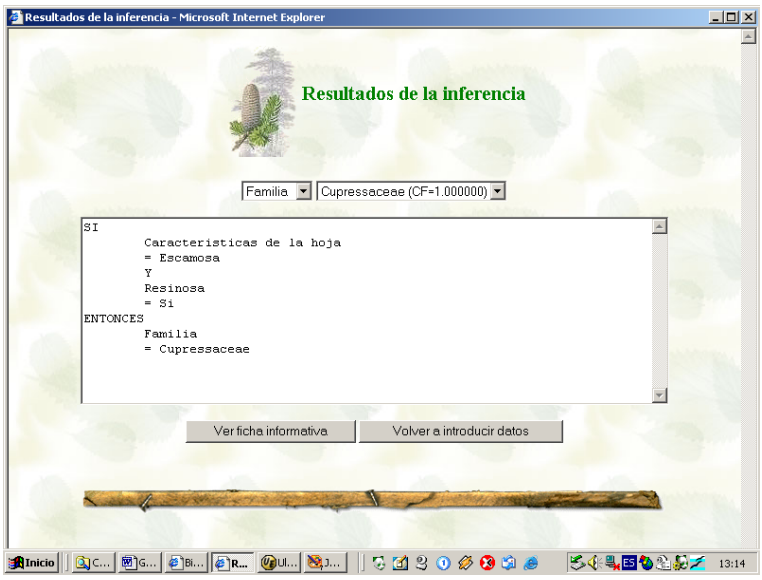


Figura 3-9. *Resultado "Familia=Cupressaceae".*

Si el usuario introduce además, que la fructificación es carnosa, *GREEN* concluye que el individuo observado pertenece al género *Juniperus"* (ver Figura 3-10).

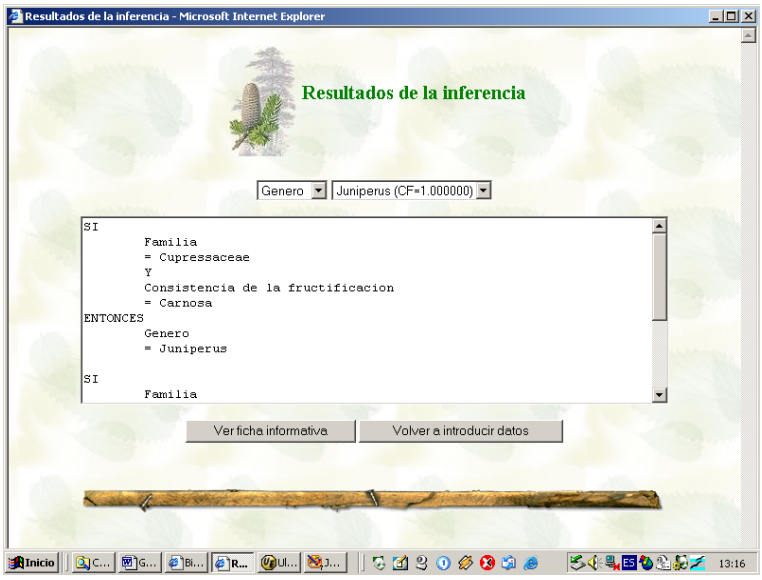


Figura 3-10. *Resultado: "Género=Juniperus".*

Si se completa la información dada al sistema incluyendo que "*el número de semillas de la arcéstida es de 2 a 4*" y que "*el arbusto es de tipo postrado (con*

un factor de certeza de 0.7)", el sistema infiere que el individuo observado es un *Juniperus sabina* con un factor de certeza de 0.7 (ver Figura 3-11).

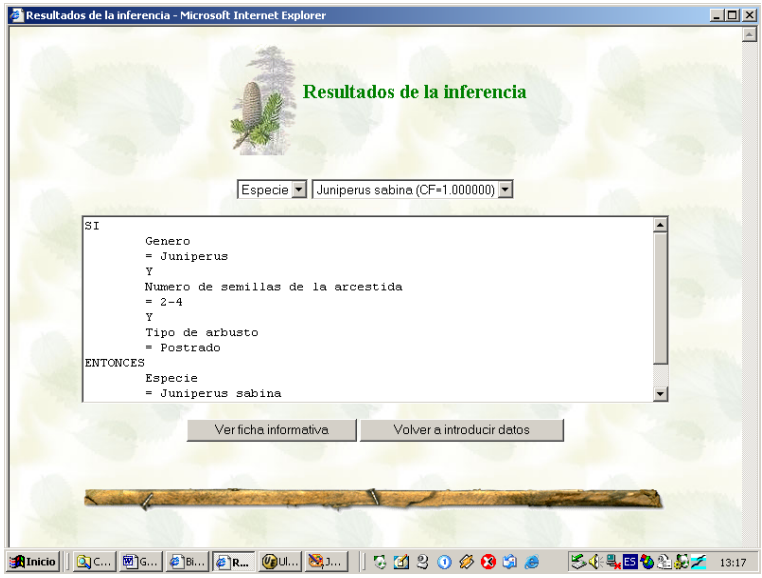


Figura 3-11. *Resultado: "Especie=Juniperus sabina".*

Puesto que hemos generado varios árboles de decisión, el sistema podría haber alcanzado el mismo objetivo con una entrada diferente de datos. Por ejemplo, puede concluir que se trata de un individuo de la familia *Cupressaceae* a partir de "*fructificación es carnosa*", "*Semilla con arilo=No*", "*Hoja persistente*", "*individuo dioico*", "*semillas numerosas = No*" y "*Hoja de tono parduzco*".

Después de introducir la información, el sistema ejecuta el motor de inferencia devolviendo al usuario un conjunto de resultados ordenados en función de su factor de certeza. Para cada resultado se muestra su justificación, esto es, el conjunto de reglas que llevó a su conclusión (ver Figura 3-11).

Si el usuario lo desea, es posible completar esta información con información de carácter general sobre el individuo (nombres vulgares, ecología, distribución geográfica, etc.). Para ello *GREEN* facilita el acceso al catálogo multimedia (ver Figura 3-12).

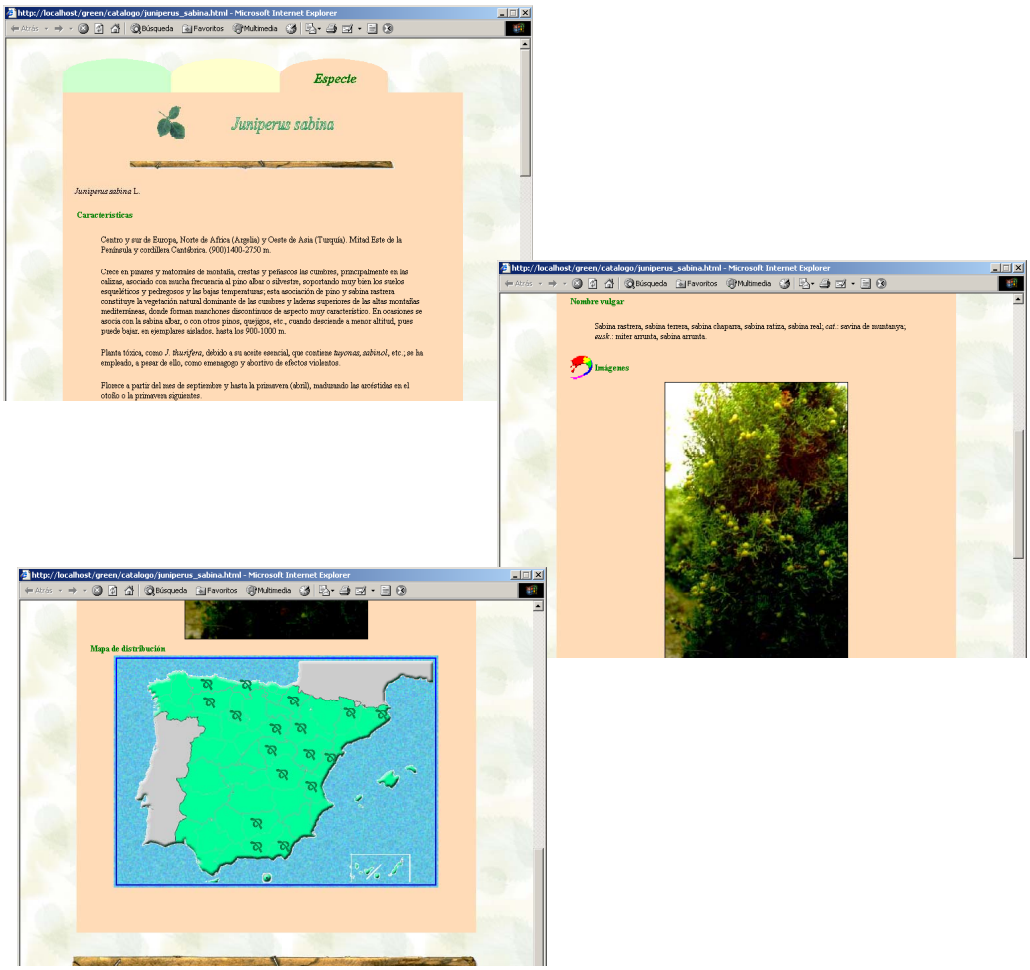


Figura 3-12. *Catálogo multimedia del sistema GREEN.*

2. Estandarización del sistema *GREEN*.

Los paquetes software tradicionales basados en la recopilación de información taxonómica estructurada utilizan modelos de datos diseñados específicamente para dichos sistemas. Esto es fuente de importantes problemas:

- Mantienen diferentes estructuras de datos, y en aquellos casos en que incluyen herramientas de traducción entre modelos se realiza con pérdidas.

- El mantenimiento de los programas de traducción es complejo ya que deben contemplar los cambios hechos a los modelos en ambas direcciones.

- Cuando las plataformas software dejan de ser actualizadas, la información que almacenan se convierte en un legado de conocimiento difícil de mantener y utilizar.

Por todo esto, el papel como estándar del modelo *SDD* descrito con anterioridad es fundamental para permitir el intercambio sin pérdidas de información entre las distintas plataformas dedicadas a la identificación, *data-mining* y bases de datos federadas (ver Figura 3-13).

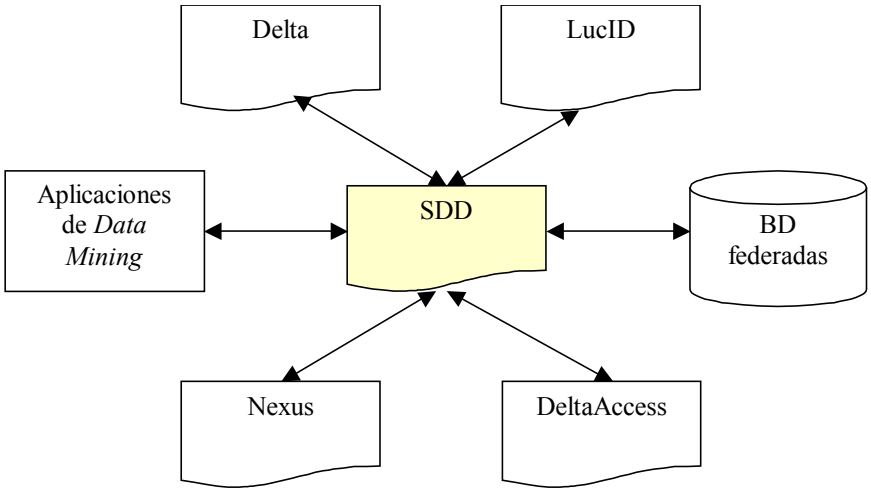


Figura 3-13. *El papel del estándar SDD como intermediario para el intercambio de información taxonómica entre sistemas.*

Debido a lo reciente de este estándar[13], durante el diseño de *GREEN* no se tuvieron en cuenta los aspectos relacionados con el intercambio de información con otros modelos. Tras la aparición de *SDD*, las herramientas relacionadas con la representación estructurada de conocimiento taxonómico deben dar a los sistemas las capacidades de interpretar conocimiento descrito con *SDD* y exportar el conocimiento al nuevo modelo.

[13] La primera versión se fija en la reunión anual del TDWG en Brasil en marzo de 2002.

2.1 Utilidades para la estandarización.

Al estudiar detalladamente la descripción de los distintos modelos de representación y de *SDD*, advertimos que aunque todos representan el mismo tipo de información, no existe una correspondencia directa entre estos. Las principales diferencias son las siguientes.

- *SDD* es un modelo multilingüe que además distingue diferentes niveles de experiencia en los usuarios, de forma que en un mismo proyecto puede cohabitar información en diferentes idiomas y para diferentes tipos de usuarios (expertos, estudiantes universitarios, incluso niños). Esta característica no se presenta en ninguno de los modelos estudiados.
- En *SDD* la información se estructura en forma de árbol, mientras que otros modelos se basan en un modelo de base de datos relacional (*DeltaAccess*, *KeyManager*), o en descripciones textuales más o menos estructuradas (*Nexus*, *Delta*).
- Los diferentes modelos incluyen información con carácter opcional que puede ser obligatoria en otros. Hay que determinar cuales son estos casos y, si falta dicha información, pedirla al usuario y filtrarla adecuadamente.

Los sistemas desarrollados para el Herbario, al igual que el resto de sistemas de identificación taxonómica estudiados, se encuentran en una etapa de actualización para acomodar estos nuevos requerimientos relacionados con la estandarización. Por esto hemos desarrollado dos herramientas:

1. ***KMtoSDD***. Realiza la traducción de *KeyManager* a *SDD* minimizando las pérdidas de información. De este modo, podremos utilizar conjuntos de datos desarrollados para *KeyManager* con aplicaciones basadas en *SDD*.
2. ***SDDtoKM***. Traduce conjuntos de datos escritos en formato *SDD* al modelo de *KeyManager* habilitándolo para utilizar conjuntos de datos desarrollados tomando como base el nuevo estándar.

Dado lo reciente de *SDD*, el esquema de la Figura 3-13 no es aún una realidad. Para incrementar la capacidad de intercambio de información de *GREEN* hemos incorporado otros dos módulos de comunicación:

3. ***DAtoSDD***. Traduce entre *DeltaAccess* y *SDD* y ha sido descrita con anterioridad. Solamente añadir que, dado que *DeltaAccess* es capaz de importar *Delta*, facilita además la traducción entre *Delta* y *SDD*.
4. ***KMtoDelta***. Traduce de *KeyManager* a *Delta* y permite que nuestras herramientas utilicen aplicaciones basadas en *Delta* sin necesidad de rescribir manualmente los conjuntos de datos y minimizando las pérdidas de información.

Con el diseño y desarrollo de estos cuatro protocolos de traducción, comunicamos *GREEN* y *KeyManager* con el nuevo estándar y los principales modelos de representación de información taxonómica (Figura 3-14.).

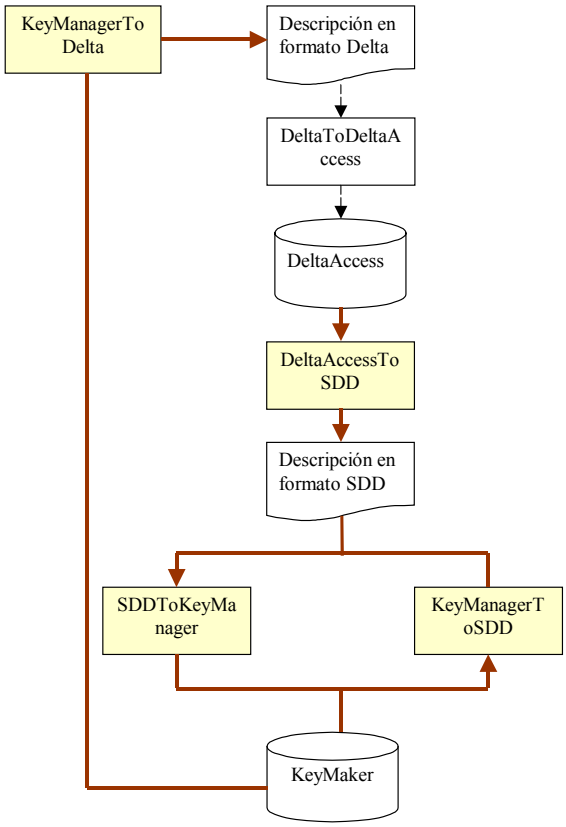


Figura 3-14. *Esquema de los caminos de comunicación establecidos con el desarrollo de las herramientas de traducción.*

Estas facilidades de comunicación permiten la generación de bases de conocimiento para *GREEN* a partir de otros modelos de representación como *Delta*, *DeltaAccess* y *SDD*. Para facilitar la lectura del documento hemos

trasladado la correspondencia entre estos modelos de representación del conocimiento al Apéndice A.

JSDD, UN PAQUETE PARA ALMACENAR *SDD* EN UNA ESTRUCTURA ORIENTADA A OBJETOS.

Para procesar documentos *SDD*, hemos desarrollo un conjunto de clases en Java que realizan la lectura del documento y lo almacenan en memoria en una estructura orientada a objetos. Estas clases se han agrupado en un paquete de gran utilidad, JSDD, que actúa como intermediario entre XML y cualquier otro tipo de aplicación. Por ejemplo, las herramientas *KMToSDD*, *SDDtoKM* y *XKey* (que veremos más adelante) hacen uso del mismo.

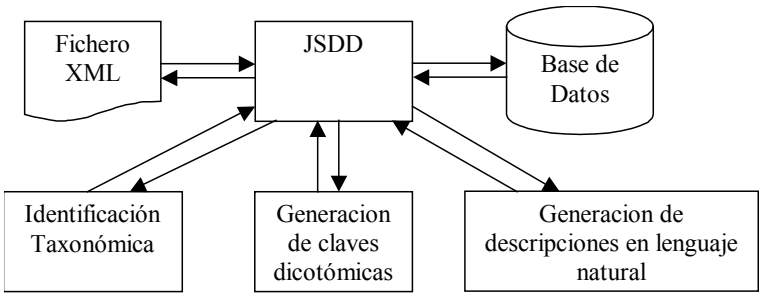


Figura 3-15. *Utilidad del paquete JSDD.*

2.2 Generación automática de la interfaz.

Al realizar una revisión de *GREEN* se advierte que, aunque los módulos para la generación de reglas y para la inferencia de conocimiento son genéricos y modulares, la interfaz está desarrollada y adaptada para trabajar con un grupo muy concreto. Este aspecto:

- Implica generar otra interfaz específica para identificar otros grupos taxonómicos.
- Puede conllevar continuos cambios en la interfaz del sistema, por ejemplo, al añadir o eliminar un carácter o estado.

Para dar a *GREEN* mayor flexibilidad hemos desarrollado una actualización que le permite generar su interfaz de forma automática, independientemente del conjunto de datos. El sistema utiliza un fichero de configuración que incluye toda la información necesaria para lograr este objetivo:

- El número de niveles de inferencia del mismo así como su nombre (por ejemplo, tres niveles de inferencia: familia, género y especie).
- El nombre del fichero con las reglas generadas para el conjunto de datos.
- El nombre del fichero de traducción para el conjunto de datos.
- El nombre del fichero con información para generar de forma automática el catálogo para el conjunto de datos.
- El nombre del fichero que será la hoja de estilo que determine el aspecto final de la interfaz para el conjunto de datos.

El resultado es una *shell* operativa y compatible con la primera versión del sistema y con los modelos de representación de conocimiento más relevantes, lo que permite realizar consultas de identificación sobre un amplio espectro de grupos taxonómicos (ver Figura 3-16 y Figura 3-17).

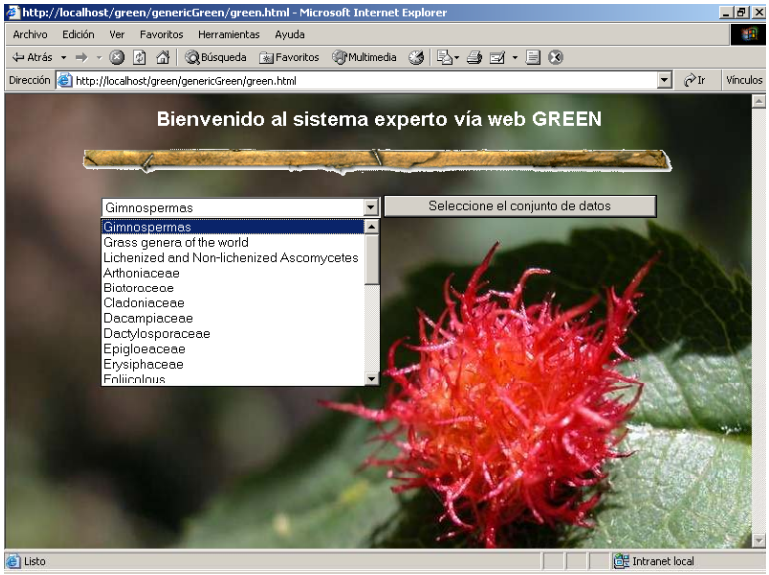


Figura 3-16. *Interfaz de selección del conjunto de datos.*

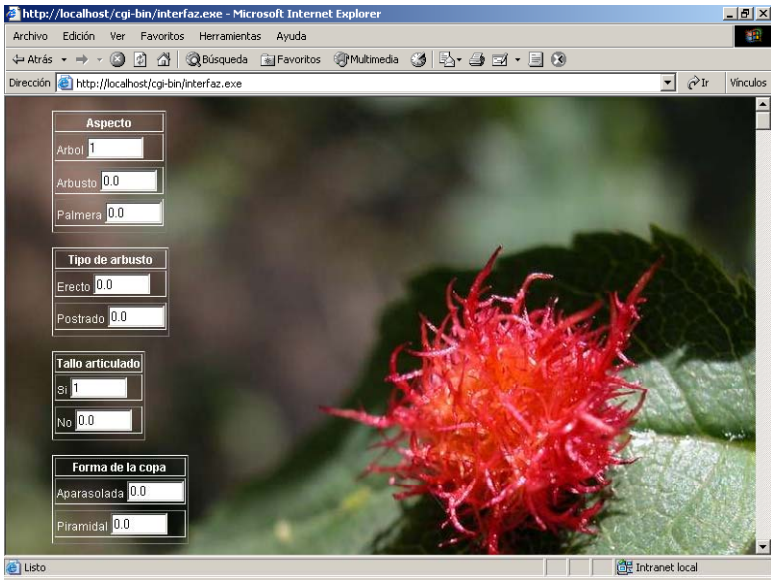


Figura 3-17. *Interfaz genérica del sistema Green.*

3. Generación interactiva de claves de identificación: la herramienta *XKEY*.

Hasta ahora, los esfuerzos se están centrando en el diseño de un *SDD* completo y funcional y, debido a lo reciente de este proyecto, no existen herramientas completamente funcionales basadas en el mismo; más bien se han desarrollado protocolos de comunicación entre modelos. La evolución natural de esta línea de investigación es el diseño y desarrollo de herramientas completamente basadas en el estándar. La representación de la información taxonómica nos proporciona un mecanismo para describir aquello que conocemos, el paso siguiente es utilizar estas descripciones para desarrollar claves de identificación que nos permitan determinar qué estamos observando en un momento determinado.

Las claves de identificación tienen especial importancia en proyectos como la elaboración de guías de campo y floras en los que es imprescindible la inclusión de un conjunto de claves de identificación. Particularmente, el

Departamento de Biología de la Universidad de Granada coordina el proyecto de elaboración de la flora de Andalucía oriental, proyecto íntimamente ligado a la representación de información taxonómica. El Herbario es el responsable de centralizar e informatizar la información para este proyecto.

Por lo general, el experto botánico diseña estas claves sin utilizar ninguna herramienta informática de apoyo. Esto implica invertir mucho tiempo en el desarrollo de una clave, y un alto coste cuando se detecta algún error en la misma: cuanto más tarde se detecte un error, más costoso será subsanarlo.

Por todo lo expuesto, hemos desarrollado *XKey*, una herramienta totalmente basada en *SDD* que combina técnicas de Inteligencia Artificial de probada eficacia con las necesidades de funcionalidad específicas del área de conocimiento en la que estamos trabajando. *XKey* además:

- Genera bases de conocimiento para el sistema *GREEN* y otras *shells* de sistemas expertos.
- Opera directamente con conjuntos de datos *SDD* sin ninguna transformación previa.
- Mantiene la compatibilidad con el trabajo previamente desarrollado porque permite operar, además, a partir de conjuntos de datos representados con la estructura requerida por *KeyManager*.

3.1 Generación de árboles de decisión.

El proceso de construcción de claves de identificación conlleva la generación de un árbol de decisión. Un árbol de decisión se construye de forma recursiva siguiendo una estrategia descendente, desde conceptos generales hasta ejemplos particulares. El método de construcción de árboles de decisión mediante particiones recursivas del conjunto de casos de entrenamiento tiene su origen en el trabajo de Hunt a finales de los años 50. Este algoritmo *divide y vencerás* es simple y elegante:

- Si existen uno o más casos en el conjunto de entrenamiento y todos ellos corresponden a objetos de la misma clase $c \in Dom\ (C)$, el árbol de decisión es una hoja etiquetada con la clase *c*. Hemos alcanzado un nodo puro.
- Si no encontramos ninguna forma de seguir ramificando el árbol o se cumple alguna condición de parada (*regla de parada*), no se sigue expandiendo el árbol por la rama actual. Se crea un nodo hoja etiquetado con la clase más común del conjunto de casos de entrenamiento que corresponden al nodo actual. Si el conjunto de casos de entrenamiento queda vacío, la clasificación adecuada ha de determinarse utilizando información adicional (por ejemplo, C4.5 opta por la clase más frecuente del nodo padre).
- Cuando en el conjunto de entrenamiento hay casos de distintas clases, éste se divide en subconjuntos que sean o conduzcan a agrupaciones uniformes de casos, entendiendo por éstas conjuntos de casos correspondientes a una misma clase. Utilizando los casos de entrenamiento disponibles, hemos de seleccionar una pregunta para ramificar el árbol de decisión. Dicha pregunta, basada en los valores que toman los atributos predictivos del conjunto de entrenamiento, ha de tener dos o más respuestas alternativas mutuamente excluyentes R_i. De todas las posibles alternativas, se selecciona una empleando una regla heurística a la que se denomina *regla de división*. El árbol de decisión resultante consiste en un nodo que identifica la pregunta realizada del cual cuelgan tantos hijos como respuestas alternativas existan. El mismo método utilizado para el nodo se utiliza recursivamente para construir los subárboles correspondientes a cada hijo del nodo, teniendo en cuenta que al hijo H_i se le asigna el subconjunto de casos de entrenamiento correspondientes a la alternativa R_i.

En resumen, cuando se construye o expande un nodo, se considera el subconjunto de casos de entrenamiento que pertenecen a cada clase. Si todos los ejemplos pertenecen a una clase o se verifica alguna regla de parada, el nodo es una hoja del árbol. En caso contrario, se selecciona una pregunta basada en los atributos del conjunto de entrenamiento (utilizando una regla de división heurística), se divide el conjunto de entrenamiento en subconjuntos y se aplica el mismo procedimiento a cada subconjunto del conjunto de entrenamiento.

Por lo general, se busca la obtención de un árbol de decisión que sea compacto. Un árbol de decisión pequeño nos permite comprender mejor el modelo de clasificación obtenido y, además, es probable que el clasificador más simple sea el correcto, de acuerdo con el principio de economía de Occam (también conocido como navaja de Occam): *"los entes no han de multiplicarse innecesariamente"*. Este principio, si bien permite la construcción de modelos fácilmente comprensibles, no garantiza que los modelos así obtenidos sean mejores que otros aparentemente más complejos [Domingos, 1998; Domingos, 1999].

CRITERIOS DE DIVISIÓN.

Una regla de división muy utilizada es la entropía. Como ya hemos comentado, permite obtener un árbol de decisión de longitud mínima que clasifica los objetivos a partir del menor número posible de atributos. En este sentido, la entropía refleja la forma de trabajo del experto, que con poca información es capaz de discriminar entre distintos objetivos.

Por otra parte, esta heurística suele construir árboles de decisión con un grado de ramificación elevado (favorece aquéllas preguntas que tienen más resultados posibles). Este aspecto no siempre refleja la metodología de trabajo del experto botánico, quien dependiendo del objetivo con que desarrolla la clave puede tener otros criterios diferentes a la minimización de la longitud. Existen otros criterios de división, a continuación describimos dos muy utilizados.

El criterio de proporción de ganancia.

Es una medida basada también en la entropía. La idea es recurrir nuevamente a la Teoría de la Información para normalizar de algún modo la ganancia de información obtenida. El contenido de un mensaje que nos indique la respuesta a la pregunta realizada (no la clase a la que pertenece cada caso) es igual a $-\sum p(A_{ij})\log_2 p(A_{ij})$. Utilizando este resultado podemos definir el siguiente criterio de división:

$$R(A_i) = \frac{H(C) - \sum_{j=1}^{M_i} p(A_{ij}) H(C \mid A_{ij})}{-\sum p(A_{ij}) \log_2 p(A_{ij})} = \frac{H(C) - \sum_{j=1}^{M_i} p(A_{ij}) \left(-\sum_{k=1}^{J} p(C_k \mid A_{ij}) \log_2 p(C_k \mid A_{ij}) \right)}{-\sum p(A_{ij}) \log_2 p(A_{ij})}$$

Fórmula 3-6. *Expresión para calcular el criterio de proporción de ganancia.*

Este criterio de división es el utilizado por C4.5 [Quinlan, 1993]. Cuando la división realizada del conjunto de casos de entrenamiento es trivial, el denominador de $R(A_i)$ es cercano a cero. Por tanto, se ha de escoger el atributo que maximice el cociente $R(A_i)$ siendo su ganancia, al menos tan grande como la ganancia media de todas las alternativas analizadas.

Dado que en la práctica hemos de disponer de muchos más casos de entrenamiento que clases diferentes, el criterio de proporción de ganancia evitará la construcción de árboles de decisión que clasifiquen los casos utilizando sus claves.

Se ha observado que el criterio de proporción de ganancia tiende a la construcción de árboles poco equilibrados, característica que hereda de la regla de división de la que se deriva (ganancia de información). Ambas heurísticas se basan en una medida de entropía que favorece particiones del conjunto de entrenamiento muy desiguales en tamaño cuando alguna de ellas es de gran pureza (todos los casos que incluye corresponden a una misma clase) aun siendo poco significativa (es decir, abarcando muy pocos casos de entrenamiento).

El índice de diversidad de Gini.

El índice de diversidad de Gini es una medida de la diversidad de clases en un nodo del árbol que trata de minimizar la impureza existente en los subconjuntos de casos de entrenamiento generados al ramificar el árbol de decisión. Es una medida de impureza muy utilizada en distintos algoritmos de

construcción de árboles de decisión. En concreto es la utilizada en CART [Breiman *et al.*, 1984].

La función empleada es la siguiente:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C \mid A_{ij}) = \sum_{j=1}^{M_i} p(A_{ij}) \left(- \sum_{k=1}^{J} p(C_k \mid A_{ij}) p(\neg C_k \mid A_{ij}) \right)$$
$$= \sum_{j=1}^{M_i} p(A_{ij}) \left(1 - \sum_{k=1}^{J} p^2 (C_k \mid A_{ij}) \right)$$

Fórmula 3-7. *Expresión para calcular el índice de diversidad de Gini.*

Donde:

- A_i es el atributo para ramificar el árbol.
- J es el número de clases del problema.
- M_i es el número de valores diferentes del atributo A_i.
- $p(A_{ij})$ es la probabilidad de que el atributo *i* tome su *j-ésimo* valor.
- $p(C_k \mid A_{ij})$ es una estimación de la probabilidad de que un ejemplo pertenezca a la clase C_k cuando su atributo A_i toma su *j-ésimo* valor.
- $p(\neg C_k \mid A_{ij})$ es $1 - p(C_k \mid A_{ij})$.

3.2 Características de *XKey*.

SELECCIÓN DEL CRITERIO DE RAMIFICACIÓN.

XKey opera con cuatro criterios de ramificación:

- Entropía [Quinlan, 1986]. Esta heurística, suele favorecer la construcción de árboles de decisión con un grado de ramificación elevado (favorece aquéllas preguntas que tienen más resultados posibles).
- Proporción de ganancia [Quinlan, 1993]. Pretende normalizar la ganancia obtenida para evitar la construcción de árboles de decisión que clasifiquen los casos utilizando sus claves.

- Índice de diversidad de Gini [Breiman *et al.*, 1984]. Esta medida trata de minimizar la impureza existente en los subconjuntos de casos de entrenamiento generados al ramificar el árbol de decisión.
- El mejor atributo desde el punto de vista de la teoría de la información no siempre es el mejor atributo desde el punto de vista taxonómico. Por este motivo *XKey* pretende integrar la visión del problema con el punto de vista de la Taxonomía. En esta área se han descrito pocos criterios para generar claves de identificación. En este sentido sólo podemos citar a el utilizado por Dallwitz en su herramienta *Key* [Dallwitz *et al.*, 2000]. *XKey* permite utilizar la regla de división propuesta en esta herramienta.

TRATAMIENTO DE VALORES NULOS.

Significado de los valores nulos en taxonomía.

Al generar una clave es común la aparición de valores nulos en los conjuntos de datos. El resultado final del proceso depende de la interpretación que el algoritmo hace de los mismos. En Taxonomía podemos dar 3 interpretaciones a la aparición de valores nulos:

1. El valor no aparece porque el atributo es inaplicable para un taxon determinado. Si durante la ramificación se utiliza un atributo con valores nulos, el taxon al que corresponde el valor nulo puede queda sin clasificar (ver Figura 3-18).
2. El valor no aparece porque es variable. Esto quiere decir que, para un taxon determinado el atributo puede tomar cualquier valor válido. Si utilizamos este carácter para ramificar el árbol de decisión, el taxon con el valor nulo se añade a cada una de las ramas del árbol.
3. El valor no aparece porque es desconocido. En tal caso y para evitar pérdida de información, se actúa igual que en el caso 2.

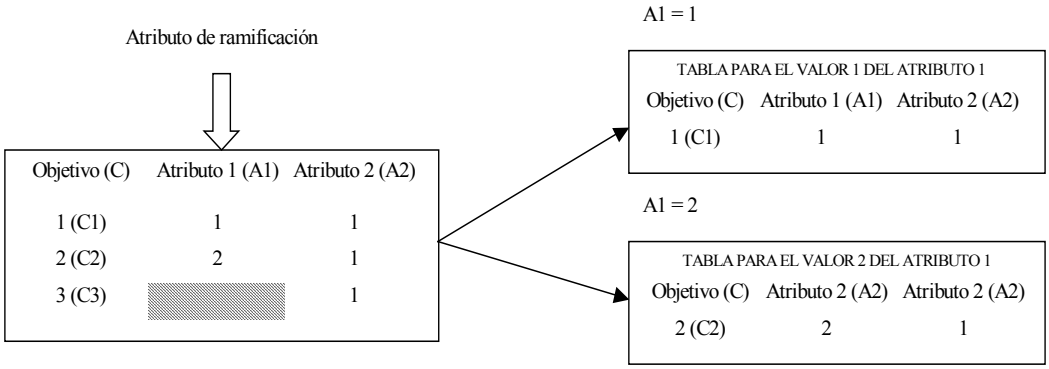


Figura 3-18. *Ejemplo de atributo con valor nulo. El objeto C3 queda sin clasificar.*

REPRESENTACIÓN DE VALORES NULOS EN *SDD*.

SDD permite dos formas de representación de valores nulos:

- Representación explícita (ver Figura 3-19). Con anterioridad hemos descrito que *SDD* admite dos estados especiales como valores de atributos para indicar de forma explícita la presencia de valores nulos y su significado. Se trata de los estados globales "*Unknown*" y "*NotApplicable*" (ver Página 101). Un atributo con el valor especial "*Unknown*" será tratado por *XKey* como desconocido (caso 3) y un atributo con el valor especial "*NotApplicable*" será tratado como inaplicable (caso 1) independientemente de la interpretación general con que se haya configurado *XKey*.

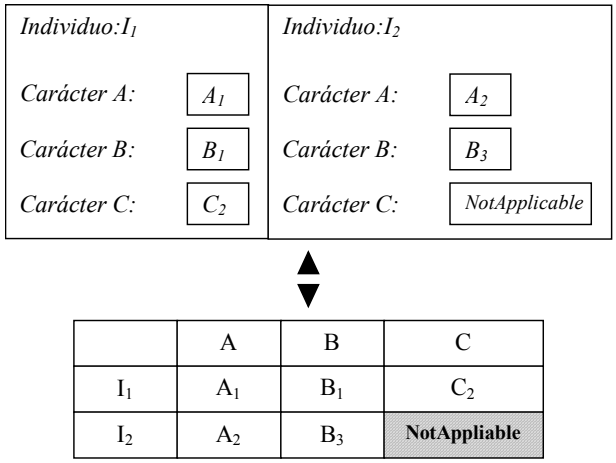
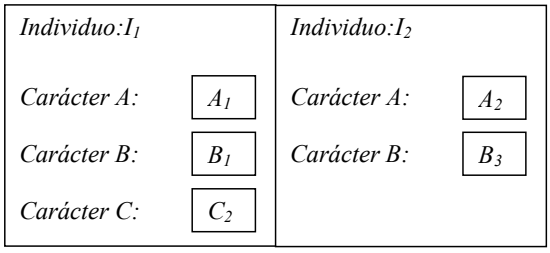


Figura 3-19. *Representación explícita de valores nulos en SDD y su correspondencia con una tabla de decisión.*

- Representación implícita (ver Figura 3-20). Además de la representación explícita, es posible que la descripción de un individuo no tenga ningún valor asignado para un determinado carácter. Esta situación equivale a un valor nulo en una tabla de decisión. *XKey* permite configurar la interpretación general de los valores nulos representados de forma implícita en un determinado conjunto de datos. Concretamente se puede elegir entre interpretarlos como valores desconocidos o como valores inaplicables.



	A	B	C
I_1	A_1	B_1	C_2
I_2	A_2	B_3	

Figura 3-20. *Representación implícita de valores nulos en SDD y su correspondencia con una tabla de decisión.*

UTILIDAD DE UN CARÁCTER.

En Taxonomía no todos los caracteres tienen la misma relevancia: unos son más importantes que otros, por ejemplo, por la facilidad con que se observan, por ser caracteres diferenciadores, etc. Este aspecto se denomina la utilidad (*reliability*) de un carácter [Dallwiz *et al.*, 2000]. Si el experto aporta información sobre qué carácter es más útil para ramificar, la clave resultado es mucho más adecuada.

Cuando *XKey* calcula el valor del criterio de división y aparecen varios atributos como candidatos para la ramificación, se seleccionará el que tenga mayor utilidad. Distinguimos los siguientes casos para determinar la utilidad de un carácter:

Caso 1.

Los conjuntos de datos de *KeyManager* incluyen un valor de prioridad para los atributos. Si este valor está disponible, se selecciona el atributo de mayor prioridad. En caso de no estar disponible, o de que todos los atributos tengan la misma utilidad, *XKey* actúa igual que en el caso 2.

Caso 2.

Los conjuntos de datos *SDD* no incluyen información sobre la utilidad de un carácter. Al no tener ningún valor asociado, en los casos en que dos atributos tienen el mismo valor del criterio de división, la construcción del árbol depende del orden en que los datos están representados en el conjunto de datos. Para evitar esta situación, *XKey* muestra al usuario las diferentes alternativas, el estado actual del árbol y los objetivos que quedan por clasificar para que sea este quien seleccione el atributo que considera más adecuado.

También es posible asignar en tiempo de ejecución un valor de prioridad a un determinado atributo. Este peso será recordado por el sistema durante toda la ejecución y se utilizará para decidir entre varios atributos de ramificación con el mismo valor del criterio de división (ver Figura 3-21).

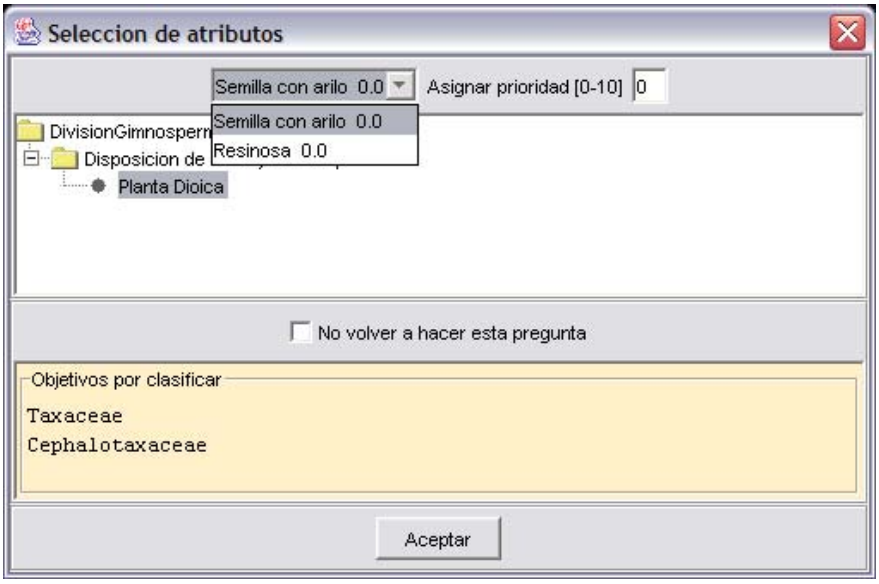


Figura 3-21. *Selección de atributos en tiempo de ejecución.*

En algunas ocasiones, el conjunto de datos puede resultar demasiado grande para aplicar esta estrategia, o bien al experto no le interesa realizar este tipo de elecciones. Por este motivo, el usuario puede indicar a *XKey* en cualquier momento que no quiere ser preguntado más sobre este aspecto.

CARACTERES DE CONFIRMACIÓN.

En una clave es habitual incluir más de un carácter en el mismo nodo del árbol de clasificación. Es decir, el nodo contiene un carácter principal y un conjunto de caracteres denominados *caracteres de confirmación* (*confirmatory characters*) [Dallwitz *et al.*, 2000] caracterizados por tener la misma distribución, dentro del grupo de *taxa* que se está siendo considerado, que el carácter principal seleccionado en un punto determinado de la clave. Esto implica que, dentro del grupo de taxa en consideración, los dos caracteres:

- Tienen el mismo valor del criterio de división.
- Tienen el mismo número de valores diferentes.
- Generan grupos de *taxa* idénticos al ser utilizados como carácter de ramificación.

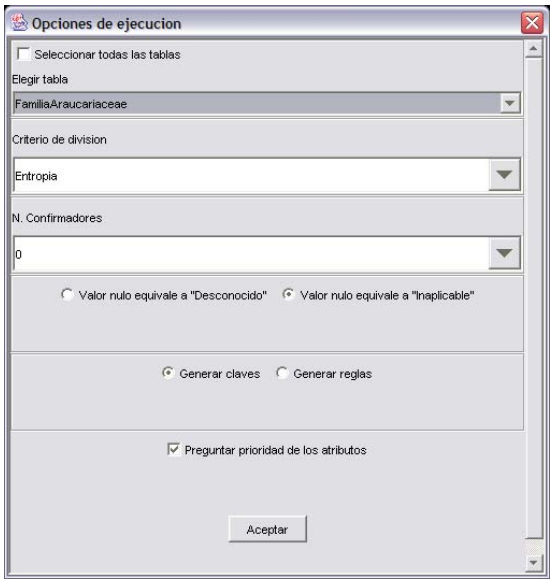


Figura 3-22. *Opciones de ejecución de XKey.*

El carácter principal y los caracteres de confirmación son equivalentes, se puede utilizar cualquiera de ellos para ramificar el árbol. *XKey* incluye una opción para que el usuario indique si desea incluir este tipo de caracteres y el número máximo permitido en cada nodo del árbol.

FORMATOS DE SALIDA.

Una vez generada la clave de identificación, se presenta al usuario en forma de árbol. Es posible salvar esta clave en los formatos que se describen a continuación:

- Formato texto. La clave se salva en un fichero de texto plano. Este formato facilita la modificación de la clave con otros editores de texto más sofisticados, y a la vez no limita la utilización de ninguno de estos.
- Formato XML. La clave se salva en un fichero en formato XML, muy próximo al que se prevé será el formato de claves que quede integrado en *SDD*. Este formato permite la edición de las claves una vez generadas y el intercambio sencillo de información.
- Formato CLIPS. La clave se salva en forma de reglas para esta conocida *shell* de sistemas expertos, con lo que logramos intercomunicar el formato *SDD* con una de las *shells* más populares dentro del campo de los sistemas expertos.
- Formato *GREEN*. Hemos comentado que además de generar claves, *XKey* puede generar conjuntos de reglas directamente utilizables por el sistema *GREEN*.

MODOS DE OPERACIÓN.

Una diferencia entre el aprendizaje de modelos de clasificación tradicionales y las claves de identificación es la interactividad en la selección de caracteres. El mejor carácter desde el punto de vista del criterio de división puede no ser el mejor carácter desde el punto de vista del experto. En ocasiones los expertos prefieren eliminar las excepciones en los primeros pasos de la clave mientras que un algoritmo clásico deja estos casos para los últimos pasos. Otras veces, el mejor carácter desde el punto de vista del criterio de división, puede ser

muy difícil de visualizar, etc. Por estos motivos, *XKey* ofrece varios modos de operación:

- Modo automático. Aplica el algoritmo para la generación de árboles de decisión sin contar con el usuario y utilizando como criterio de ramificación el que se haya preseleccionado en el menú de opciones.

- Modo semi-automático. Aplica el algoritmo para generación de árboles de decisión de forma automática, pero consulta al usuario en aquellos casos en que se producen empates en la selección del mejor atributo de ramificación.

- Modo interactivo. Aplica el algoritmo para generar claves por pasos. En cada paso el usuario selecciona el nodo que quiere ramificar y el atributo que será utilizado para dicha ramificación. *XKey* muestra la lista de todos los caracteres que se pueden utilizar para la ramificación, ordenados en función del criterio de división. De este modo aporta una información estadística de la que el experto no dispone cuando genera claves de forma manual.

Generación automática de claves.

El algoritmo para la generación automática de claves de *XKey* responde a las directrices generales de un algoritmo *TDIDT*. La aportación *XKey* es su capacidad de:

1. Añadir caracteres de confirmación.
2. Tratar los valores nulos según la interpretación determinada por el usuario.
3. Seleccionar, en tiempo de ejecución, el mejor atributo de ramificación (en el caso en que varios atributos tienen el mismo valor del criterio de división).

Generación interactiva de claves.

La generación interactiva de claves sigue el mismo esquema de la generación automática, añadiendo la facilidad de que en cada paso el usuario decida qué nodo ramificar y qué atributo utilizar para realizar la ramificación. Las operaciones permitidas durante la ejecución en modo interactivo son:

- Añadir nodo. El usuario selecciona el nodo a ramificar y el atributo de ramificación y genera el subárbol correspondiente a dicho nodo.

- Borrar nodo. Se puede eliminar un paso determinado en la clave. En este caso se elimina el nodo y todos sus descendientes.

- Finalizar. Esta opción permite al usuario fijar el lugar de aparición de determinados caracteres y pasar después el control al sistema para que termine de forma automática. Cuando se selecciona esta operación, *XKey* analiza el modelo de clasificación para detectar los nodos hoja que no contienen clases del ejemplo (nodos intermedios del modelo de clasificación) y terminar la construcción de esos nodos de forma automática.

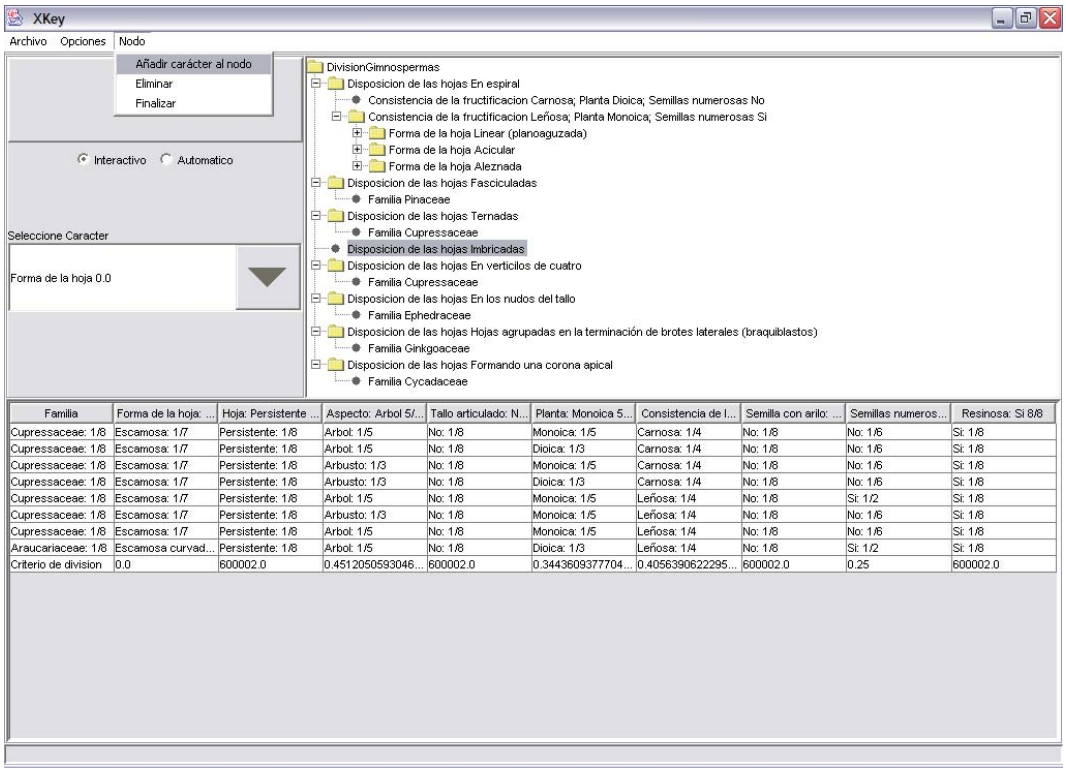


Figura 3-23. *Selección interactiva de atributos con XKey.*

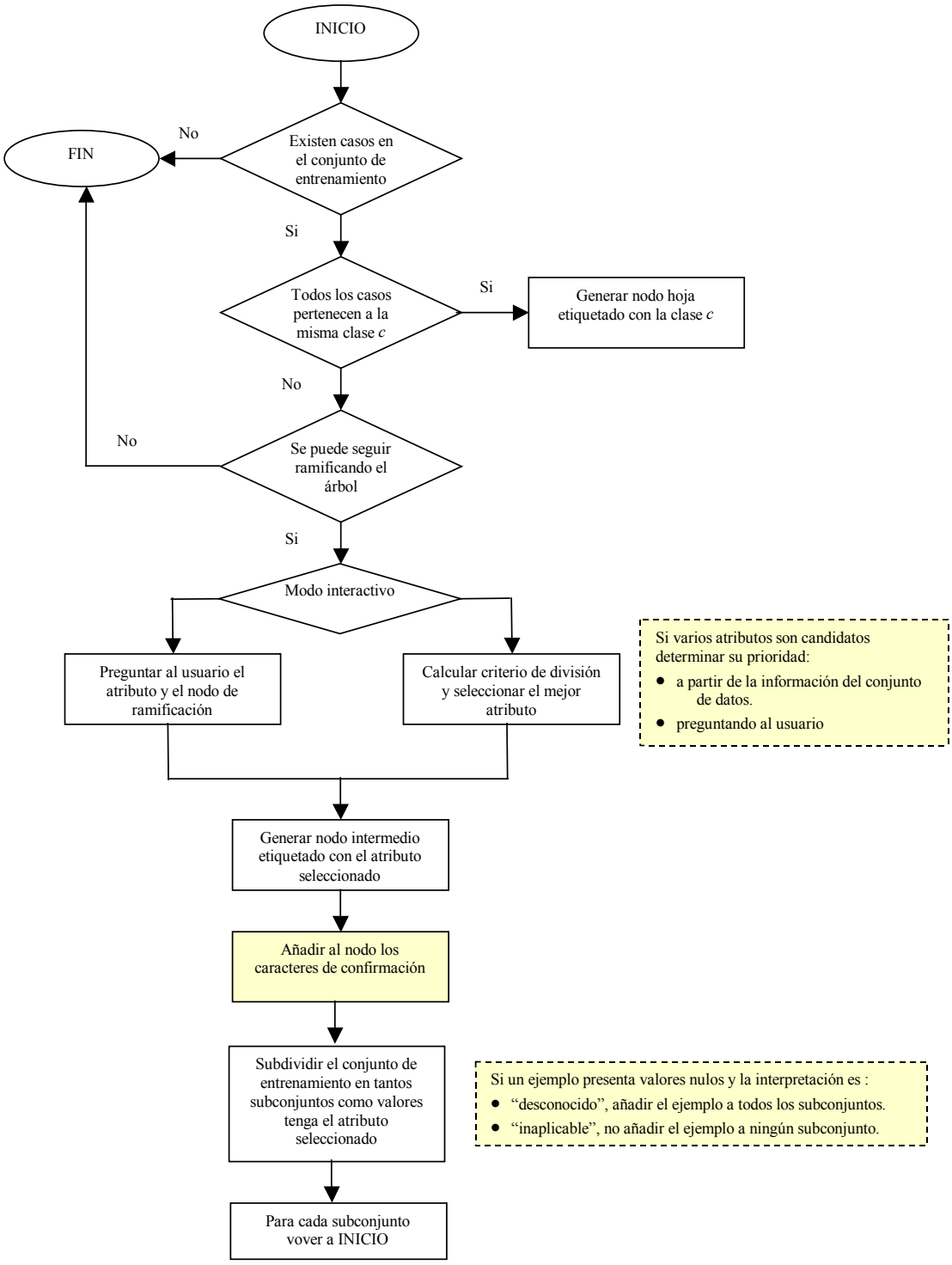


Figura 3-24. *Generación automática de claves de identificación.*

RESULTADOS DE EJECUCIÓN.

Además de la clave, *XKey* ofrece información cuantitativa que permite analizar los resultados obtenidos. Esta información incluye:

- Longitud media de la clave. Indica el número medio de preguntas que hay que responder en la clave para llegar a una identificación.
- Desviación típica de la longitud de la clave. Nos da una medida de lo equilibrado que está el árbol. Si la desviación típica es grande, los diferentes caminos del árbol tienen longitudes muy diferentes. Para poder comparar unas claves con otras, el sistema devuelve el coeficiente de variación de Pearson $\left(\frac{\sigma_x}{\bar{x}}\right)$.
- Longitud máxima y mínima de la clave: Indica el número máximo y mínimo de preguntas que hay que responder para llevar a cabo una identificación.
- Número de nodos hoja. Es una medida de la ramificación del árbol. Si hay muchos más nodos hoja que objetivos a clasificar, el árbol está muy ramificado.
- Número de nodos internos. Es una medida del tamaño del árbol.
- Número medio de opciones en los nodos hoja y la desviación típica. Indica, en el caso de nodos hoja con más de una opción, el número medio de opciones de cada uno de ellos.
- Número de nodos OR. Indica el número de nodos en que no es posible llegar a una identificación clara.
- Número de nodos exclusivos. Indica el número de nodos en los que llega a una identificación completa.
- Número de atributos totales en el conjunto de datos y número de atributos utilizados en la clave. Permite determinar qué proporción de los atributos del conjunto de datos se han utilizado para generar la clave y cuáles han sido estos atributos.
- Número de atributos de confirmación utilizados en la clave. Permite determinar cuantos atributos confirmadores diferentes se han podido incluir en la clave y cuáles han sido estos atributos.

3.3 Construcción del modelo de probabilidad a partir de la descripción orientada a objetos de *SDD*.

Las reglas de división utilizadas por *XKey* (ver Página 144) determinan el valor del criterio de división para un atributo en función de un conjunto de frecuencias:

- $p(A_{ij})$ es la probabilidad de que el atributo *i* tome su *j-ésimo* valor.
- $p(C_k \mid A_{ij})$ es una estimación de la probabilidad de que un ejemplo pertenezca a la clase C_k cuando su atributo A_i toma su *j-ésimo* valor. Esta estimación no es más que la frecuencia relativa, $f(C_k \mid A_{ij})$, en el conjunto de entrenamiento utilizado.

El modelo *SDD* no recoge de forma explícita estos valores de frecuencia, pero podemos obtenerlos a partir del propio conjunto de datos haciendo una abstracción previa sobre el modelo *SDD* que nos permita relacionar su estructura orientada a objetos con la estructura de una tabla de decisión.

ABSTRACCIÓN DEL MODELO *SDD*.

Un documento *SDD*, visto desde el punto de vista de la generación de claves, es una tupla $D = \{LC, LI\}$, donde:

- LC es la lista de los atributos válidos en el documento $LC = \{A_1, A_2, ..., A_m\}_{i=1}^{m}$. Cada atributo es una lista $A_i = LV = \{A_{i1}, A_{i2}, ..., A_{ip}\}_{j=1}^{p}$ que recoge valores permitidos para este.
- LI es la lista de descripciones de individuos $LI = \{I_1, I_2, ..., I_p\}_{k=1}^{q}$. Cada elemento, I_k, está formado por una lista, $I_k = LC' \subseteq LC$, con los caracteres que lo describen, formada a su vez por una lista, $LV' \subseteq LV \quad \forall I_k$, de atributos válidos para esos caracteres.

Si consideramos la variable estadística bidimensional (I, A_i), la distribución margial de A_i nos da las frecuencias marginales: $f(A_{ij}) = \frac{n_{.j}}{n}$.

I\A$_i$	**A$_{i1}$**	**A$_{i2}$**	**...**	**A$_{ij}$**	**...**	**A$_{ip}$**	
I$_1$	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}	$n_{1.}$
I$_2$	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}	$n_{2.}$
...	...	...	...	...	...	...	...
I$_k$	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kp}	$n_{k.}$
...	...	...	...	...	...	...	...
I$_q$	n_{q1}	n_{q2}		n_{qj}	...	n_{qp}	$n_{q.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.p}$	**n**

Distribución marginal de I

Distribución marginal de A$_i$

Tabla 3-3. *Variable estadística bidimensional (I, A_i).*

A partir de esta tabla obtenemos la distribución de *I* condicionada al valor *j* de A_i.que nos proporciona la estimación de la frecuencia relativa $f(I_1 \mid A_{ij}) = \frac{n_{kj}}{n.j}$

I\A$_{ij}$	**A$_{ij}$**
I$_1$	n_{1j}
I$_2$	n_{2j}
...	...
I$_k$	n_{kj}
...	...
I$_q$	n_{qj}
	$n_{.j} = \sum_{k=1}^{q} n_{kj}$

Tabla 3-4. *Distribución de I condicionada al valor j del atributo A_i.*

XKey obtiene, de forma incremental (a medida que procesa cada I_k), los valores n_{kj} para cada atributo A_i. Para ello, analiza el primer individuo I_1 y estima n_{1j}., el primer sumando que contribuirá a la frecuencia marginal $n_{.j}$.(ver Tabla **3-5**)

Para cada I_k del documento *D*:
  factor = 1
  Para cada carácter A_i de la descripción de I_k.
    Para cada estado A_{ij} de A_i
      Asignar a n_{kj} el valor *factor*
      *factor = factor*p*
      Actualizar frecuencias condicionadas de los atributos $A_1,...,A_{i-1}$: $n_{kj} = n_{kj} * factor$

Tabla 3-5. *Cálculo de las frecuencias relativas.*

Finalizado el cálculo de las frecuencias marginales, la Tabla 3-6 muestra el proceso para calcular la sumatoria para obtener *n.j*, que nos permitirá obtener $f(A_{ij})$.

Para cada carácter A_i , de la descripción de I_k $n_{.j}$=0 Para cada valor *j* de A_i hacer: $n_{.j} = n_{.j}+n_{kj}$

Tabla 3-6. *Cálculo de las frecuencias marginales.*

Cuando un atributo presenta más de un valor para un determinado carácter indica que la descripción representa más de un ejemplo (o fila) en una tabla de decisión que debe propagarse a los valores n_{kj} previamente calculados. El valor *factor*, controla el número de filas que hay que añadir cada vez.

EJEMPLO.

El siguiente ejemplo ilustra el algoritmo que acabamos de detallar. Partimos de las descripciones de dos individuos I_1, e I_2.

Individuo:I_1				*Individuo:*I_2		
Carácter A:	A_1			*Carácter A:*	A_2	
Carácter B:	B_1	B_2	B_3	*Carácter B:*	B_1	B_3
Carácter C:	C_1	C_2		*Carácter D:*	D_1	D_2

Figura 3-25. *Abstracción de la descripción SDD de dos individuos I_1, e I_2.*

El algoritmo procesa el primer individuo, I_1 (inicialmente *factor=1)*. Analiza el primer atributo de I_1 y asigna a n^{A}_{11} el valor de *factor*.

Clase	Carácter A
I_1	A_1
factor = 1	$n^{A}_{11} = 1$

Tabla 3-7. *Resultado parcial para I_1 tras añadir el atributo A.*

A continuación analiza el segundo atributo, *B*, y asigna a $\left(n_{1j}^{B}\right)_{j=1}^{3}$ el valor de *factor*. El carácter *B* tiene 3 valores diferentes para I_1, entonces actualiza *factor* y los valores n_{kj} del atributo *A* multiplicando por 3 (ver Tabla 3-8).

Clase	Carácter A	Carácter B
I_1	A_1	B_1
I_1	A_1	B_2
I_1	A_1	B_3
factor =3	$n_{11}^{A}=3$	$n_{11}^{B}=1$ $n_{12}^{B}=1$ $n_{13}^{B}=1$

Tabla 3-8. *Resultado parcial para I_1 tras añadir los atributos A y B.*

Para terminar con I_1, analiza el tercer atributo, *C*, haciendo $\left(n_{1j}^{C}\right)_{j=1}^{2}=3$ (el valor de *factor*). El carácter *C* tiene 2 valores diferentes para I_1, luego actualiza el valor de *factor* y de los n_{kj} previos doblando su magnitud (ver Tabla 3-9). A continuación procesa del mismo modo el individuo I_2. El resultado se muestra en la Tabla 3-10.

Clase	Carácter A	Carácter B	Carácter C
I_1	A_1	B_1	C_1
I_1	A_1	B_2	C_1
I_1	A_1	B_3	C_1
I_1	A_1	B_1	C_2
I_1	A_1	B_2	C_2
I_1	A_1	B_3	C_2
factor = 6	$n_{11}^{A}=6$	$n_{11}^{B}=2$ $n_{12}^{B}=2$ $n_{13}^{B}=2$	$n_{11}^{C}=3$ $n_{12}^{C}=3$

Tabla 3-9. *Resultado parcial para I_1 tras añadir los atributos A, B y C.*

Clase	Carácter A	Carácter B	Carácter C	Carácter D
I_1	A_1	B_1	C_1	
I_1	A_1	B_2	C_1	
I_1	A_1	B_3	C_1	
I_1	A_1	B_1	C_2	
I_1	A_1	B_2	C_2	
I_1	A_1	B_3	C_2	
I_2	A_2	B_1		D_1
I_2	A_2	B_3		D_1
I_2	A_2	B_1		D_2
I_2	A_2	B_3		D_2

Tabla 3-10. *Estructura tabular de los ejemplos de la* **Figura 3-25**.

Tras procesar I_2, suma las frecuencias relativas obtenidas con los valores obtenidos al analizar I_1.

	A		B			C		D	
I_1	$n^{A}_{11}=6$	$n^{A}_{12}=0$	$n^{B}_{11}=2$	$n^{B}_{12}=2$	$n^{B}_{13}=2$	$n^{C}_{11}=3$	$n^{C}_{12}=3$	$n^{D}_{11}=0$	$n^{D}_{12}=0$
I_2	$n^{A}_{21}=0$	$n^{A}_{22}=4$	$n^{B}_{21}=2$	$n^{B}_{22}=0$	$n^{B}_{23}=2$	$n^{C}_{21}=0$	$n^{C}_{22}=0$	$n^{D}_{21}=2$	$n^{D}_{22}=2$
	$n^{A}_{.1}=6$	$n^{A}_{.2}=4$	$n^{B}_{.1}=4$	$n^{B}_{.2}=2$	$n^{B}_{.3}=4$	$n^{C}_{.1}=3$	$n^{C}_{.2}=3$	$n^{D}_{.1}=2$	$n^{D}_{.2}=2$

Tabla 3-11. *Resultados parciales de probabilidad tras analizar I_2.*

Capítulo 4. Resultados experimentales.

Este capítulo, como su nombre indica, está dedicado al análisis de los resultados obtenidos de la experimentación con el conjunto de herramientas que hemos desarrollado y a determinar la calidad de los resultados obtenidos con nuestras herramientas.

Para la prueba de las herramientas, hemos utilizado los siguientes conjuntos de datos:

- *Grass Genera of the world* (géneros de pasto mundiales) [Watson *et al.*, 1986]. Este conjunto de datos está desarrollado en Delta y ha sido utilizado para comprobar el funcionamiento de *GREEN* y *XKey* con conjuntos de datos de tamaño y procedencia aleatorios.

- *Gimnospermas ibéricas*. Este conjunto de datos ha sido desarrollado en colaboración con el Herbario. Las Gimnospermas ibéricas son un grupo muy conocido por los expertos implicados en la evaluación del sistema, lo que nos permite realizar una evaluación de la calidad de los resultados. Se compone de los subconjuntos de datos de la Tabla 4-1.

División Gymnospermae	*Familia Araucariaceae*
Familia Cephalotaxaceae	*Familia Cupressaceae*
Familia Cycadaceae	*Familia Ephedraceae*
Familia Ginkgoaceae	*Familia Pinaceae*
Familia Taxaceae	*Familia Taxodiaceae*
Género Abies	*Género Calocedrus*
Género Cedrus	*Género Chamaecyparis*
Género Cryptomeria	*Género Cupressus*
Género Cycas	*Género Ephedra*
Género Juniperus	*Género Larix*
Género Picea	*Género Pinus*
Género Platycladus	*Género Pseudotsuga*
Género Sequoia	*Género Sequoiadendron*
Género Taxodium	*Género Tetraclinis*

Tabla 4-1. *Subgrupos del conjunto de datos Gimnospermas ibéricas.*

Recordemos que esta memoria ha introducido una *shell* para generación de sistemas expertos (*GREEN*) y una herramienta para generar claves de identificación y reglas (*XKey*). Cada una de estas herramientas requiere aplicar un conjunto de técnicas de prueba diferentes.

1. Evaluación del sistema *GREEN*.

Siguiendo la metodología habitual para la evaluación de sistemas expertos, hemos realizado la evaluación del sistema *GREEN* en dos fases: validación y verificación [Cabrero-Carnosa *et al.*, 2003; Mahaman *et al.*, 2002]. La verificación pretende localizar posibles errores en el sistema experto y garantizar que funciona como se espera; mientras que la validación pretende evaluar el grado en que el sistema se comporta realmente como un experto en la materia, es decir, si damos un conjunto de datos al sistema experto y a un experto humano ¿Proporcionan ambos los mismos resultados? [O'Keefe *et al.*, 1987]. En estas dos etapas de evaluación han intervenido tres expertos botánicos, dos de ellos ajenos al proyecto.

VERIFICACIÓN.

Como hemos definido, el objetivo de la verificación es establecer una correspondencia entre las especificaciones del sistema y las operaciones realizadas por este. Al referirnos a un sistema experto, esta verificación debe alcanzar tanto al motor de inferencia como a la base de conocimiento. La verificación conlleva dos pasos: la comprobación de la adecuación a las especificaciones y la búsqueda de errores semánticos y sintácticos en la base de conocimiento [Gonzalez & Dankel, 1993].

Comprobación de la adecuación a las especificaciones.

Este proceso consiste en la realización y comprobación de un informe que describe cómo da respuesta el diseño del sistema al siguiente conjunto de preguntas [Gonzalez & Dankel, 1993]:

- *¿Se ha implementado el método de representación de conocimiento elegido?* **Sí**.
  En la etapa de diseño hemos optado por el modelo de reglas porque proporcionan una estructuración del conocimiento análoga a las claves utilizadas por los expertos botánicos y comprensible por el usuario. Además, facilitan la representación jerárquica del conocimiento, lo que permite dar respuestas multinivel y hacer búsquedas con poca información.

- *¿Se ha utilizado la técnica de razonamiento adecuada?* **Sí**.
  Se ha seleccionado una buena técnica de razonamiento que combina el razonamiento hacia delante con el razonamiento hacia atrás. La técnica de razonamiento hacia delante es adecuada en aquellos casos en se parte de un conjunto de observaciones a partir del cual se infiere la conclusión. En otros casos, el usuario prefiere confirmar una determinada hipótesis, lo que hace más adecuado el razonamiento hacia atrás. *GREEN* permite alternar entre estos métodos de razonamiento durante una misma sesión de consulta. Otro aspecto importante es la capacidad del sistema de seguir varias líneas de inferencia al mismo tiempo, de esta forma infiere siempre el máximo conocimiento a partir de los datos de entrada disponibles.

- *¿Es modular su diseño e implementación?* **Sí**.

  El sistema ha sido diseñado e implementado de forma modular. De hecho, la separación de módulos y el desacoplamiento entre el conocimiento y el mecanismo de inferencia nos ha permitido incluir de forma sencilla la consulta de Gimnospermas accediendo por rango taxonómico (familia y género) y la consulta de otros grupos diferentes, por ejemplo, los géneros de pastos mundiales.

- *¿La interfaz del sistema corresponde a los requerimientos?* **Parcialmente**.

  La interfaz se genera de forma automática, con independencia del conjunto de datos, y presenta al usuario los caracteres que pueden conducir a la identificación. Debido al tratamiento de la incertidumbre, también se incluyen las opciones de introducir un nivel de certeza para las observaciones y seleccionar varios estados para un mismo carácter. Por otro lado, se establece un diálogo con el usuario en los casos en que no se puede llegar a una conclusión. Este diálogo consiste en preguntas sobre datos desconocidos por el sistema, que en caso de conocerse harían cierto el antecedente de una regla. El usuario es quien fija el umbral para el número de atributos desconocidos por regla, que determina si realiza o no la pregunta sobre un determinado carácter. Con todo esto, hemos dado al sistema las principales capacidades de interacción. Para mejorar la interactividad sería interesante añadir entradas de glosario, ordenar los caracteres de acuerdo a su capacidad de separación ante una determinada entrada de datos y permitir la eliminación de la lista de caracteres, de aquellos que no son válidos para la identificación.

- *¿El módulo de explicación es apropiado para los usuarios finales?* **Sí**.

  El módulo justificador presenta al usuario un esquema de las reglas disparadas para alcanzar cada conclusión. La representación del conocimiento en forma de reglas facilita al usuario la comprensión del razonamiento seguido por el sistema. Además, los resultados se ordenan en función de su certeza, indicando qué opción resultado tiene mayor correspondencia con la observación que se ha realizado.

- *¿Se cumplen los requerimientos de tiempo real del sistema?* **Parcialmente**.
  El lenguaje de programación utilizado (*C*), es un lenguaje compilado que proporciona rapidez de ejecución y la reducción del tiempo de respuesta. También hemos reducido el trasiego de información preguntando al usuario cada vez el máximo de información de que dispone y hemos optimizado la programación del motor de inferencia. Todo esto conduce a respuestas eficientes. Sin embargo, no podemos hacer una afirmación rotunda con respecto a los requerimientos de tiempo real, pues al no tratarse de un aspecto crítico, no se ha realizado una verificación formal de este requisito.

- *¿El grado de mantenimiento del sistema es el deseado?* **Sí**.
  El diseño modular del sistema contribuye a facilitar su mantenimiento. La independencia entre la base de conocimiento y el motor de inferencia permite incorporar con facilidad nuevos conjuntos de datos y reparar inconsistencias en el conocimiento sin necesidad del alterar ni el motor de inferencia ni el contenido de otros conjuntos de datos. Esta separación también facilita el mantenimiento y la actualización del motor de inferencia.

- *¿El sistema cumple las especificaciones de seguridad?* **Sí**.
  Las especificaciones de seguridad se desarrollan en la configuración del servidor, que oculta la base de conocimiento y el módulo de inferencia e impide el acceso a usuarios no autorizados al directorio donde reside.

Consistencia y completitud.

En la etapa de adquisición del conocimiento es habitual que se introduzcan, de forma inconsciente, errores semánticos y sintácticos en la base de conocimiento. Estos errores afectan a la completitud y consistencia y deben ser detectados y eliminados durante la verificación. Una base de conocimiento completa y consistente no es condición suficiente para que el sistema proporcione las respuestas correctas, pero garantiza que se ha diseñado e implementado correctamente.

La búsqueda de errores sintácticos en la base de conocimiento requiere de la comprobación de reglas redundantes, conflictivas, subsumidas, circulares, reglas con condiciones IF innecesarias, reglas sin salida, reglas perdidas y reglas inalcanzables [González & Dankel, 1993]. Dadas las particularidades de la obtención automática de reglas, algunos de estos casos no se presentan en nuestro sistema, por lo que esta comprobación se reduce a reglas redundantes, conflictivas, subsumidas y condiciones IF innecesarias (ver Página 124).

VALIDACIÓN.

La validación, el último control de calidad de un sistema experto, es un proceso más complicado que la verificación. Pretende garantizar que la salida del sistema es correcta y evaluar en qué grado el sistema se comporta realmente como un experto. Para esto, hemos aplicado diversas técnicas de validación de sistemas basados en el conocimiento, descritas en [González & Dankel, 1993] que detallamos a continuación.

Validación informal.

Hemos realizado una *validación informal* para refinar el funcionamiento del sistema durante la fase de desarrollo. La opinión del experto botánico implicado en el proyecto sobre la validez de las conclusiones del sistema permitió detectar algunas deficiencias para las que planteamos un conjunto de mejoras iniciales.

Fundamentalmente, encontramos algunas incorrecciones en el conjunto de datos de Gimnospermas ibéricas, errores semánticos que no pudieron ser encontrados con técnicas automáticas al comprobar la consistencia y la completitud. Su estudio revela que se debió a:

- La consulta de varias fuentes bibliográficas. Diferentes autores, incluso un mismo autor en obras diferentes, pueden utilizar de forma distinta la terminología botánica. Este aspecto pone de manifiesto la dificultad de

adquirir conocimiento taxonómico y la importancia de realizar una homogeneización previa de la terminología.

- La inclusión de especies cultivadas en el conjunto de datos de Gimnospermas. Estos casos especiales presentan con frecuencia caracteres que no se observan en otros grupos, lo que obliga a complicar la terminología y es una fuente de errores.

- La inclusión de reglas adicionales cuando aparecen caracteres diferenciadores (ver 120). El algoritmo de clasificación añade de forma automática estas reglas, pero no todas las reglas que cumplen las condiciones estructurales necesarias son reglas con caracteres diferenciadores. Esta decisión sólo puede tomarla un experto; la solución es proponer un conjunto de reglas candidatas sobre las que el experto tomará una decisión.

Además de esto, el experto implicado en el proyecto propuso una serie de mejoras de la interfaz orientadas a facilitar la introducción de datos de entrada. Sus propuestas incluyeron:

- El diálogo con el usuario en los casos en que el sistema no puede alcanzar una solución con los datos de entrada introducidos. La finalidad de este diálogo es obtener información extra que podría conducir a la identificación.
- La modularización de la base de conocimiento. De esta forma, los caracteres se presentan ordenados por niveles, lo que evita introducir información no necesaria y reduce el tiempo de búsqueda.
- La Incorporación de un modo de consulta para la prueba de hipótesis. En ocasiones el usuario desea comprobar una determinada hipótesis, este modo de consulta va de los objetivos a los datos y requiere de un proceso de encadenamiento hacia atrás (ver Página 127).

La validación informal es de gran utilidad durante el desarrollo del sistema, pero no es suficiente para la validación completa del mismo. Por esto, hemos llevado a cabo una segunda etapa, posterior al desarrollo, en la que hemos realizado una *validación por test de prueba* y un *test de sensibilidad*.

Validación por test de prueba.

El resultado del *test de prueba* es la estimación numérica de algún criterio de evaluación. La *adecuación* y la *aceptabilidad* son dos criterios muy utilizados.

- La *adecuación*[14] mide la cantidad del dominio del problema cubierta por el sistema. Por ejemplo, un sistema de clasificación de insectos que clasifica correctamente 120 de un total de 145 especies tiene una adecuación del 83%. También se puede establecer esta medida ponderando las respuestas, de esta forma tendrán más peso las más críticas. Nuestro modelo clasificador siempre cubre todo el dominio del problema (*XKey* siempre genera al menos una regla por objetivo), por lo que la adecuación del sistema es del 100%.

- La *aceptabilidad* es la proporción de respuestas que concuerdan con las que propondría un experto al plantearle el mismo problema [Marcot, 1987]. El sistema es jerárquico y proporciona respuestas en tres niveles (familia, género y especie), por esto la aceptabilidad se mide en función del rango taxonómico hasta el que es capaz de afinar una solución: al menos de debe dar respuesta al nivel de familia y género, y es menos importante llegar hasta la especie. Durante el desarrollo del sistema experto es necesario especificar previamente un umbral mínimo de aceptabilidad [O'Keefe *et al.*, 1987]; es habitual situarlo alrededor del 80% [Batchelor *et al.*, 1989].

La validación por *test de prueba* requiere de un conjunto de casos de prueba. Los casos de prueba de *GREEN* han consistido en una selección de pliegos de herbario revisados por los autores (expertos) de las Gimnospermas para la edición de la *Flora Ibérica* [López & Do Amaral, 1986], obra de referencia para la botánica española. Dichos pliegos fueron sometidos a un proceso de identificación por expertos botánicos que utilizaron como herramienta para ello el sistema *GREEN*. De esta forma hemos comparado los resultados alcanzados por

[14] También de denomina cobertura, en inglés *coverage* [Marcot, 1987].

GREEN con los proporcionados por los expertos utilizando un sistema de identificación tradicional (la clave impresa).

Análisis de sensibilidad

El *análisis de sensibilidad* pretende evaluar el impacto en los resultados producido por las variaciones en las entradas. Es decir, al introducir un pequeño cambio en las entradas ¿son muy grandes las variaciones en las salidas del sistema? Para obtener pequeñas variaciones en las entradas hemos utilizado distintos pliegos pertenecientes a un mismo taxon. Los pliegos correspondientes al mismo taxon, aun cuando mantienen un conjunto de características comunes, son ligeramente diferentes debido a la edad del individuo, el estado de conservación, la época de recolección, etc. Este análisis se ha realizado en paralelo con el *test* de prueba y sus resultados se describen en la discusión de la Página 177.

Resultados experimentales de la validación.

Los experimentos que hemos realizado están descritos de forma detallada en el apéndice de la Página 273. La Tabla 4-2 muestra la precisión de los resultados obtenidos por *GREEN*.

A partir de los resultados de los experimentos, hemos obtenido dos valores de aceptabilidad:

- El primero de estos valores sólo considera aciertos los casos en que el sistema determina la especie correctamente.
- El segundo es un valor de aceptabilidad ponderada que tiene en cuenta el rango taxonómico hasta el que el sistema es capaz de precisar su respuesta. Así, una respuesta correcta sólo al nivel de género tiene la mitad del valor de una respuesta correcta al nivel de especie. Del mismo modo, una respuesta correcta sólo al nivel de familia representa un cuarto del valor de una respuesta correcta a nivel específico.

Los valores de aceptabilidad y aceptabilidad ponderada obtenidos son, respectivamente, 83.33% y 89.16%. Estos resultados superan el umbral del 80% fijado previamente (ver Página 168).

Resumen de los resultados			
EXPERIMENTO	**PRECISIÓN DEL RESULTADO**	**EXPERIMENTO**	**PRECISIÓN DEL RESULTADO**
1	Especie	16	Especie
2	Especie	17	Género
3	Especie	18	Especie
4	Especie	19	Especie
5	Especie	20	Especie
6	Especie	21	Familia
7	Especie	22	Especie
8	Especie	23	Especie
9	No resuelto	24	Especie
10	Especie	25	Especie
11	Género	26	Especie
12	Especie	27	Especie
13	Género	28	Especie
14	Especie	29	Especie
15	Especie	30	Especie

Total de casos	**Nivel especie**	**Nivel género**	**Nivel Familia**	**Casos no resueltos**
30	25	3	1	1

Aceptabilidad	83.33%
Aceptabilidad ponderada	89.16%

Tabla 4-2. *Resultados experimentales sobre la aceptabilidad de GREEN.*

Intervalo de confianza para la aceptabilidad.

La aceptabilidad mide el número de veces, *K*, que el sistema da una respuesta correcta al probarlo con *N* ejemplos. Si consideramos la variable aleatoria *número de éxitos observados en los N ensayos*, esta variable es un experimento binomial, pues posee las siguientes propiedades [Spiegel, 2000]:

1. Consta de un número determinado, *N*, de ensayos idénticos.
2. Cada ensayo tiene dos resultados posibles.
3. La probabilidad de tener éxito en un ensayo es igual a algún valor p, y permanece constante de un ensayo a otro. La probabilidad de fracaso es

$q = 1 - p$. Al ser p y q desconocidos, para valores grandes de N ($N>=30$), podemos tomar la estimación $p \sim P = K/N$[15].

4. Los ensayos son independientes.
5. La variable aleatoria bajo estudio es el número de éxitos observados en N ensayos.

A partir de los datos de la Tabla 4-2 pretendemos determinar un intervalo de confianza que contenga el valor K/N con una probabilidad alta, por ejemplo del 95%. Para calcular este intervalo, con una confianza del 95% fijamos un valor $\alpha = 0.05$ $(1 - \alpha = 0.95)$.

Si el tamaño de N es suficientemente grande ($N \geq 30$), la distribución de los valores K/N es aproximadamente normal [Spiegel, 2000]. Cuando esto sucede, el error estándar de la proporción es:

$$S_P = \sqrt{\left(\frac{(K/N) * (1 - K/N)}{N} \right)}$$

Fórmula 4-1. *Error estándar de la proporción K/N.*

Donde K/N es la probabilidad de éxito y $1-K/N$ es la probabilidad de fracaso. Si el estadístico P es la proporción de éxitos en una muestra de tamaño N sacada de una población binomial en la que p es la proporción de éxitos (o sea, la probabilidad e éxito), los límites de confianza para p vienen dados por $P \pm z_{\alpha/2} S_P$.

$$\left[K/N - z_{\alpha/2} * \sqrt{\frac{K/N(1 - K/N)}{N}} \quad , \quad K/N + z_{\alpha/2} * \sqrt{\frac{K/N(1 - K/N)}{N}} \right]$$

Fórmula 4-2. *Intervalo de confianza para la proporción K/N cuando N≥30.*

[15] El Teorema Central del límite establece que si el tamaño de la muestra es grande ($N \geq 30$), las distribuciones de muestreo son normales o casi normales. Cuando los parámetros de la población (σ, p, μ, etc.) son desconocidos y las muestras son grandes, pueden ser estimados con sus estadísticos muestrales.

Desarrollando la expresión obtenemos el intervalo de confianza de la Fórmula 4-2, donde $z_{\alpha/2}$ es el $(1-\alpha/2)-cuantil$ de la N(0,1).

Esto nos da los intervalos $[0.699 \quad , \quad 0.966]$ en el caso de la aceptabilidad sin ponderar y $[0.780 \quad , \quad 1]$ para la aceptabilidad ponderada (con $\alpha = 0.05$). Este último intervalo mejora los límites de confianza obtenidos en el primer caso, y nos indica que la probabilidad de que el intervalo [78%, 100%] de aciertos contenga a la proporción de aciertos es del 95%.

Para muestras pequeñas (*N<30*), esta aproximación no es buena y empeora al decrecer *N*, de modo que son precisas ciertas modificaciones. En este caso, los intervalos de confianza vienen dados por la Fórmula 4-3.

$$\left[K/N - t_{n-1,\alpha/2} * \sqrt{\frac{K/N(1-K/N)}{N}} \quad , \quad K/N + t_{n-1,\alpha/2} * \sqrt{\frac{K/N(1-K/N)}{N}} \right]$$

Fórmula 4-3. *Intervalo de confianza para la proporción K/N cuando N<30.*

Para pequeñas muestras sustituimos los valores $z_{\varepsilon/2}$ (obtenidos de la distribución normal) por $t_{n-1,\alpha/2}$ (obtenidos de la distribución de Student) de forma que cuando *N* crece, ambos métodos tienden a coincidir. Por ejemplo, en este caso, los límites de confianza obtenidos son $[0.694 \quad , \quad 0.972]$ para la aceptabilidad sin ponderar. Estos límites de confianza son mayores que los obtenidos por métodos de grandes muestras. Era de esperar, porque la precisión disponible con pequeñas muestras es menor que con muestras grandes.

Contraste de hipótesis para la aceptabilidad.

Queremos determinar si las respuestas proporcionadas por el sistema son o no aleatorias. Para ello, realizamos un contraste (o test) de hipótesis. Esta prueba estadística se basa en hacer una suposición o hipótesis de trabajo (hipótesis nula, H_0), si suponemos que dicha hipótesis es cierta pero vemos que los resultados hallados en una muestra aleatoria difieren notablemente de los esperados bajo tal

hipótesis, entonces diremos que las diferencias observadas son significativas y nos veremos inclinados a rechazar la hipótesis y aceptar una hipótesis alternativa (H_1). El estimador o estadístico de prueba, *z*, es una función de las mediciones de la muestra que sirve de fundamento para la toma de esta decisión. La región de rechazo, especifica los valores del estadístico de prueba para los que la hipótesis nula se rechaza a favor de la hipótesis alternativa.

Si *p* es la probabilidad de que el sistema devuelva una respuesta correcta, en el problema que nos ocupa hemos de decidir entre estas dos hipótesis:

H_0: p=0.8, los aciertos del sistema se deben al azar.

H_1: p>0.8, los aciertos del sistema son reales.

Como estamos interesados en el caso en que el sistema consiga muchos aciertos, escogeremos un contraste de una cola. Para un contraste unilateral con un 95% de confianza (al nivel de significación 0.05), debemos tomar el valor para el estadístico de modo que el área en la región crítica sea 0.05. En este caso, el área entre 0 y *z* es 0.450 y z=1.645 (ver Figura 4-1).

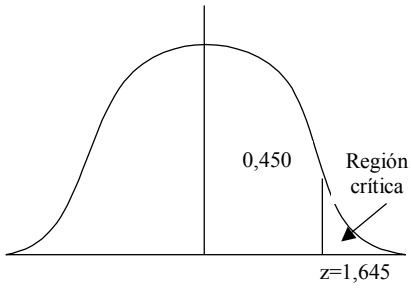


Figura 4-1. *Región de rechazo para un contraste de hipótesis de una cola.*

Luego, nuestra regla de decisión es:

- Si el *z* observado es mayor que 1.645, el resultado es significativo al nivel 0.05 y las respuestas del sistema no son aleatorias.
- En caso contrario, el resultado se debe al azar (no es significativo al nivel 0.05).

Al hacer un contraste sobre la proporción de éxitos en una muestra, el valor z viene dado por la Fórmula 4-4.

$$z = \frac{P - p}{\sqrt{pq/N}}$$

Fórmula 4-4. *Estimador de prueba para la proporción de éxitos en la muestra.*

En el caso $P=K/N$, donde K es el número real de éxitos de la muestra viene dado por la expresión de la Fórmula 4-5.

$$z = \frac{X - N_p}{\sqrt{Np(1-p)}}$$

Fórmula 4-5. *Estimador de prueba para la proporción de éxitos en la muestra cuando P=K/N.*

El resultado es significativo al nivel de significación 0.05 cuando la aceptabilidad es del 77.8% y la aceptabilidad ponderada del 83%.

Intervalo de confianza para la distancia entre los resultados del sistema y los resultados esperados.

El sistema experto ofrece soluciones en varios niveles (familia, género y especie), podemos estimar un intervalo de confianza para la distancia media entre las soluciones del sistema y la solución esperada al nivel de especie.

Si tomamos el espacio muestral del conjunto de las posibles combinaciones de respuestas del experto y del sistema, podemos definir una variable aleatoria X = *distancia entre las respuestas del sistema y del experto*, que permite diferenciar los distintos niveles de respuesta del sistema:

$$X = \begin{cases} 0 \;\; si \;\; la \;\; especie \;\; es \;\; correcta \\ 1 \;\; si \;\; el \;\; género \;\; es \;\; correcto \\ 2 \;\; si \;\; la \;\; familia \;\; es \;\; correcta \\ 4 \;\; en \;\; otro \;\; caso \end{cases}$$

Fórmula 4-6. *Distancia de las soluciones del sistema con respecto a las soluciones esperadas.*

Para esta variable aleatoria definimos los estadísticos *media muestral* ($\overline{X}$), *varianza muestral* (S^2) y *desviación típica muestral* (S).

$$\overline{X} = \frac{1}{N}\sum_{i}^{N} X_i$$

Fórmula 4-7. *Media muestral.*

$$S^2 = \frac{1}{N-1}\sum_{i}^{N}\left(X_i - \overline{X}\right)^2$$

Fórmula 4-8. *Varianza muestral.*

$$S = +\sqrt{S^2}$$

Fórmula 4-9. *Desviación típica muestral.*

A partir de la Tabla 4-2, obtenemos los valores de la Tabla 4-3 para los estadísticos que acabamos de definir y para la distancia entre las respuestas del sistema y las respuestas esperadas.

EXPERIMENTO	DISTANCIA	EXPERIMENTO	DISTANCIA
1	0	16	0
2	0	17	1
3	0	18	0
4	0	19	0
5	0	20	0
6	0	21	2
7	0	22	0
8	0	23	0
9	4	24	0
10	0	25	0
11	1	26	0
12	0	27	0
13	1	28	0
14	0	29	0
15	0	30	0
Media muestral		0.3	
Desviación típica muestral		0.83	

Tabla 4-3. *Distancia de las respuestas del sistema experto con respecto a la respuesta esperada.*

Los límites de confianza para estimar la media, μ, de una población vienen dados por $\overline{X} \pm z_{\alpha/2}\, S/\sqrt{N}$. Este resultado es válido para N≥30. Igual que en el caso del intervalo de confianza para la proporción de aciertos, para N<30, la aproximación es pobre y debe emplearse la teoría de pequeñas muestras, en cuyo caso el intervalo viene determinado por $\overline{X} \pm t_{n-1,\alpha/2}\, S/\sqrt{N}$.

El intervalo de confianza obtenido es $[0.0006, 0.5993]$. Esto indica que encontraremos la distancia media entre las respuestas del sistema y las del experto en el intervalo (entorno a 0.5) con una probabilidad del 95%.

Contraste de hipótesis para la distancia media de las respuestas del sistema.

A partir de la información de la muestra, queremos determinar el valor de la distancia media para la población con un nivel de confianza del 95% (o un nivel de significatividad de 0.05). Nuestras hipótesis de contraste son las siguientes:

H_0: $\mu = 0.5$

H_1: $\mu \neq 0.5$

Puesto que $\mu \neq 0.5$ incluye valores mayores y menores que la media, usaremos un contraste de dos colas. De este modo, la regla de decisión es la siguiente:

- Rechazar la hipótesis nula si el z de la media muestral está fuera del rango $[-1.96, 1.96]$.
- Aceptar la hipótesis nula en caso contrario.

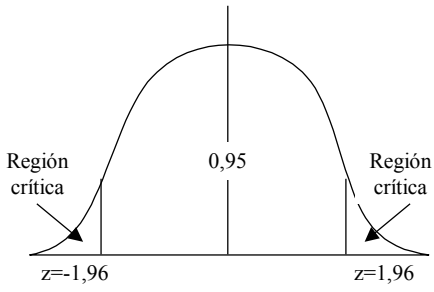


Figura 4-2. *Regiones de rechazo para un contraste de hipótesis de dos colas.*

El estadístico bajo consideración es la media muestral. El valor de *z* viene dado por la expresión de la Fórmula 4-10.

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{N}}$$

Fórmula 4-10. *Estimador de prueba para la distancia media de las respuestas del sistema.*

Usando la desviación típica muestral como estimación de σ obtenemos el valor 1.93 para *z*. Este valor está dentro del rango $[-1.96, 1.96]$, luego aceptamos H_0 al nivel de significación 0.05.

Análisis cualitativo de los experimentos realizados y los resultados obtenidos.

Los niveles de aceptabilidad y los intervalos de confianza obtenidos, son aceptables. Analicemos ahora el significado cualitativo y biológico de algunos ejemplos especialmente significativos.

En el experimento 9, el sistema debía haber concluido la especie *Abies pinsapo*, pero debido a la falta de información inicial no es capaz de llegar a ninguna conclusión, ni siquiera al nivel de familia. En estos casos, el sistema pregunta al usuario caracteres que podrían ser de utilidad. Esta característica ha permitido incluir nuevos datos (concretamente que se trata de una planta monoica) y se ha podido llegar a la determinación correcta.

En el experimento 11, *Pinus nigra*, el sistema da una respuesta satisfactoria para el género y la familia, pero no llega a la identificación correcta de la especie al concluir *Pinus pinea*. Esto se ha debido a diferencias en la interpretación del carácter "características de la apófisis". El experto que elaboró el conjunto de datos consideró que era *poco prominente*, mientras que el observador que ha realizado las pruebas ha considerado que era *convexa*.

En el experimento 18, *Ephedra fragilis*, el sistema devuelve dos resultados: *Ephedra fragilis* por el grosor de las ramillas y las ramas fácilmente

desarticulables y *Ephedra nebrodensis* por el color de las ramillas. También es un buen resultado, porque incluye entre sus respuestas la identificación correcta.

En los experimentos 19 y 20, *Larix decidua* y *Araucaria heterophylla*, el observador tenía dudas y ha seleccionado varios estados para un mismo carácter. En este caso, el sistema lanza varias líneas de inferencia, recuperando información por varios caminos y devolviendo varios resultados. El usuario es quien toma la decisión final a partir de la información proporcionada en la justificación, los niveles de certeza de los resultados y las descripciones diagnósticas.

En el experimento 13, *Pinus canariensis*, no es posible alcanzar una solución aceptable para la especie. Esto se debe a la falta información sobre la piña en el pliego. Lo mismo sucede en el experimento 21, *Sequoiadendron giganteum*, en el que solo llegamos hasta familia *Taxodiaceae*. En este último caso, necesitábamos información sobre la corteza y la piña para seguir avanzando, pero el pliego sólo daba información sobre la hoja. El no disponer de suficiente información complica el proceso de determinación. Esto también es común que suceda a los expertos, cuando se encuentran ante muestras incompletas, por ejemplo, la información de un pliego puede estar muy limitada. En el caso de pruebas con información en vivo esta casuística se simplifica ya que se dispone de todo el individuo.

Hemos presentado al sistema un caso de prueba con una especie que no se encontraba recogida en la base de conocimiento, la *Ephedra scoparia* (experimento 17). Es obvio que no se puede llegar a la determinación de la especie, pero el sistema concluyó adecuadamente la familia y el género. Se trata de un buen resultado, porque aún sin estar sin estar registrada la especie, la respuesta se ha aproximado mucho a la realidad. Debido a los datos sobre la "consistencia de la fructificación" y la "caducidad de la hoja", también ha devuelto como resultados las familias *Taxodiaceae* y *Pinaceae*. Pero dado que estas características pueden darse en ambas familias, la certeza de los resultados es menor.

En el ejemplo 29, *Cupressus sempervirens*, también concluye la identificación al nivel de especie, pero con un nivel de certeza bajo. Esto es debido a que el tamaño de la piña no está recogido correctamente en la base de datos.

DISCUSIÓN Y POSIBLES MEJORAS.

La evaluación del sistema ha producido, en general, buenos resultados, lo que conduce a reafirmar que la elección de técnicas para el diseño de implementación de *GREEN* ha sido adecuada. Estas técnicas se acoplan fácilmente unas con otras y contribuyen a la robustez y seguridad del sistema.

La verificación informal ha facilitado la depuración del funcionamiento del sistema y del contenido del conjunto de ejemplos. Disponer de una herramienta como *XKey* para generar la base de conocimiento de forma automática facilita la tarea de depuración de los conjuntos de datos.

Superada la verificación informal, que ha servido para realizar la puesta a punto del sistema, la mayor parte de los casos en que el sistema no era capaz de devolver la identificación correcta se ha debido la falta de información en la muestra, caso ante el cual un experto tampoco hubiera podido dar una conclusión.

Durante la prueba hemos detectado que, dado el volumen de caracteres manejado y la terminología empleada, el usuario puede introducir información innecesaria para la determinación. Aprovechando la capacidad genérica de que se ha dotado a *GREEN*, hemos generado una base de conocimiento modular (ver Figura 4-3), y hemos incluido entradas al sistema para cada uno de los rangos taxonómicos de Gimnospermas. De este modo, es posible acceder a la consulta directamente desde un nivel taxonómico preseleccionado y reducir el número de datos a introducir.

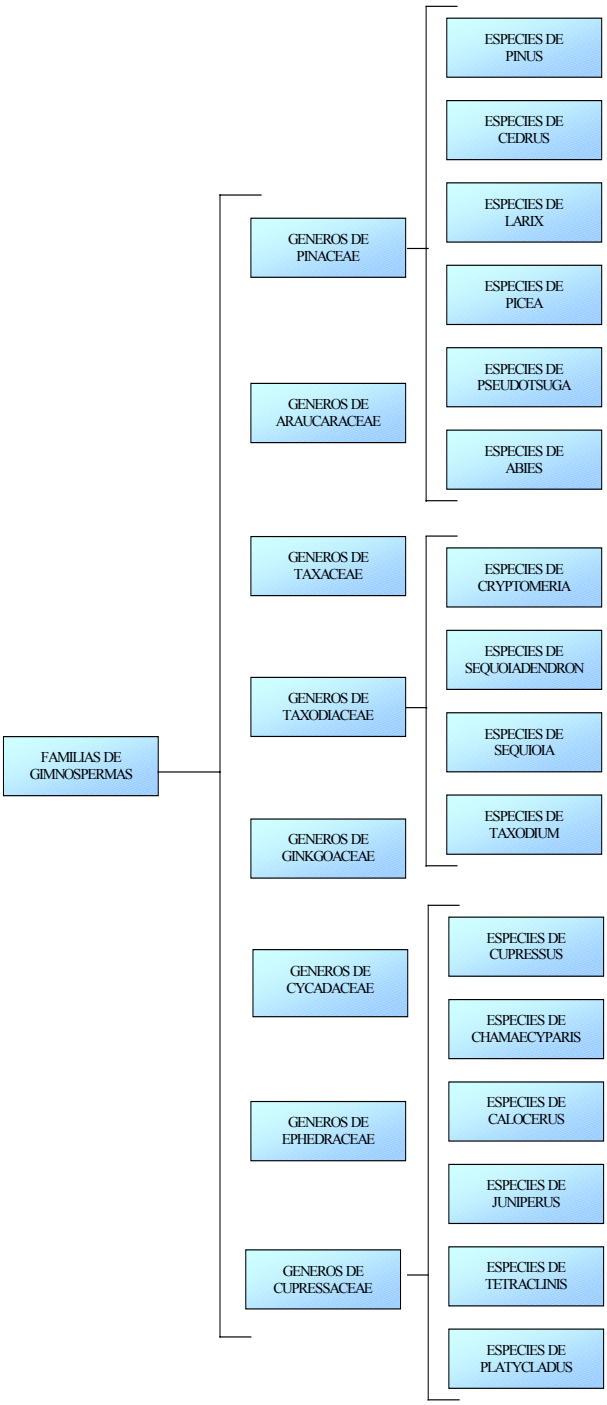


Figura 4-3. *Descomposición modular de reglas de Gimnospermas ibéricas.*

En algunas ocasiones, el usuario quiere comprobar una hipótesis sobre una muestra (por ejemplo, ¿se trata de un *Pinus pinea*?), para agilizar este tipo de consultas y optimizar el número de preguntas a responder, hemos incorporado la posibilidad de razonamiento hacia atrás.

En lo que a la calidad de los resultados se refiere, las pruebas estadísticas realizadas, revelan que se encuentran dentro de parámetros aceptables y tienen unos niveles de confianza razonables. La investigación se ha centrado en la aplicación de técnicas de Inteligencia Artificial al problema de la identificación taxonómica, estos datos podrían verse incrementados mediante la mejora de la interfaz para realizar una búsqueda más guiada que evite problemas como la interpretación errónea de los descriptores.

2. Prueba de la herramienta *XKey*.

DATOS DE EJECUCIÓN DE LA HERRAMIENTA *XKEY*.

La ejecución de *XKey* devuelve un conjunto de medidas con las que podemos comparar las claves obtenidas para un mismo conjunto de datos al utilizar, por ejemplo, diferentes criterios de división. A continuación presentamos y analizamos los resultados experimentales obtenidos con el conjunto de datos de Gimnospermas. La evaluación de estas claves generadas con *XKey* para este grupo nos ofrece un conjunto de pruebas suficientemente amplio para determinar qué criterios, en qué casos o condiciones y por qué permiten generar la clave más idónea. Las tablas Tabla 4-4 a Tabla 4-14 muestran el resumen de los resultados de ejecución.

Además de la interpretación cuantitativa de los resultados, otro aspecto relacionado con el dominio de aplicación con el que estamos trabajando, es el significado biológico de las claves obtenidas. Cada criterio de división es una heurística que pretende favorecer la construcción de un árbol con unas características determinadas, pero, es posible que la clave más eficiente no sea la más adecuada.

División Gymnospermae.

DIVISIÓN GYMNOSPERMAE					
NÚMERO DE OBJETIVOS	9				
NÚMERO DE ATRIBUTOS	10				
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**	**EXPERTO**
Longitud media de la clave	1.923	3.0	5.045	2.214	1.846
Desviación de la longitud	1.718	0.942	1.358	1.711	1.292
Longitud máxima de la clave	3	4	7	4	3
Longitud mínima de la clave	1	2	2	1	1
Número de hojas	13.0	12.0	22.0	14.0	13.0
Número de nodos internos	17.0	17.0	37.0	20.0	17.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0	0
Número de nodos exclusivos	13	12	22	14	13
Número de atributos utilizados en la clave	4	4	9	6	4
Número de caracteres confirmadores	3	1	2	3	3
Caracteres confirmadores diferentes	3	1	2	3	3

Tabla 4-4. *Medidas de la clave de la división Gymnospermae.*

Familia Cupressaceae.

FAMILIA CUPRESSACEAE				
NÚMERO DE OBJETIVOS	6			
NÚMERO DE ATRIBUTOS	10			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	2.5	2.5	4.615	2.857
Desviación de la longitud	0.748	0.748	1.319	0.916
Longitud máxima de la clave	3	3	7	4
Longitud mínima de la clave	1	1	2	1
Número de hojas	6.0	6.0	13.0	7.0
Número de nodos internos	9.0	9.0	24.0	11.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	6	6	13	7
Número de atributos utilizados en la clave	4	4	7	4
Número de caracteres confirmadores	3	3	4	3
Caracteres confirmadores diferentes	3	3	4	3

Tabla 4-5. *Medidas de la clave de la familia Cupressaceae.*

Familia Pinaceae.

FAMILIA PINACEAE				
NÚMERO DE OBJETIVOS	6			
NÚMERO DE ATRIBUTOS	9			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	2.166	2.5	3.571	2.857
Desviación de la longitud	0.776	0.748	1.036	0.916
Longitud máxima de la clave	3	3	5	4
Longitud mínima de la clave	1	1	1	1
Número de hojas	6.0	6.0	7.0	7.0
Número de nodos internos	9.0	9.0	12.0	11.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	6	6	7	7
Número de atributos utilizados en la clave	4	4	6	5
Número de caracteres confirmadores	2	1	3	2
Caracteres confirmadores diferentes	2	1	3	2

Tabla 4-6. *Medidas de la clave de la familia Pinaceae.*

Familia Taxodiaceae.

FAMILIA TAXODIACEAE				
NÚMERO DE OBJETIVOS	4			
NÚMERO DE ATRIBUTOS	13			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	2.25	2.25	2.8	2.0
Desviación de la longitud	0.737	0.737	0.931	1.0
Longitud máxima de la clave	3	3	4	3
Longitud mínima de la clave	1	1	1	1
Número de hojas	4.0	4.0	5.0	5.0
Número de nodos internos	6.0	6.0	8.0	7.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	4	4	5	5
Número de atributos utilizados en la clave	3	3	4	3
Número de caracteres confirmadores	9	9	5	8
Caracteres confirmadores diferentes	9	9	5	8

Tabla 4-7. *Medidas de la clave de la familia Taxodiaceae.*

Género Pinus.

GÉNERO PINUS						
NÚMERO DE OBJETIVOS	8					
NÚMERO DE ATRIBUTOS	16					
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**	**EXPERTO 1**	**EXPERTO 2**
Longitud media de la clave	1.25	1.25	5.689	3.636	2.9	2.8
Desviación de la longitud	0.979	0.979	1.687	0.893	0.327	0.845
Longitud máxima de la clave	2	2	8	5	3	4
Longitud mínima de la clave	1	1	2	2	2	2
Número de hojas	8.0	8.0	29.0	11.0	10.0	10.0
Número de nodos internos	9.0	9.0	48.0	18.0	15.0	14.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0	0	0
Número de nodos exclusivos	8	8	29	11	10	10
Número de atributos utilizados en la clave	2	2	10	7	4	5
Número de caracteres confirmadores	3	3	10	9	5	6
Caracteres confirmadores diferentes	3	3	10	9	5	6

Tabla 4-8. *Medidas de la clave del género Pinus.*

Género Juniperus.

GÉNERO JUNIPERUS						
NÚMERO DE OBJETIVOS	11					
NÚMERO DE ATRIBUTOS	19					
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**	**EXPERTO 1**	**EXPERTO 2**
Longitud media de la clave	1.545	3.272	5.647	4.65	2.545	3.363
Desviación de la longitud	1.068	0.874	1.541	1.673	0.648	0.760
Longitud máxima de la clave	2	4	8	7	3	4
Longitud mínima de la clave	1	1	2	2	2	2
Número de hojas	11.0	11.0	34.0	20.0	11.0	11.0
Número de nodos internos	14.0	16.0	62.0	31.0	18.0	19.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0	0	0
Número de nodos exclusivos	11	11	34	20	11	11
Número de atributos utilizados en la clave	4	5	10	7	6	8
Número de caracteres confirmadores	7	6	15	9	5	8
Caracteres confirmadores diferentes	7	6	15	9	5	8

Tabla 4-9. *Medidas de la clave del género Juniperus.*

Género Cupressus.

GÉNERO CUPRESSUS				
NÚMERO DE OBJETIVOS	4			
NÚMERO DE ATRIBUTOS	8			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	1.8	1.75	2.6	1.8
Desviación de la longitud	0.496	0.494	0.688	0.496
Longitud máxima de la clave	2	2	3	2
Longitud mínima de la clave	1	1	1	1
Número de hojas	5.0	4.0	5.0	5.0
Número de nodos internos	7.0	5.0	8.0	7.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	5	4	5	5
Número de atributos utilizados en la clave	3	2	4	3
Número de caracteres confirmadores	3	0	5	3
Caracteres confirmadores diferentes	3	0	5	3

Tabla 4-10. *Medidas de la clave del género Cupressus.*

Género Ephedra.

GÉNERO EPHEDRA				
NÚMERO DE OBJETIVOS	3			
NÚMERO DE ATRIBUTOS	6			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	1.0	1.666	2.8	1.0
Desviación de la longitud	0.0	0.489	0.931	0.0
Longitud máxima de la clave	1	2	4	1
Longitud mínima de la clave	1	1	1	1
Número de hojas	3.0	3.0	5.0	3.0
Número de nodos internos	3.0	4.0	8.0	3.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	3	3	5	3
Número de atributos utilizados en la clave	1	2	4	1
Número de caracteres confirmadores	1	2	2	1
Caracteres confirmadores diferentes	1	2	2	1

Tabla 4-11. *Medidas de la clave del género Ephedra.*

Género Pinus.

GÉNERO ABIES				
NÚMERO DE OBJETIVOS	2			
NÚMERO DE ATRIBUTOS	5			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	1.0	1.0	1.0	1.0
Desviación de la longitud	0.0	0.0	0.0	0.0
Longitud máxima de la clave	1	1	1	1
Longitud mínima de la clave	1	1	1	1
Número de hojas	2.0	2.0	2.0	2.0
Número de nodos internos	2.0	2.0	2.0	2.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	2	2	2	2
Número de atributos utilizados en la clave	1	1	1	1
Número de caracteres confirmadores	4	4	4	4
Caracteres confirmadores diferentes	4	4	4	4

Tabla 4-12. *Medidas de la clave del género Abies.*

Género Cycas.

GÉNERO CYCAS				
NÚMERO DE OBJETIVOS	2			
NÚMERO DE ATRIBUTOS	3			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	1.0	1.0	1.0	1.0
Desviación de la longitud	0.0	0.0	0.0	0.0
Longitud máxima de la clave	1	1	1	1
Longitud mínima de la clave	1	1	1	1
Número de hojas	2.0	2.0	2.0	2.0
Número de nodos internos	2.0	2.0	2.0	2.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	2	2	2	2
Número de atributos utilizados en la clave	1	1	1	1
Número de caracteres confirmadores	2	2	2	2
Caracteres confirmadores diferentes	2	2	2	2

Tabla 4-13. *Medidas de la clave del género Cycas.*

Género Cedrus.

GÉNERO CEDRUS				
NÚMERO DE OBJETIVOS	2			
NÚMERO DE ATRIBUTOS	6			
CRITERIO DE DIVISIÓN	**ENTROPÍA**	**PROPORCIÓN DE GANANCIA**	**DIVERSIDAD DE GINI**	**MEDIDA DE DALLWIITZ**
Longitud media de la clave	1.0	1.0	1.666	1.75
Desviación de la longitud	0.0	0.0	0.692	0.699
Longitud máxima de la clave	1	1	2	2
Longitud mínima de la clave	1	1	1	1
Número de hojas	2.0	2.0	6.0	8.0
Número de nodos internos	2.0	2.0	8.0	10.0
Número medio de opciones en nodos hoja	1.0	1.0	1.0	1.0
Desviación del numero de opciones	0.0	0.0	0.0	0.0
Número de nodos OR	0	0	0	0
Número de nodos exclusivos	2	2	6	8
Número de atributos utilizados en la clave	1	1	2	2
Número de caracteres confirmadores	2	2	3	0
Caracteres confirmadores diferentes	2	2	3	0

Tabla 4-14. *Medidas de la clave del género Cedrus.*

INTERPRETACIÓN CUANTITATIVA DE LOS RESULTADOS OBTENIDOS

Observaciones sobre la longitud de las claves generadas.

En cuanto a la longitud de las claves, un factor común es la poca adecuación de las que utilizan como criterio de división el índice de diversidad de Gini. Son demasiado largas y la selección de caracteres no resulta adecuada. En las tablas resumen y en las gráficas se observa que producen siempre las claves más largas (ver Figura 4-4) y los árboles más grandes, es decir, con mayor número de nodos internos (ver Figura 4-6) y externos (ver Figura 4-5). Prácticamente en todos los casos superan el promedio de estos valores.

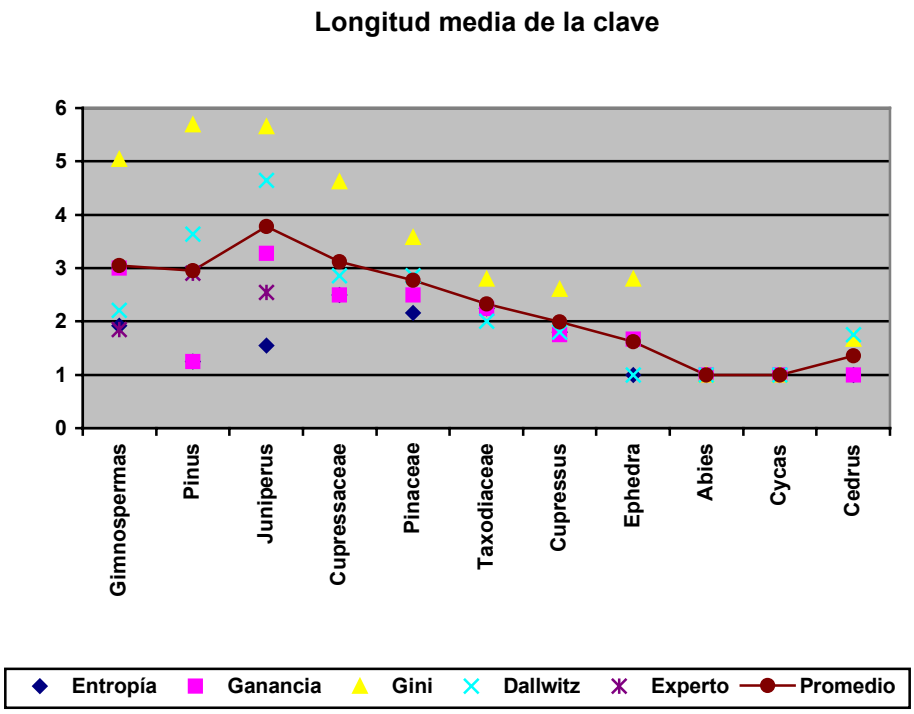


Figura 4-4. *Comparación de la longitud media de las claves.*

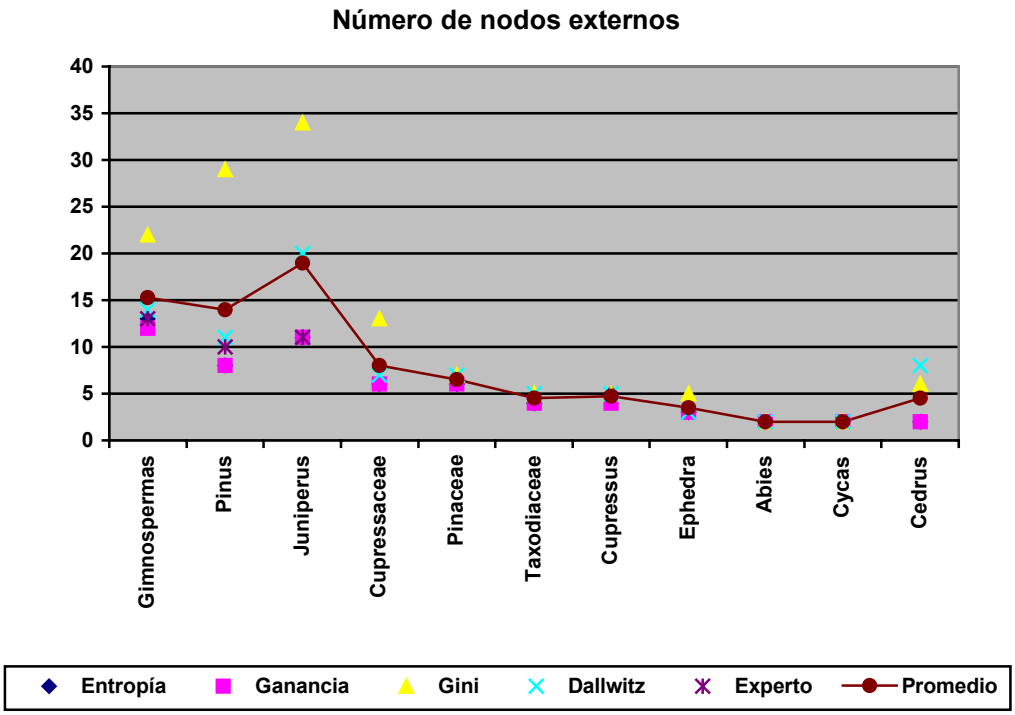


Figura 4-5. *Número de nodos externos.*

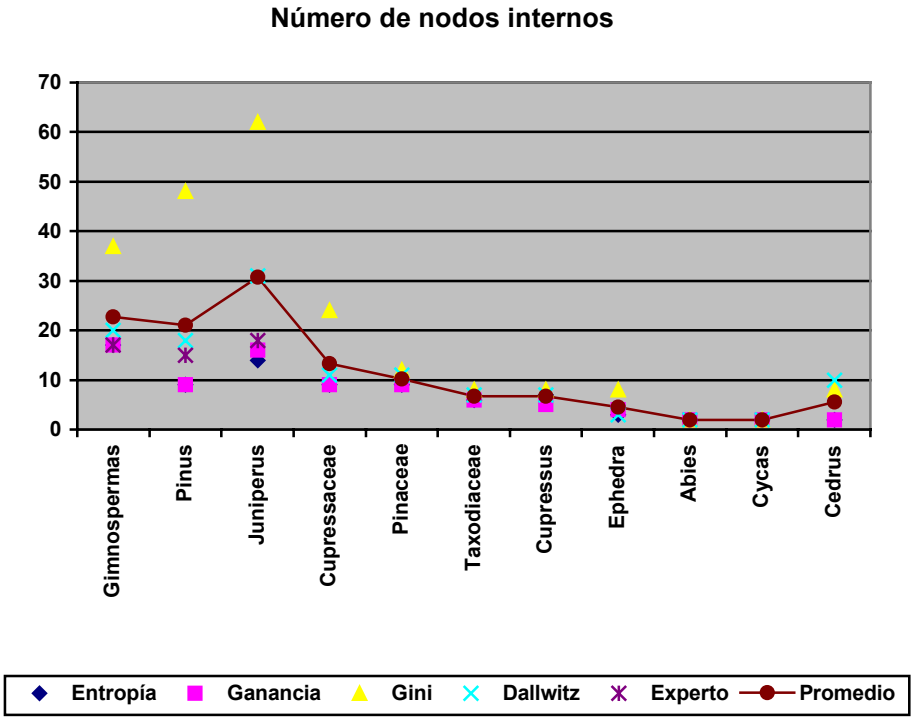


Figura 4-6. *Número de nodos internos.*

Si tenemos en cuenta la desviación con respecto a la longitud mínima de las claves, nuevamente la entropía es el mejor criterio, esto es, las claves generadas con esta regla de división son prácticamente en todos los casos de longitud mínima (Figura 4-7). En segundo lugar está el criterio de la proporción de ganancia, seguido del criterio de Dallwitz, que no produce tan buenos resultados, debido a que la medida que utiliza para contabilizar la variabilidad intra-taxon es más grosera que la utilizada por los dos criterios anteriores. Dallwitz solamente tiene en cuenta el número de *taxa* distintos en cada partición, mientras que los otros dos contabilizan además la frecuencia de cada *taxa*.

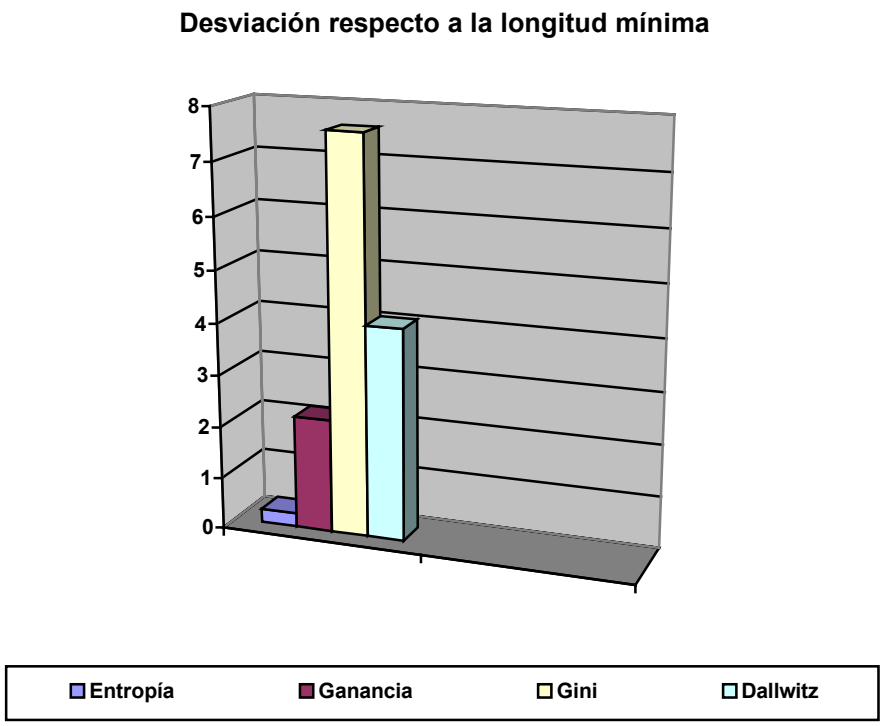


Figura 4-7. *Desviación respecto a la longitud mínima.*

Observación sobre el equilibrio de los árboles.

La desviación de la longitud de las ramas del árbol con respecto de la media es también mayor para la medida de Gini, lo que indica que la longitud de las ramas de los árboles generados con este criterio es más variable Figura 4-8. Las claves más equilibradas son las generadas con el criterio de la proporción de ganancia (Figura 4-9), seguidas de las generadas con la entropía y el criterio de división de Dallwitz.

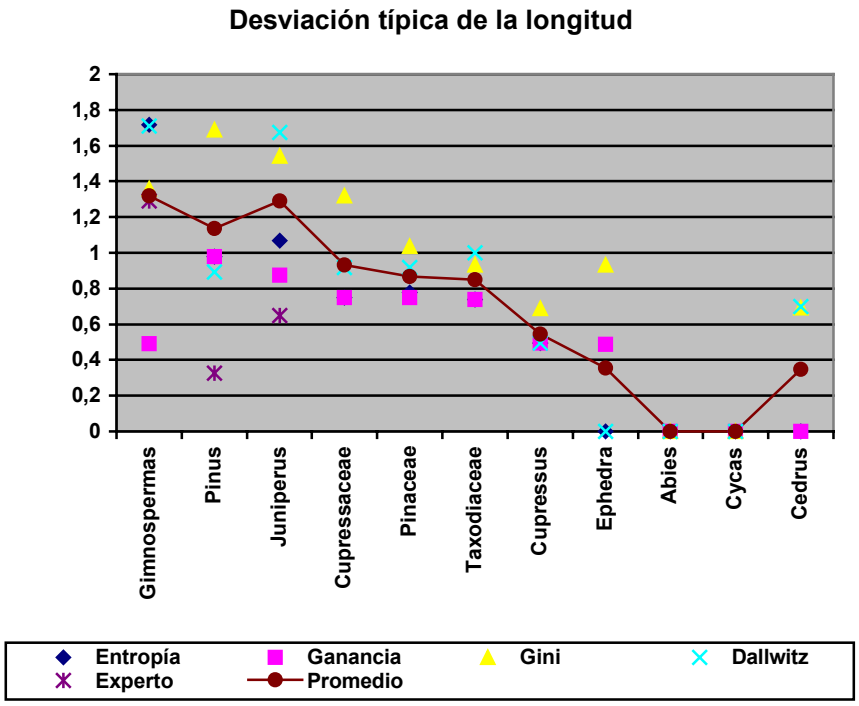


Figura 4-8. *Desviación típica de la longitud.*

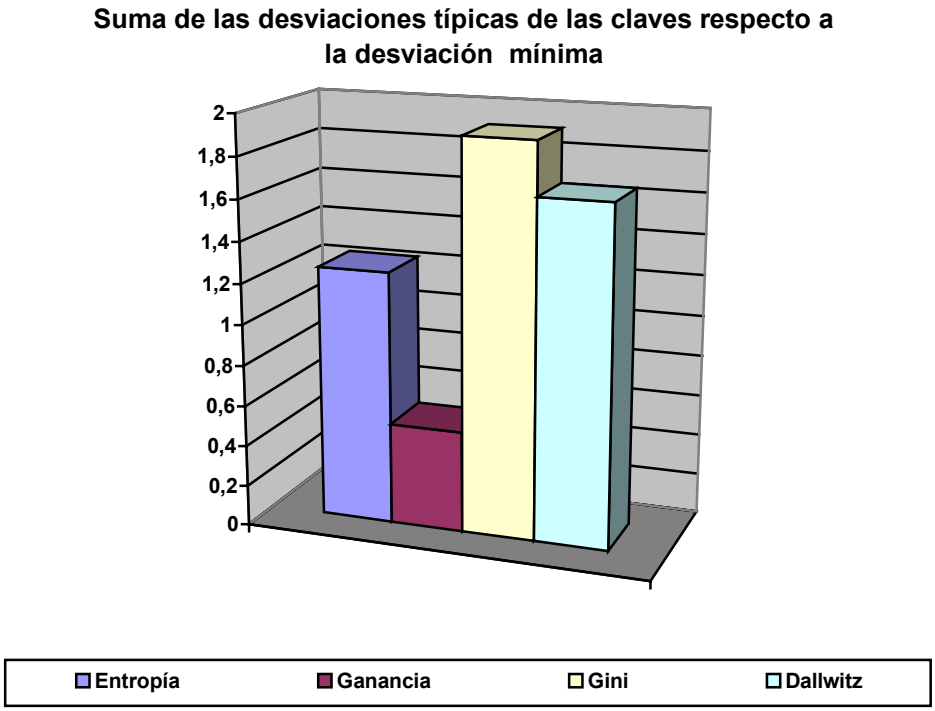


Figura 4-9. *Suma de las desviaciones típicas respecto de la desviación mínima.*

Observaciones sobre el número de caracteres confirmadores incluidos.

El índice de diversidad de Gini parece que permite incluir un mayor número de caracteres confirmadores (Figura 4-10). No obstante, esto es debido a que el árbol generado es mayor que con otros criterios. El criterio de la entropía es el que produce una mejor relación *tamaño del árbol / número de atributos confirmadores* (Figura 4-11), seguido del criterio de proporción de ganancia.

La Figura 4-12 reafirma este resultado. Si observamos la desviación respecto del máximo de la relación *confirmadores / nodos internos* en cada caso, vemos como la entropía es la que presenta una desviación menor, seguida del criterio de proporción de ganancia y, de lejos, por los criterios de división de Dallwitz y Gini.

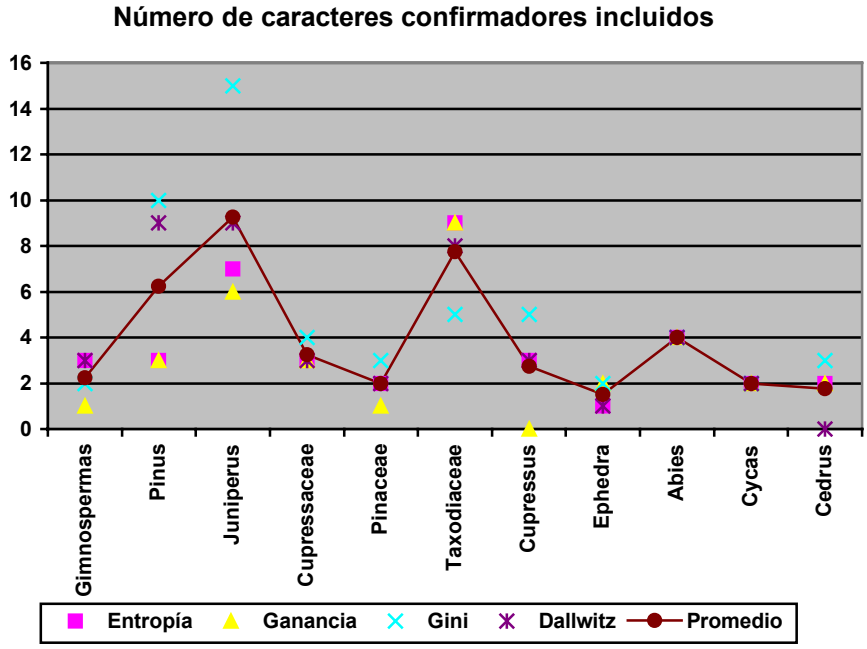


Figura 4-10. *Número de caracteres confirmadores incluidos.*

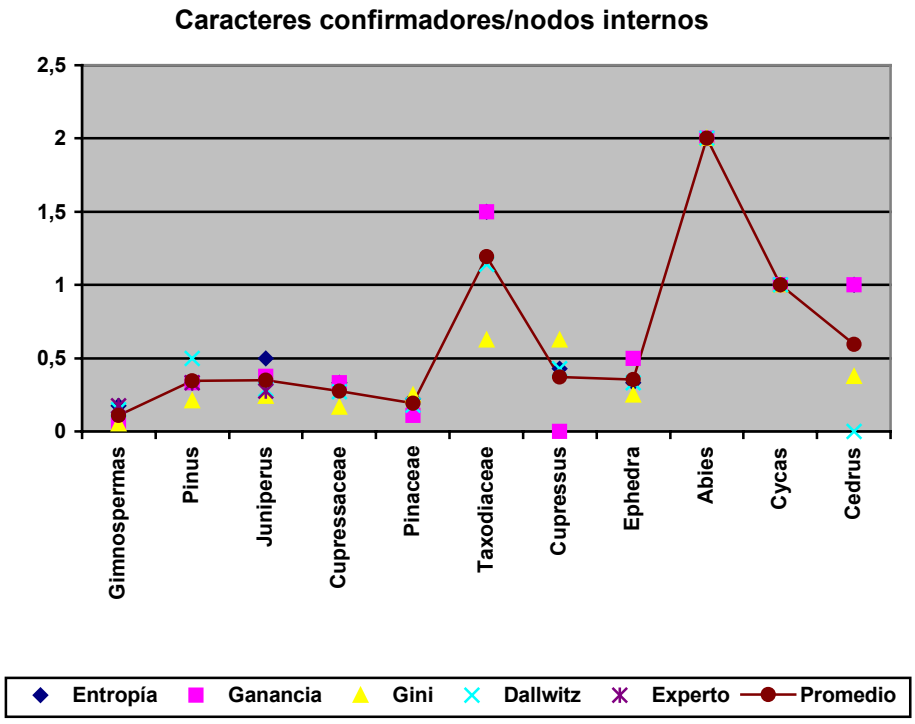


Figura 4-11. *Número de caracteres confirmadores por nodo interno.*

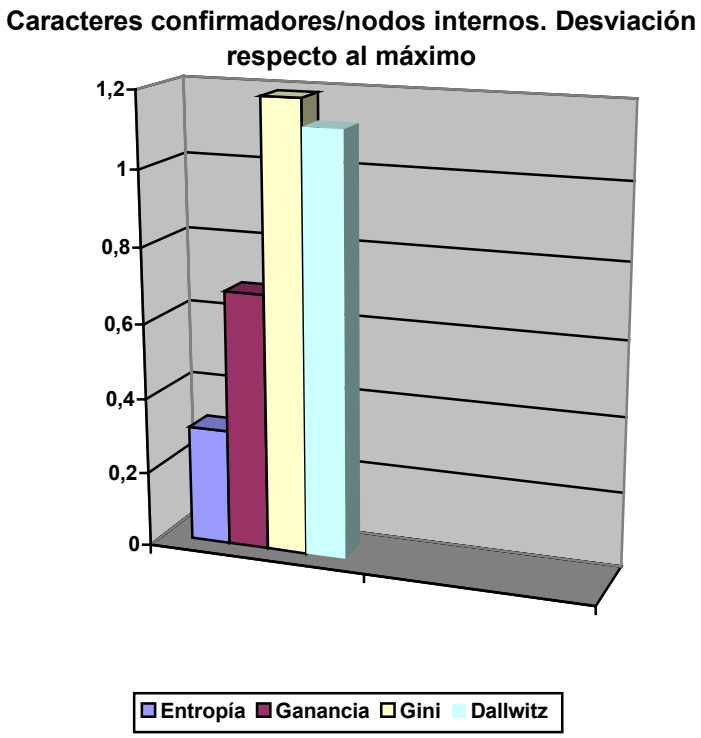


Figura 4-12. *Número de caracteres confirmadores por nodo interno. Desviación respecto al máximo.*

Observaciones sobre el número de caracteres utilizados.

En lo que al número de caracteres diferentes utilizados, la entropía es nuevamente el criterio que utiliza un menor número de caracteres en la generación de sus claves. Esto es, localiza la menor cantidad de información para llevar a cabo la identificación (Figura 4-13). En segundo lugar encontramos al criterio de proporción de ganancia, seguido del criterio de Dallwitz. El índice de diversidad de Gini, además de producir las claves más largas, incluye un mayor número de atributos en la generación de claves.

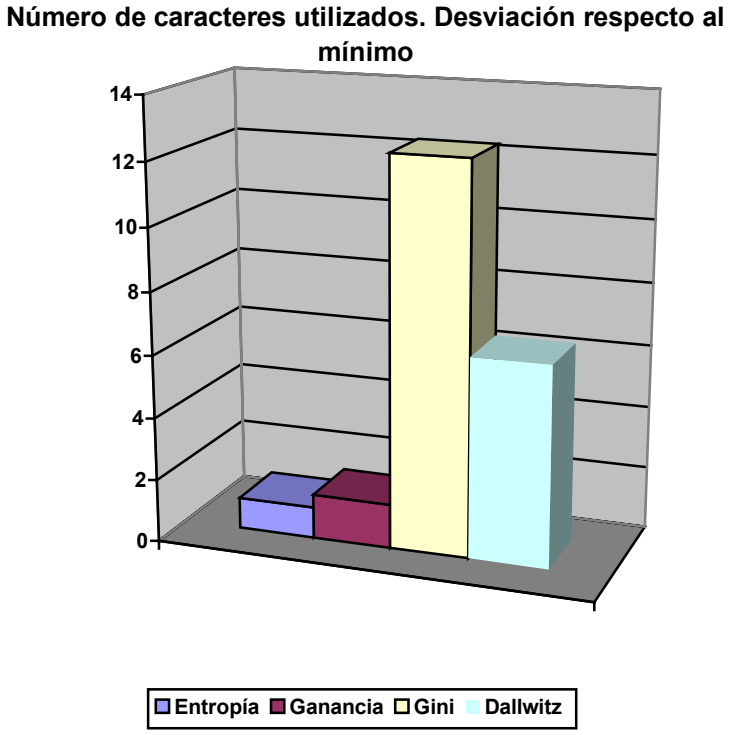


Figura 4-13. *Número de caracteres utilizados. Desviación respecto al mínimo.*

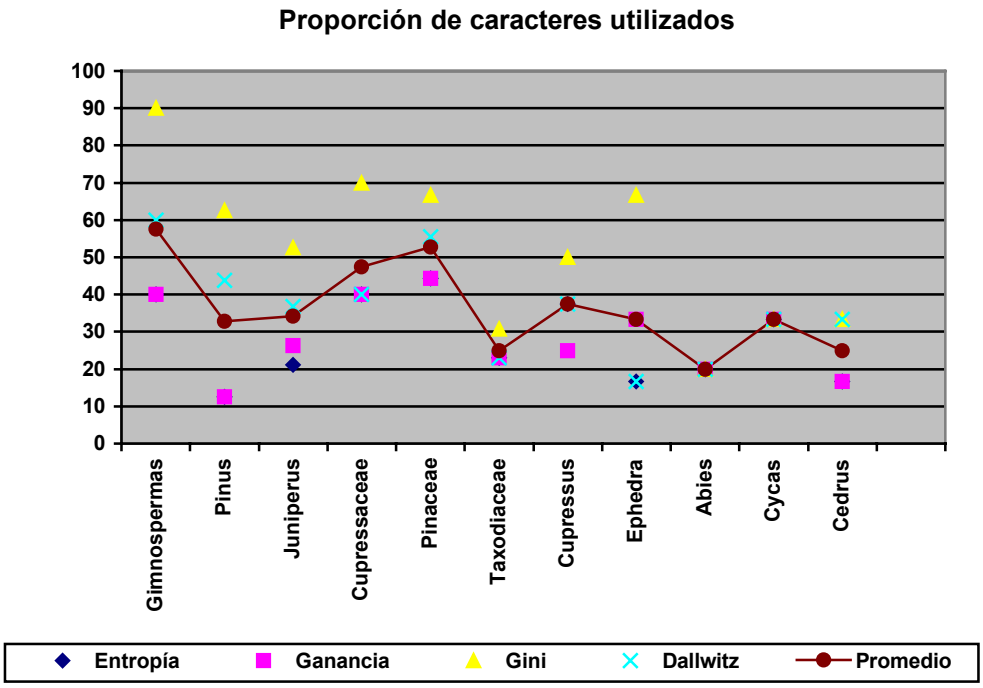


Figura 4-14. *Proporción de caracteres utilizados.*

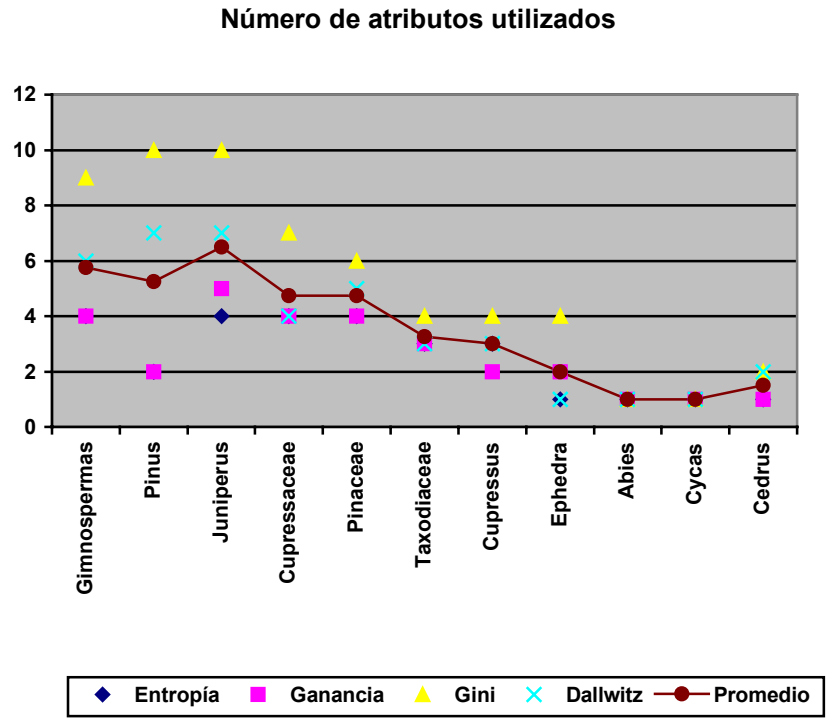


Figura 4-15. *Número de atributos utilizado en la clave.*

Comparación práctica de los criterios.

El criterio de proporción de ganancia produce resultados bastante buenos en cuanto a la longitud media de la clave, poder generador de caracteres confirmadores, número de atributos utilizados, etc. Las claves no resultan tan económica con las producidas con el criterio de la entropía, debido a que esta regla de división intenta equilibrar más los árboles. Esto le lleva a pasos redundantes no necesarios. Es más eficaz la entropía, que apura más la capacidad de discriminación.

La medida utilizada por Dallwitz también produce buenos resultados, pero en algunos casos no deseados, puesto que mide de forma más grosera la variabilidad intra-taxon. La entropía realiza una medida más precisa. La diferencia con el criterio de división de Dallwitz se ve clara en el caso de las claves de la familia *Taxodiaceae* (Tabla 4-15 y Tabla 4-16) y del género *Cedrus* (Tabla 4-17 y Tabla 4-18).

FAMILIA *TAXODIACEAE* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave:3	*Total de caracteres confirmadores:9*
Atributos incluidos en la clave: *Hoja: 1 vez* *Forma de la hoja: 1 vez* *Tronco de 10 a 12 m de diámetro: 1 vez*	*Caracteres confirmadores incluidos:* *Con raíces aéreas: 1 vez* *Base del tronco: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Envés con dos bandas: 1 vez* *Corteza rojiza al desprenderse en placas: 1 vez* *Corteza fibrosa: 1 vez* *Tamaño de la piña (largo): 1 vez* *Escamas de la piña peltadas: 1 vez*
(0) Hoja Caduca; Con raíces aéreas Si; Base del tronco Ensanchada .....Género *Taxodium* (0) Hoja Persistente; Con raíces aéreas No; Base del tronco No ensanchada.....1 (1) Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Envés con dos bandas Si .....Género *Sequoia* (1) Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Envés con dos bandas No.....2 (2) Tronco de 10 a 12 m de diámetro Si; Corteza rojiza al desprenderse en placas No; Corteza fibrosa No; Tamaño de la piña (largo) Entre 4 y 6 cm; Escamas de la piña peltadas Si .....Género *Sequoiadendron* (2) Tronco de 10 a 12 m de diámetro No; Corteza rojiza al desprenderse en placas Si; Corteza fibrosa Si; Tamaño de la piña (largo) Entre 2 y 4 cm; Escamas de la piña peltadas No .....Género *Cryptomeria*	

Tabla 4-15. *Clave para la familia Taxodiaceae generada según el criterio de Mínima Entropía.*

FAMILIA *TAXODIACEAE* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:3	*Total de caracteres confirmadores:8*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 1 vez* *Hoja: 1 vez* *Forma de la hoja: 1 vez*	*Caracteres confirmadores incluidos:* *Envés con dos bandas: 1 vez* *Con raíces aéreas: 1 vez* *Corteza rojiza al desprenderse en placas: 1 vez* *Base del tronco: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Escamas de la piña peltadas: 1 vez* *Escamas de la piña recurvadas, con 4-6 esquinas: 1 vez*
(0) Tamaño de la piña (largo) Entre 1 y 2 cm .....Género *Taxodium* (0) Tamaño de la piña (largo) Entre 2 y 4 cm.....1 (1) Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Escamas de la piña peltadas Si; Escamas de la piña recurvadas, con 4-6 esquinas No.....2 (2) Hoja Caduca; Envés con dos bandas No; Con raíces aéreas Si; Corteza rojiza al desprenderse en placas No; Base del tronco Ensanchada .....Género *Taxodium* (2) Hoja Persistente; Envés con dos bandas Si; Con raíces aéreas No; Corteza rojiza al desprenderse en placas Si; Base del tronco No ensanchada .....Género *Sequoia* (1) Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Escamas de la piña peltadas No; Escamas de la piña recurvadas, con 4-6 esquinas Si .....Género *Cryptomeria* (0) Tamaño de la piña (largo) Entre 4 y 6 cm .....Género *Sequoiadendron*	

Tabla 4-16. *Clave para la familia Taxodiaceae generada según el criterio de Dallwitz.*

Al utilizar el criterio de la entropía, la clave solo presenta un camino para cada taxon de forma que en la clave no aparece un taxon (género) como objetivo final (nodo hoja) más de una vez. Al utilizar el criterio de división de Dallwitz aparecen 5 objetivos finales para cuatro taxones, de forma que un taxon aparece para dos caminos estando a nivel de género.

GÉNERO *CEDRUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Guía del árbol: 1 vez* *Ramas colgantes: 1 vez*
(0) Tamaño 6000 cm; Guía del árbol Recurvada; Ramas colgantes Si .....Especie *Cedrus deodara* (0) Tamaño 5000 cm; Guía del árbol No recurvada; Ramas colgantes No .....Especie *Cedrus atlántica*	

Tabla 4-17. *Clave para el género Cedrus generada según el criterio de Mínima Entropía.*

Esta diferencia entre ambos criterios queda aún más patente en el caso de las claves generadas para el género *Cedrus*. Para separar dos *taxa*, Dallwitz da 5 caminos para *C. atlántica* y 3 para *C. deodara*. Con la entropía tenemos 2 clases de *Cedrus* y dos caminos para llegar.

GÉNERO *CEDRUS* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:2	*Total de caracteres confirmadores:0*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 2 veces* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:*
(0) Tamaño de las hojas (largo) Entre 2 y 2.5 cm.....1 (1) Tamaño de la piña (largo) Entre 8 y 12 cm .....Especie *Cedrus deodara* (1) Tamaño de la piña (largo) Entre 4 y 6 cm .....Especie *Cedrus atlántica* (1) Tamaño de la piña (largo) Entre 6 y 8 cm .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 2.5 y 3 cm.....2 (2) Tamaño de la piña (largo) Entre 8 y 12 cm .....Especie *Cedrus deodara* (2) Tamaño de la piña (largo) Entre 4 y 6 cm .....Especie *Cedrus atlántica* (2) Tamaño de la piña (largo) Entre 6 y 8 cm .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 3 y 5 cm .....Especie *Cedrus deodara* (0) Tamaño de las hojas (largo) Entre 1 y 2 cm .....Especie *Cedrus atlántica*	

Tabla 4-18. *Clave para el género Cedrus generada según el criterio de Dallwitz.*

DISCUSIÓN E INTERPRETACIÓN BIOLÓGICA DE LOS RESULTADOS OBTENIDOS.

La entropía produce los resultados más económicos y las claves más optimizadas en cuanto al número de preguntas. En el ámbito de la Biología, este aspecto debe ser matizado. En ocasiones, el carácter diferenciador desde el punto de vista de la teoría de la información, no tiene por qué ser el mejor carácter desde el punto de vista biológico debido, por ejemplo, a la dificultad para su observación (necesidad de observación microscópica, temporalidad, tamaño, disposición, etc). Por otro lado el no disponer de un carácter interrumpe el proceso de identificación, por lo que no es muy lógico incluir en primer lugar en una clave un carácter de difícil observación. Es necesario incluir en los primeros pasos de la

clave aquellos caracteres que, además de tener un alto poder diferenciador, son más fácilmente observables.

Hay que tener en cuenta, además de los criterios de división, las características del grupo taxonómico objeto de estudio y los futuros receptores de las claves. No es lo mismo hacer claves de expertos para expertos que de expertos para alumnos (carácter docente) o para un público general (carácter más divulgativo).

Puede suceder que claves que desde el punto de vista de la teoría de la información son más complicadas, como en el caso de las generadas con el índice de diversidad de Gini o con el criterio de Dalwitz, sean de gran validez para un público no adentrado en el conocimiento vegetal al utilizar la información de forma reiterativa. Dadas las dimensiones de las claves generadas con el criterio de Gini, no recomendamos esta opción para conjuntos de datos con demasiados caracteres.

La posibilidad de elegir entre los 4 criterios que se plantean en nuestro trabajo viene justificada por permitir elaborar distintos tipos de claves según el tipo de trabajo, investigación y usuario final. Si la clave en si misma va a ser toda la información que tengamos respecto a los taxones, necesitaremos claves largas que recojan el mayor número de caracteres posibles para la identificación de un taxon. Sin embargo, si las claves son la herramienta de acceso a la información de manera que, una vez identificado el taxon, este se acompaña de una descripción exhaustiva podremos utilizar claves muy precisas y de alto nivel de discriminación.

Por otra parte, al generar claves de expertos para expertos, lo que presupone un conocimiento avanzado tanto del grupo como de la morfología y características vegetales, si implicaría el uso de criterios como el de la entropía para la realización de las mismas. El caso de considerar las claves como una herramienta para un usuario intermedio, para el estudio de la Botánica como es nuestro caso, es el que nos llevaría muy probablemente a utilizar el criterio más óptimo según la teoría de la información, pero combinado con la interactividad.

Todo lo anteriormente expuesto se puede comprobar en las claves obtenidas para la identificación a nivel de familia de las Gimnospermas consideradas. Si observamos las cuatro claves (consultar Apéndice B), se comprueba que, como dijimos anteriormente, la entropía es la que obtiene la clave más corta y directa, frente a la obtenida mediante el criterio de Gini. A pesar de ello, puede suceder que los atributos sobre los cuales la entropía ha generado su clave no sean aquellos que un experto en el área considere más apropiados. En estos casos, la aplicación de la interactividad puede combinar ambos resultados. La combinación de la interactividad con un criterio de división como la entropía produce resultados muy satisfactorios. La entropía sugiere y ordena los caracteres, pero es el experto el que finalmente decide qué carácter seleccionar. En el caso de la clave de familias de Gimnospermas conseguimos incluso reducir la longitud media de la clave con respecto a la entropía (Tabla 4-19 y Tabla 4-20).

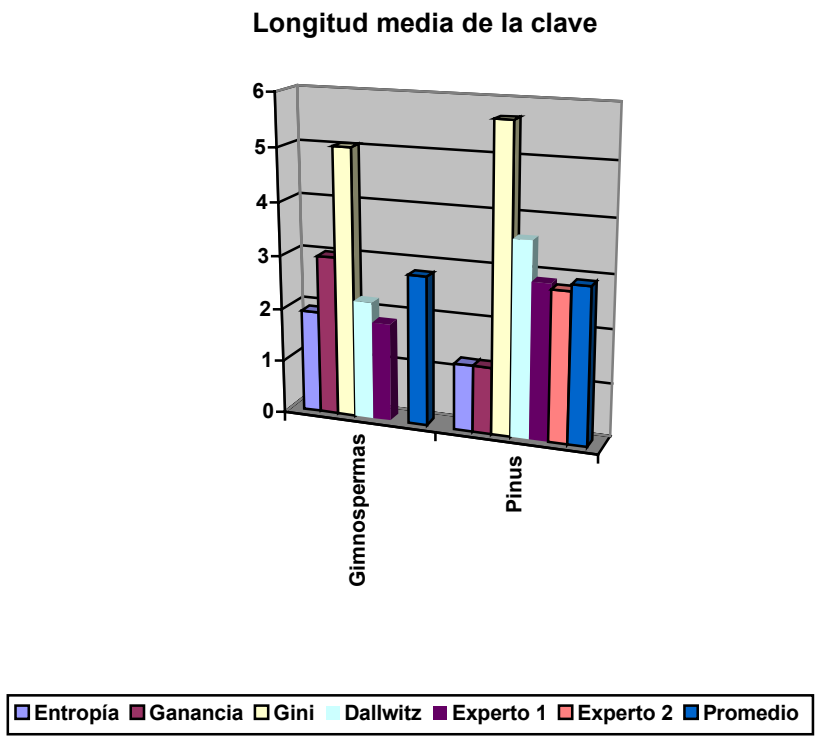


Figura 4-16. *Longitud media de las claves para la división Gymnospermae y el género Pinus.*

DIVISIÓN GYMNOSPERMAE/ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Semilla con arilo: 1 vez* *Planta: 1 vez* *Forma de la hoja: 2 veces* *Disposición de las hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Resinosa: 1 vez* *Consistencia de la fructificación: 1 vez* *Semillas numerosas: 1 vez*
(0) Disposición de las hojas En espiral.....1 (1) Planta Dioica; Consistencia de la fructificación Carnosa; Semillas numerosas No.....2 (2) Semilla con arilo Si; Resinosa No .....Familia *Taxaceae* (2) Semilla con arilo No; Resinosa Si .....Familia *Cephalotaxaceae* (1) Planta Monoica; Consistencia de la fructificación Leñosa; Semillas numerosas Si.....3 (3) Forma de la hoja Linear (planoaguzada) .....Familia *Taxodiaceae* (3) Forma de la hoja Acicular .....Familia *Pinaceae* (3) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (0) Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (0) Disposición de las hojas Ternadas .....Familia *Cupressaceae* (0) Disposición de las hojas Imbricadas.....4 (4) Forma de la hoja Escamosa .....Familia *Cupressaceae* (4) Forma de la hoja Escamosa curvada hacia el ápice .....Familia *Araucariaceae* (0) Disposición de las hojas En verticilos de cuatro .....Familia *Cupressaceae* (0) Disposición de las hojas En los nudos del tallo .....Familia *Ephedraceae* (0) Disposición de las hojas Hojas agrupadas en la terminación de brotes laterales (braquiblastos) .....Familia *Ginkgoaceae* (0) Disposición de las hojas Formando una corona apical .....Familia *Cycadaceae*	

Tabla 4-19. *Clave para la división Gymnospermae generada según el criterio de mínima entropía.*

DIVISIÓN GYMNOSPERMAE /CRITERIO EXPERTO	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Semilla con arilo: 1 vez* *Planta: 1 vez* *Forma de la hoja: 1 vez* *Disposición de las hojas: 2 veces*	*Caracteres confirmadores incluidos:* *Resinosa: 1 vez* *Consistencia de la fructificación: 1 vez* *Semillas numerosas: 1 vez*
(0) Forma de la hoja Linear (planoaguzada).....1 (1) Planta Dioica; Consistencia de la fructificación Carnosa; Semillas numerosas No.....2 (2) Semilla con arilo Si; Resinosa No .....Familia *Taxaceae* (2) Semilla con arilo No; Resinosa Si .....Familia *Cephalotaxaceae*	

DIVISIÓN GYMNOSPERMAE /CRITERIO EXPERTO
(1) Planta Monoica; Consistencia de la fructificación Leñosa; Semillas numerosas Si .....Familia *Taxodiaceae* (0) Forma de la hoja Acicular.....3 (3) Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (3) Disposición de las hojas Ternadas .....Familia *Cupressaceae* (3) Disposición de las hojas En espiral .....Familia *Pinaceae* (0) Forma de la hoja Escamosa.....4 (4) Disposición de las hojas Imbricadas .....Familia *Cupressaceae* (4) Disposición de las hojas En verticilos de cuatro .....Familia *Cupressaceae* (4) Disposición de las hojas En los nudos del tallo .....Familia *Ephedraceae* (0) Forma de la hoja En forma de abanico, con escotadura central .....Familia *Ginkgoaceae* (0) Forma de la hoja En forma de palmera .....Familia *Cycadaceae* (0) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (0) Forma de la hoja Escamosa curvada hacia el ápice .....Familia *Araucariaceae*

Tabla 4-20. *Clave para la División Gymnospermae generada según el criterio del Experto.*

Un caso similar sucede con las claves del género *Pinus*, en este caso, el carácter sobre el que sustenta la clave es "características de la apófisis", mediante el cual logra separar muy rápidamente las distintas especies de pinos. Sin embargo, es un carácter no fácilmente apreciable debido a referirse a la piña (no siempre va a estar presente), además de referirse a un aspecto que supone un conocimiento de la morfología vegetal que no tiene por qué conocer cualquier usuario (Tabla 4-21).

GÉNERO *PINUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 2	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Características de la apófisis: 1 vez* *Color de la corteza (ritidoma): 1 vez*	*Caracteres confirmadores incluidos:* *Piñas brillantes: 1 vez* *Hoja de color: 1 vez* *Con hojas: 1 vez*
(0) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (0) Características de la apófisis Poco prominente.....1 (1) Color de la corteza (ritidoma) Gris-ceniciento; Piñas brillantes Si; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus nigra subsp. salzmannii* (1) Color de la corteza (ritidoma) Pardo-rojizo; Piñas brillantes No; Hoja de color Verde claro; Con hojas Rigidas .....Especie *Pinus sylvestris* (0) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata*	

GÉNERO *PINUS* /ENTROPÍA
(0) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (0) Características de la apófisis Convexa .....Especie *Pinus pinea* (0) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (0) Características de la apófisis Prominente .....Especie *Pinus canariensis*

Tabla 4-21. *Clave para el género Pinus generada según el criterio de Mínima Entropía.*

El conocimiento del grupo taxonómico concreto por parte de los expertos justifica que, al elegir nuevamente la entropía como criterio más útil para elaborar las claves, se haga uso de la interactividad para forzar la elección de otros criterios de arranque. Según la opinión de los expertos un buen atributo sería "forma de la hoja", la clave resultante de forzar la selección de este atributo de entrada dio como resultado la clave de la Tabla 4-22. Esta clave sólo utiliza la función interactiva para el atributo de partida.

GÉNERO *PINUS* /CRITERIO EXPERTO1	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:5*
Atributos incluidos en la clave: *Características de la apófisis: 2 veces* *Con hojas: 2 veces* *Número de hojas por fascículo: 1 vez* *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Color de la corteza (ritidoma): 1 vez* *Tamaño de la piña (ancho): 1 vez* *Hoja de color: 1 vez* *Características de la apófisis: 1 vez* *Tamaño de las hojas (ancho): 1 vez*
(0) Número de hojas por fascículo 2.....1 (1) Con hojas Rígidas.....2 (2) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (2) Características de la apófisis Poco prominente .....Especie *Pinus sylvestris* (2) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (1) Con hojas Flexibles.....3 (3) Características de la apófisis Poco prominente .....Especie *Pinus nigra subsp. salzmannii* (3) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (3) Características de la apófisis Convexa .....Especie *Pinus pinea* (3) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (0) Número de hojas por fascículo 3.....4 (4) Con hojas Rígidas; Color de la corteza (ritidoma) Gris-ceniciento; Tamaño de la piña (ancho) Entre 2 y 4 cm; Hoja de color Verde oscuro .....Especie *Pinus uncinata* (4) Con hojas Flexibles; Color de la corteza (ritidoma) Pardo-rojizo; Tamaño de la piña (ancho)	

GÉNERO *PINUS* /CRITERIO EXPERTO1
Entre 4 y 7.5 cm; Hoja de color Verde intenso.....5 (5) Tamaño 4000 cm; Características de la apófisis Muy prominente y punzante; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm .....Especie *Pinus radiata* (5) Tamaño 6000 cm; Características de la apófisis Prominente; Tamaño de las hojas (ancho) Hasta 0.1 cm .....Especie *Pinus canariensis*

Tabla 4-22. *Clave para el género Pinus generada según el criterio del Experto.*

La interactividad puede utilizarse en otros niveles. Continuando con el ejemplo de *Pinus* se forzó que el segundo atributo fuera la "presencia / ausencia de piñón", lo que dio como resultado la Tabla 4-23, que vino a ser la clave que mejor respondía a las expectativas del experto.

GÉNERO *PINUS* /CRITERIO EXPERTO2	
Total de atributos diferentes incluidos en la clave: 5	*Total de caracteres confirmadores:6*
Atributos incluidos en la clave: *Características de la apófisis: 1 vez* *Color de la corteza (ritidoma): 1 vez* *Con piñón: 1 vez* *Número de hojas por fascículo: 1 vez* *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Piñas brillantes: 1 vez* *Hoja de color: 1 vez* *Con hojas: 1 vez* *Forma de la copa: 1 vez* *Semilla alada y persistente: 1 vez* *Características de la apófisis: 1 vez*
(0) Número de hojas por fascículo 2.....1 (1) Con piñón No; Forma de la copa Piramidal; Semilla alada y persistente Si.....2 (2) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (2) Características de la apófisis Poco prominente.....3 (3) Color de la corteza (ritidoma) Gris-ceniciento; Piñas brillantes Si; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus nigra subsp. salzmannii* (3) Color de la corteza (ritidoma) Pardo-rojizo; Piñas brillantes No; Hoja de color Verde claro; Con hojas Rígidas .....Especie *Pinus sylvestris* (2) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (2) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (2) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (1) Con piñón Si; Forma de la copa Aparasolada; Semilla alada y persistente No .....Especie *Pinus pinea* (0) Número de hojas por fascículo 3.....4 (4) Tamaño 2500 cm; Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (4) Tamaño 4000 cm; Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (4) Tamaño 6000 cm; Características de la apófisis Prominente .....Especie *Pinus canariensis*	

Tabla 4-23. *Clave para el género Pinus generada según el criterio del Experto.*

Todo esto apoya la validez de la herramienta, por ofertar varios criterios de división de entre los cuales el usuario puede elegir el que considere más adecuado para su trabajo. Esto combinado con la opción interactiva, da como resultado claves totalmente optimizadas y satisfactorias para los usuarios.

La herramienta *XKey* aplica técnicas inteligentes a la generación de claves de identificación. No obstante, debemos hacer notar que la adecuación de los resultados depende, además de todo lo expuesto, de las características del conjunto de datos y del problema concreto que sea abordado. Una buena selección de caracteres de identificación conducirá a claves mucho más satisfactorias y adaptadas a la realidad.

Conclusiones.

Recordemos que el objetivo de este trabajo era el de diseñar y desarrollar un sistema de identificación vegetal correcto desde el punto de vista de la Inteligencia Artificial y la Ingeniería del Software y capaz de acomodarse a las particularidades de la Biología. A este planteamiento general se añadía el requerimiento de utilizar modelos estándar de representación del conocimiento taxonómico y la generación de claves de identificación. En esta memoria se ha presentado un conjunto de herramientas para la consecución de dicho objetivo:

- *GREEN*. Es un sistema experto para la identificación interactiva y *on-line* de especímenes biológicos capaz de operar con diferentes grupos taxonómicos.
- *KeyManager* y *XKey*. Son dos herramientas que permiten obtener conjuntos de reglas y claves a partir de ejemplos.
- *KMtoDelta*, *KMtoSDD*, *SDDtoKM*, *DAtoSDD*. Son una serie de utilidades para facilitar el intercambio de información con modelos estándar de representación del conocimiento taxonómico.

Terminamos con las conclusiones sobre la solución propuesta:

1. Se ha mostrado la adecuación de técnicas de Inteligencia Artificial como el aprendizaje automático, los sistemas expertos y el tratamiento de la incertidumbre, al área de la identificación taxonómica. Los resultados experimentales avalan la adecuación de las técnicas utilizadas. Así, para los resultados proporcionados por el sistema experto, se han obtenido unos valores de aceptabilidad del 83% (aceptabilidad sin ponderar) y 89% (aceptabilidad ponderada), y unos niveles de confianza muy aceptables para estos valores. Los tests estadísticos de contraste realizados también confirman que los resultados obtenidos por el sistema no se han debido al azar.

2. Se ha probado que el problema de la identificación taxonómica se puede abordar con éxito desde la perspectiva que ofrecen los sistemas expertos. La capacidad del sistema de razonar hacia delante y hacia atrás permite la consulta para diferentes fines: mientras que el razonamiento hacia delante es una buena estrategia para realizar consultas a ciegas, el razonamiento hacia atrás es adecuado en aquellos casos en que se quiere realizar una búsqueda guiada, por ejemplo, por algún tipo de suposición o hipótesis. Todo esto se ve potenciado con la capacidad de *GREEN* de cambiar de estrategia de razonamiento en una misma sesión de identificación y de seguir simultáneamente varias líneas de razonamiento.

3. El tratamiento de la incertidumbre mediante factores de certeza también ha resultado adecuado. La utilización de esta técnica permite que los usuarios expresen su grado de seguridad en las observaciones realizadas y que el sistema ordene sus resultados en función del nivel de certidumbre que presentan.

4. Para terminar con *GREEN*, señalar que el módulo justificador del sistema también resulta de gran ayuda. Al mostrar la traza del razonamiento seguido, facilita al usuario la toma de decisiones y aumenta su confianza en el sistema. Por otro lado facilita el aprendizaje de los caracteres a observar. Todo esto se ha visto apoyado por el modelo seleccionado para la representación del conocimiento: las reglas presentan una estructura fácilmente comprensible por los usuarios y similar al esquema de las claves de identificación.

5. Se han actualizado los sistemas del Herbario con la capacidad de importar y exportar conjuntos de datos basados en el estándar *SDD*. Las utilidades desarrolladas para este fin facilitan que los investigadores compartan y reutilicen información sobre sus proyectos. En este sentido cabe destacar que *DAtoSDD* es una herramienta muy poderosa al permitir la reutilización de los conjuntos de datos desarrollados no sólo en *DeltaAccess* sino también en *Delta*, el estándar predecesor de *SDD*.

6. Se ha desarrollado JSDD, un paquete Java para el almacenamiento orientado a objetos de documentos *SDD*-XML que actúa de intermediario entre XML y otras aplicaciones, por ejemplo, la herramienta *XKey* hace uso del mismo.

7. Se ha desarrollado *XKey*, una herramienta para generar claves de identificación que opera directamente a partir de conjuntos de datos desarrollados con *SDD*. *XKey* también puede operar con conjuntos de datos de *KeyManager,* y con descripciones en los formatos *DeltaAccess* y *Delta* (mediante el uso de las utilidades de traducción). La salida de la herramienta se presenta en varios formatos: formato texto, formato XML y formato CLIPS y se complementa con información estadística que facilita el estudio y comparación de los resultados obtenidos.

8. *XKey* genera claves de identificación de forma rápida y cómoda, lo que supone un notable ahorro de tiempo para el experto. Es muy versátil, pues permite: seleccionar diferentes criterios de división, configurar el significado de los valores nulos, incluir caracteres diferenciadores y asignar pesos a los caracteres en tiempo de ejecución. Esta funcionalidad añadida produce claves con un significado biológico más adecuado que las generadas con árboles de decisión clásicos. Además de generar claves de identificación, *Xkey* también genera bases de conocimiento completas y consistentes y compatibles con el sistema *GREEN*.

9. Se ha advertido que la adecuación de la clave a la realidad depende en gran medida de los atributos seleccionados para su ramificación. Ante varias

alternativas de ramificación equivalentes, *XKey* detiene su ejecución y recurre al criterio del usuario. Además de la elección del carácter de división, el usuario puede asignar a los caracteres, en tiempo de ejecución, un peso o valor de utilidad que se utilizará para elegir entre varios descriptores en aquellos casos en que haya empate. Este modo de ejecución semi-automático puede ser desactivado para operar de forma totalmente automática.

10. Además de los modos de ejecución automático y semi-automático, se ha dotado a *XKey* de la capacidad de generar claves de forma interactiva. De esta forma, el usuario selecciona en cada momento qué nodo ramificar y qué atributo utilizar; añadir que también es posible eliminar nodos del árbol. Para ayudar al usuario, *XKey* presenta en cada paso los caracteres disponibles ordenados según el criterio de división. Esta opción es de especial relevancia, porque permite combinar la capacidad de discriminación de reglas de división como la entropía con el criterio del experto humano y conseguir que las claves obtenidas tengan un contenido biológico mucho más acertado. En cualquier momento se puede pasar de modo interactivo a modo automático para terminar la generación de la clave.

11. Por último, se ha realizado un estudio comparativo del efecto de varios criterios de división en la generación de claves dicotómicas con un grupo taxonómico real y suficientemente complejo, las Gimnospermas Ibéricas. Este estudio revela que el criterio de división de la entropía es el que produce las claves de menor longitud. Además de esto, también favorece la inclusión de caracteres diferenciadores. El criterio de proporción de ganancia genera claves algo más largas, pero a cambio están más equilibradas (la longitud de los caminos es menos variable). Estos dos criterios detectan la información más relevante para la clasificación de un conjunto de *taxa*. El criterio de Dallwitz no ofrece resultados tan buenos en cuanto a la longitud media de la clave, poder generador de caracteres confirmadores, etc. A pesar de esto, puede resultar de gran utilidad en los casos en que el objetivo no es la minimización de la longitud de la clave. Lo mismo sucede con el índice de diversidad de Gini, con la observación de que este último criterio no es aconsejable con conjuntos de datos grandes porque produce claves demasiado complicadas.

A lo largo de este trabajo hemos observado que los resultados pueden mejorarse y adaptarse mejor a la realidad biológica añadiendo metaconocimiento. Por ejemplo, la asignación de un valor de utilidad para los caracteres permite desempatar en aquellos casos en que el sistema no tiene información para tomar una decisión. Esta alternativa tiene sus inconvenientes: toda esta información debe ser recogida en el modelo de representación del conocimiento, lo que hace su diseño más complicado. A esto hay que añadir la elevada cantidad de caracteres con los que se suele tratar y que todo el metaconocimiento adicional debe ser introducido por el experto que desarrolla el conjunto de datos. En este sentido hay que encontrar un equilibrio entre la cantidad de información adicional y la funcionalidad de las herramientas.

Futuras investigaciones.

Una vez descritos los resultados más relevantes, sugerimos una serie de líneas de trabajo para investigaciones futuras.

- De forma general, podemos decir que un campo de investigación con futuro es la integración de técnicas de Inteligencia Artificial con los métodos de trabajo de los expertos para mejorar las herramientas relacionadas con la Taxonomía. Concretamente, se abre todo un abanico de posibilidades relacionadas con el nuevo modelo de representación de información basado en XML: ontologías, interfaces que se adaptan a las consultas de los usuarios, tecnología móvil, utilización de otros métodos de consulta (medidas de similitud, etc.), estudio del efecto de las dependencias entre caracteres, tratamiento de valores continuos y utilización de otras medidas de incertidumbre.

- Podemos considerar la identificación taxonómica y la generación de claves como aplicaciones particulares de un problema más general: la representación del conocimiento taxonómico. Un proyecto muy interesante es el desarrollo de una plataforma para el tratamiento integral de la información taxonómica que a partir de una misma descripción *SDD* aborde: la obtención de sistemas de

identificación, claves de identificación, descripciones diagnóstico en lenguaje natural, apoyo a la generación de guías de campo, etc. Este trabajo puede completarse con la implementación del paso inverso: la obtención de descripciones estructuradas a partir de descripciones en lenguaje natural. Además de estas funcionalidades básicas, resultaría de utilidad la capacidad de gestionar proyectos, compartir recursos y enlazar con bases de datos gráficas, bibliográficas, glosarios, etc. Así se lograría un completo aprovechamiento de la información taxonómica a partir de las descripciones estructuradas en formato XML.

- En cuanto a la generación de claves dicotómicas, es muy interesante dotar al sistema de la capacidad de comparación inteligente y automática de diversas claves. Para esto es necesario establecer alguna medida de distancia que tenga sentido desde el punto de vista matemático y desde el punto de vista botánico. También sería de interés la utilización de XML para establecer un formato estándar de intercambio de reglas de identificación y claves, ente sistemas.

- Como hemos comentado, existen varias versiones de *SDD* y las herramientas que hemos presentado se basan en la versión 0.5 del estándar *SDD*. Un aspecto que no podemos olvidar es la adaptación de nuestro trabajo a medida que se publiquen nuevas versiones estables del mismo.

Apéndice A. Descripción de los protocolos de estandarización.

En el Capítulo 3, hemos presentado varias herramientas para el intercambio de datos entre los distintos modelos de representación de conocimiento taxonómico. Este apéndice describe la correspondencia entre estos modelos que es implementada por cada herramienta.

1. La herramienta *KMtoSDD*.

1.1. Elemento <Generator>.

Recordemos que <*Generator*> almacena la información relacionada con la aplicación que genera el documento XML. En este caso la aplicación que genera el documento XML es el programa *KMtoSDD*. La Tabla A- I muestra el resumen de los valores asignados a este elemento.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
Application	Sí	Valor por defecto: "KMtoSDD"
Version	Sí	Valor por defecto: "0.5"
LastUpdateDate	Sí	Fecha de la última actualización del programaKMtoSDD.
Authors	No	Valor por defecto: "Eva Lucrecia Gibaja Galindo"
Institution	No	Valor por defecto: "University of Granada"

Tabla A- I. *KMtoSDD: Valores por defecto del elemento <Generator>.*

1.2. Elemento <ProyectDefinition>.

KeyManager está orientado a la generación de conjuntos de reglas, por lo que no recoge información de carácter general relacionada con el conjunto de datos que se describe en el documento. Esta información será suministrada por el usuario a través de la interfaz; en caso contrario se asignarán los valores por defecto que especifica la Tabla A- II.

SUBELEMENTOS	VALOR POR DEFECTO
<Author>	"Author names"
<Editor>	"Editor names"
<FirstPublicationDate>	Fecha del sistema
<LastRevisionDate>	Fecha del sistema
<Version>	"*project version*"
<Title>	"*ProjectTitle*"
<Rights>	"*© año -xxx*"
<ProjectIconUri>	"*http://projectIcon.uri*"
<DefaultAudience>	"en5" (idioma inglés y nivel de conocimientos 5)

Tabla A- II. *KMtoSDD: Valores por defecto de los subelementos de <ProjectDefinition>.*

1.3. Elemento <Descriptions>.

El elemento *<Descriptions>* contiene al menos un **<***ItemDefinition***>** que define el nombre de un taxon. Esta información es almacenada por KeyManager en la tabla de traducción y la correspondencia con *SDD* puede verse en la Tabla A- III.

ATRIBUTOS	OBLIGATORIO EN *SDD*	TRADUCCIÓN
Identification	Sí	Traduccion.nombreAtributo si y solo si Traduccion.isTaxa es "*true*".
Rank	No	Traduccion.valorAtributo si y solo si Traduccion.isTaxa es "*true*".
SUBELEMENTOS	**OBLIGATORIO EN *SDD***	**TRADUCCIÓN**
<*Identification*>	Sí	Traduccion.nombreAtributo si y solo si Traduccion.isTaxa es "*true*".

Tabla A- III. *KMtoSDD: Traducción de los atributos y subelementos de <Descriptions>.*

ELEMENTO <CODED DESCRIPTION>.

En *KeyManager* la descripción de los *taxa* tiene forma tabular: cada fila corresponde a un taxon y las columnas corresponden a los atributos que los describen. En *SDD* <*Character*> tiene un atributo obligatorio, *keyref* cuyo valor se corresponde con el nombre de la columna en cuestión.

La aparición de un <*Character*>, implica que habrá al menos un estado, <*State*>, de ese carácter. El elemento <*State*> tiene también un atributo *keyref* cuyo valor tomamos del contenido de la columna para el ítem que se está describiendo.

Especie	Tamaño de las hojas (largo)	Hojas escurridas sobre el raquis	Forma de las hojuelas
Cycas revoluta	Hasta 100 cm	No	Con el borde revuelto
Cycas circinalis	Mas de 100 cm	Sí	Acintadas y planas

Tabla A- IV. *KMtoSDD: Descripción de las especies del género Cycas.*

La Figura A- Imuestra la traducción del taxon "*Cycas circinalis*" descrito en la Tabla A- IV.

```
<Item Identification="Cycas circinalis" Rank="Especie">
  - <ItemDefinition>
      <Identification>Cycas circinalis</Identification>
    </ItemDefinition>
  - <CodedDescription>
    - <Character keyref="Tamaño de las hojas (largo)">
        <State keyref="Mas de 100 cm" />
      </Character>
    - <Character keyref="Hojas escurridas sobre el raquis">
        <State keyref="Si" />
      </Character>
    -<Character keyref="Forma de las hojuelas">
        <State keyref="Acintadas y planas" />
      </Character>
  </CodedDescription>
</Item>
```

Figura A- I. *KMtoSDD: Traducción del taxon "Cycas circinalis".*

1.4. Elemento <Terminology>.

El elemento <Terminology> contiene siete subelementos obligatorios: <*AudienceDefintions*>, <*GlobalMeasureDefinitions*>, <*GlobalStateDefinitions*>, <*FrequencyDefinitions*>, <*ModifierDefinitions*>, <*CharacterDefinitions*> y <*CharacterGroupDefinitions*> cuya correspondencia con KeyManager se detalla a continuación.

ELEMENTO <AUDIENCEDEFINITIONS>.

<*AudienceDefinition*> describe el tipo de público por defecto. *KeyManager* no incluye este tipo de información, por lo que hemos incluido de forma automática el un conjunto de <*AudienceDefinition*> más habitual.

ELEMENTO <GLOBALMEASUREDEFINITIONS>.

Este elemento modeliza medidas de carácter global predefinidas por el estándar, por lo que se incluyen de forma automática en la traducción.

ELEMENTO <GLOBALSTATEDEFINITIONS>.

El elemento <*GlobalStateDefinitions*> define tres tipos de estados globales:

1. Estados globales especiales (*SpecialStates*).
2. Estados globales de cálculo (*ComputedSpecialStates*).
3. Estados globales definidos por el usuario.

Los dos primeros son establecidos por defecto por el estándar, por lo que se incluyen automáticamente en la traducción. *KeyManager* considera locales todos los estados definidos por el usuario, pero por un error de diseño, *SDD* fuerza la aparición de un conjunto de estados globales definidos por el usuario. Por este motivo, se han incluido un conjunto definido por defecto.

ELEMENTOS <FRECUENCYDEFINITIONS> Y <MODIFIERDEFINITIONS>.

SDD distingue dos tipos de modificadores, los modificadores de frecuencia y los que no lo son. La aparición de la definición de modificadores de frecuencia es de carácter obligatorio en *SDD*, por lo que hemos incluido en la traducción un conjunto de modificadores por defecto.

ELEMENTO <CHARACTERDEFINITIONS>.

Este elemento contiene la definición de todos los caracteres que pueden aparecer en la descripción de un taxon en el documento. Consta de uno o más <*CharacterDefinition*>.

El elemento **<***CharacterDefinition***>** tiene un atributo obligatorio, *key*, cuyo valor se corresponde con el valor del campo "*TRADUCCION.nombreAtributo*" donde "*TRADUCCION.isTaxa*" tiene el valor "*false*". En la traducción se han incluído sus dos hijos obligatorios:

- <*LinguisticSets*> Almacena el nombre del carácter y su valor lo tomamos del campo "*TRADUCCION.nombreAtributo*".
- <*DescriptorDefinitions*> define todos los posibles estados, <*LocalStateDefinitions*>, para un determinado carácter. Cada uno de estos estados será representado por uno o varios elementos <*StateDefinition*>.
- <*StateDefinition*> tiene un atributo *key*, y una etiqueta <*Label*> cuyo valor coincide y es el del campo "*TRADUCCION.valorAtributo*".

ELEMENTO <CHARACTERGROUPDEFINITIONS>.

SDD no permite representar jerarquías de elementos más que utilizando grupos de caracteres. Por esto utilizamos el elemento <*CharacterGroupDefinition*> para exportar a *SDD* la jerarquía de *taxa* que se representa en el modelo de *KeyManager*.

Para esto, dentro del elemento <*CharacterGroupDefinitions*> se incluirá un elemento <*CharacterGroupDefinition*> donde el atributo obligatorio, *key*, tiene el valor "*taxaHierarchy*" y el atributo obligatorio, *type*, tiene el valor "*CharacterArray*". Cada rango taxonómico será representado por un elemento <*CharacterGroupItem*>.

2. La herramienta *SDDtoKM*.

2.1. Reconstruir la tabla de información.

Esta tabla contiene dos tipos de información: información general sobre una tabla de datos (campos nombre de la tabla, número de columnas, etc) e información sobre la jerarquía taxonómica representada en la base de datos (campos *ordenJerarquía*, *nombreAtributoPadre*, *valorAtributoPadre*, etc.). La Tabla A- V recoge el esquema para reconstruir la tabla de información.

CAMPOS	TRADUCCIÓN
nombreTabla	Se obtiene del atributo *key* del elemento <*CharacterGroupItem*>
nombreAtributo	Se obtiene del atributo *key* del elemento <*CharacterGroupItem*>
valorAtributo	Se obtiene del atributo *key* del elemento <*CharacterGroupItem*>
ordenJerarquia	Campo calculado.
numeroColumnas	Campo calculado.
tablaFinal	Campo calculado.
nombreAtributoPadre	Se obtiene del padre del elemento <*CharacterGroupItem*>
valorAtributoPadre	Se obtiene del padre del elemento <*CharacterGroupItem*>

Tabla A- V. *SDDtoKM: Reconstrucción de la tabla "Información".*

Como se indicó en la página 224 , al hacer la traducción de *KeyManager* a *SDD*, la jerarquía taxonómica queda representada en la sección <*CharacterGroupItems*>. Esta información debe ser recuperada justamente en sentido inverso. En caso de no existir, se considera que todos los elementos <*Item*> descritos dentro del documento pertenecen al mismo rango taxonómico. El Figura A- II presenta la jerarquía taxonómica de los géneros y especies de la familia *Cupressaceae* en formato *SDD*. La Tabla A- VI corresponde a la reconstrucción de la tabla de información para los datos de dicho ejemplo.

NOMBRETABLA	NOMBRE ATRIBUTO	VALOR ATRIBUTO	ORDEN JERARQUIA	NUMERO COLUMNAS	TABLA FINAL	NOMBRE ATRIBUTOPADRE	VALOR ATRIBUTO PADRE
DivisionGimnospermas			0	16	No		
FamiliaCupressaceae	Familia	Cupressaceae	1	11	No		
GeneroCalocedrus	Genero	Calocedrus	2	1	Si	Familia	Cupressaceae
GeneroChamaecyparis	Genero	Chamaecyparis	2	1	Si	Familia	Cupressaceae
GeneroCupressus	Genero	Cupressus	2	9	Si	Familia	Cupressaceae
GeneroJuniperus	Genero	Juniperus	2	20	Si	Familia	Cupressaceae
GeneroPlatycladus	Genero	Platycladus	2	1	Si	Familia	Cupressaceae
GeneroTetraclinis	Genero	Tetraclinis	2	1	Si	Familia	Cupressaceae

Tabla A- VI. ***SDDtoKM:*** *Ejemplo de tabla de información.*

```
- <CharacterGroupDefinitions>
  - <CharacterGroupDefinition key="TaxaHierarchy" type="PartHierarchy">
    - <LinguisticSets>
      - <LinguisticSet keyref="en1">
          <Label />
        </LinguisticSet>
      </LinguisticSets>
    - <CharacterGroupItem key="Division Gimnospermas">
      - <CharacterGroupItem key="Familia Cupressaceae">
        - <CharacterGroupItem key="Genero Calocedrus">
            <CharacterGroupItem key="Especie Calocedrus decurrens" />
          </CharacterGroupItem>
        - <CharacterGroupItem key="Genero Chamaecyparis">
            <CharacterGroupItem key="Especie Chamaecyparis lawsoniana" />
          </CharacterGroupItem>
        - <CharacterGroupItem key="Genero Cupressus">
            <CharacterGroupItem key="Especie Cupressus arizonica" />
            <CharacterGroupItem key="Especie Cupressus lusitanica" />
            <CharacterGroupItem key="Especie Cupressus macrocarpa" />
            <CharacterGroupItem key="Especie Cupressus sempervirens" />
          </CharacterGroupItem>
        - <CharacterGroupItem key="Genero Juniperus">
            <CharacterGroupItem key="Especie Juniperus communis subsp. alpina" />
            <CharacterGroupItem key="Especie Juniperus communis subsp. communis" />
            <CharacterGroupItem key="Especie Juniperus communis subsp. hemisphaerica" />
            <CharacterGroupItem key="Especie Juniperus navicularis" />
            <CharacterGroupItem key="Especie Juniperus oxycedrus subsp. badia" />
            <CharacterGroupItem key="Especie Juniperus oxycedrus subsp. macrocarpa" />
            <CharacterGroupItem key="Especie Juniperus oxycedrus subsp. oxycedrus" />
            <CharacterGroupItem key="Especie Juniperus phoenicea subsp. phoenicea" />
            <CharacterGroupItem key="Especie Juniperus phoenicea subsp. turbinata" />
            <CharacterGroupItem key="Especie Juniperus sabina" />
            <CharacterGroupItem key="Especie Juniperus thurifera" />
          </CharacterGroupItem>
        - <CharacterGroupItem key="Genero Platycladus">
            <CharacterGroupItem key="Especie Platycladus orientalis" />
          </CharacterGroupItem>
        - <CharacterGroupItem key="Genero Tetraclinis">
            <CharacterGroupItem key="Especie Tetraclinis articulata" />
          </CharacterGroupItem>
        </CharacterGroupItem>
        </CharacterGroupItem>
      </CharacterGroupItem>
    </CharacterGroupDefinition>
  </CharacterGroupDefinitions>
```

Figura A- II. *SDDtoKM:* *Jerarquía taxonómica de los géneros y especies de la familia Cupressaceae en formato SDD.*

2.2. Reconstruir la tabla de traducción.

KeyManager almacena la tabla de traducción los nombres que se consideran válidos. Estos nombres se refieren a:

- Nombres de caracteres y sus posibles valores. En un documento *SDD*, esta información la encontramos en los elementos <*CharacterDefinition*>, <*LocalStateDefinition*> y <*LocalStateSetReference*>.
- Nombres de objetivos y sus posibles valores. En un documento *SDD*, esta información la encontramos en el elemento <*Item*>.

NOMBRES Y VALORES DE LOS CARACTERES

Un carácter está definido por el nombre del carácter y el conjunto de valores permitidos para este. En *SDD* encontramos esta información en el elemento <*CharacterDefinition*>, concretamente el nombre del carácter se toma del subelemento <*Label*>. En la Figura A- IV, el nombre del carácter es "*stipule color*".

Los valores permitidos para un carácter se definen mediante:

- Estados locales <*LocalStateDefinition*>. En este caso se tomará como valor permitido el contenido dentro de la etiqueta <*Label*>. En la Figura A- IV para el carácter "*stipule colo*r" se define el estado local "*black*".
- Estados globales <*GlobalStateSetReference*>. Este elemento contiene una referencia, *keyref*, a un estado global. El valor del estado es el correspondiente a la etiqueta <*Label*> del estado global. En la Figura A- IV, se definen dos estados para el atributo "*stipule color*" que se basan en el grupo de estados globales "*color*".

```
<GlobalStateDefinitionSet key="color">
  - <SetLabelLinguisticSets keyref="en5">
      <Label>standard color states</Label>
    </SetLabelLinguisticSets>-
  - <StateDefinition key="white">
    - <LinguisticSets>
      - <LinguisticSet keyref="en5">
          <Label>white</Label>
        </LinguisticSet>
      </LinguisticSets>
    </StateDefinition>
</GlobalStateDefinitionSet>
```

Figura A- III. *SDDtoKM: **Ejemplo** de estado global.*

En algunos casos, podemos considerar como valores de estados globales, gamas de colores, los valores genéricos "desconocido" (*unknown*), "no aplicable" (*notApplicable*), "vacío" (*empty*). Es posible que la descripción de los valores que puede tomar un determinado carácter incluya alguna referencia a estos estados de carácter global. La posibilidad de definir estados o valores de carácter global, está directamente relacionada con la flexibilidad para definir la terminología válida en el documento. No obstante, al abordar la descripción de un taxon particular, los valores de un carácter (sean globales o no) se asocian de forma local a la

descripción de un individuo. Por este motivo hemos de considerar que todos los estados asociados a los caracteres son de carácter local.

```
- <CharacterDefinition key="stipule-color" type="ordinal-interval">
  - <LinguisticSets>
    - <LinguisticSet keyref="en5">
        <Label>stipule color</Label>
      </LinguisticSet>
    </LinguisticSets>
  -<DescriptorDefinitions>
    - <GlobalStateSetReference keyref="color">
      - <Selections>
          <DescriptorSelection keyref="white" />
        </Selections>
      </GlobalStateSetReference>
    - <LocalStateDefinitions>
      - <StateDefinition key="black">
        - <LinguisticSets>
          - <LinguisticSet keyref="en5">
              <Label>black</Label>
            </LinguisticSet>-
          </LinguisticSets>
        </StateDefinition>
      </LocalStateDefinitions>
    </DescriptorDefinitions>
  </CharacterDefinition>
```

Figura A- IV. *Carácter con estados globales.*

NOMBRES Y VALORES DE LOS OBJETIVOS.

En *KeyManager*, el nombre de un objetivo está compuesto por el rango taxonómico al que pertenece y el nombre científico del objetivo (por ejemplo, "especie *Abies pinsapo*"). En *SDD*, esta información se describe en el elemento <*Item*>. El valor del rango taxonómico se obtiene del atributo "*Rank*". Es un atributo opcional, de modo que si no aparece, se considera por defecto que el valor del rango taxonómico es la cadena de caracteres "*Rank*". El nombre científico del objetivo se toma del atributo obligatorio"*Identification*".

En la Figura A- V el rango taxonómico es "*species*" (especie) y el nombre del objetivo es "*Discaria pubescens*".

```
<Item Identification="Discaria pubescens" Rank="species">
```

Figura A- V. *SDDtoKM:* *Ejemplo de identificación de un ítem.*

RECONSTRUIR LAS TABLAS DE DATOS.

Una descripción en *SDD* se corresponde con un elemento <*Item*>. A media que se leen los elementos del documento XML, su información es almacenada en una estructura de datos en memoria que refleja propiedades generales del taxon o individuo (identificación, rango taxonómico) y una descripción formada por un vector de caracteres y sus respectivos atributos.

La estructura de la información en *SDD* guarda más relación con el concepto de clase y objeto, mientras que *KeyManager* almacena la información sobre los objetivos en forma tabular. Necesitamos hacer una transformación entre estas dos estructuras.

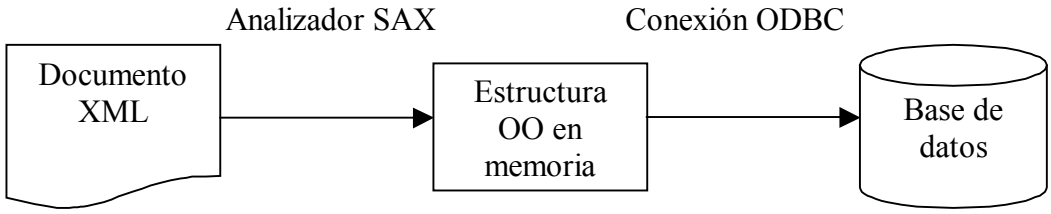


Figura A- VI. *SDDtoKM: Esquema de la transformación de una estructura XML a una estructura relacional.*

Una vez que se tiene en memoria el modelo orientado a objetos del documento, dicho modelo se escribe en una base de datos con la estructura requerida por el programa *keyManager*. La Tabla A- VIII ilustra esta idea:

La Tabla A- VII contiene una breve descripción de la familia *Taxaceae* en forma tabular, mientras que la Tabla A- VIII contiene la misma descripción en formato *SDD*. Esta familia admite individuos tanto con aspecto de árbol como de arbusto. Esta información se refleja en la tabla en la aparición de dos tuplas (una por cada posible combinación de valores), mientras que *SDD* almacena esta información en una misma descripción en la que para el atributo "*aspecto*" incluye dos estados: "*Árbol*" y "*Arbusto*".

FAMILIA	ASPECTO	RESINOSA	PERSISTENCIA DE LA HOJA	SEMILLA CON ARILO
Taxaceae	Arbol	No	Persistente	Si
Taxaceae	Arbusto	No	Persistente	Si

Tabla A- VII. *SDDtoKM: Descripción tabular de la familia Taxaceae..*

```
<Item Identification="Taxaceae" Rank="Familia">
  - <ItemDefinition>
      <Identification>Taxaceae</Identification>
    </ItemDefinition>
  - <CodedDescription>
    - <Character keyref="Aspecto">
        <State keyref="Arbol" />
        <State keyref="Arbusto" />
      </Character>
    - <Character keyref="Resinosa">
        <State keyref="No" />
      </Character>
    - <Character keyref="Persistencia de la hoja">
        <State keyref="Persistente" />
      </Character>
    - <Character keyref="Semilla con arilo">
        <State keyref="Si" />
      </Character>
  </CodedDescription>
</Item>
```

Tabla A- VIII. *.SDDtoKM: Descripción* XML *de la familia Taxaceae*

3. La herramienta *KMtoDelta*.

El sistema *Delta* trabaja con tres tipos de documentos:

- Documento de especificaciones: Contiene información de tipo general sobre el conjunto de datos.
- Documento de caracteres: Donde se describen los caracteres válidos y sus valores.
- Documento de ítem: Contiene las descripciones de los ítem en función de los caracteres antes citados.

A continuación describimos el protocolo de construcción de estos tres ficheros a partir de la información disponible en un conjunto de datos de *KeyManager*.

```
*SHOW ~ Dataset specifications.
*DATA BUFFER SIZE 4000
*NUMBER OF CHARACTERS 88
*MAXIMUM NUMBER OF STATES 47
*MAXIMUM NUMBER OF ITEMS 9
*NUMBERS OF STATES 1,3 2,2 3,2 4,2 5,2 6,13 7,2 8,3 9,2 10,2 11,3 12,13 13,2 14,2 15,2
16,2 17,2 18,2 19,2 20,2 21,2 22,2 23,2 24,2 25,2 26,2 27,3 28,3 29,2 30,2 31,2 32,2 33,2 34,2
35,2 36,2 37,2 38,2 39,2 40,3 41,2 42,2 43,2 44,2 45,2 46,2 47,2 48,3 49,5 50,2 51,2 52,3 53,3
54,2 55,2 56,3 57,2 58,2 59,3 60,9 61,3 62,2 63,3 64,2 65,2 66,2 67,2 68,2 69,2 70,2 71,2 72,2
73,7 74,2 75,2 76,2 77,2 78,3 79,2 80,4 81,2 82,4 83,2 84,2 85,2 86,9 87,18 88,47
```

Figura A- VII. *KMtoDelta. Documento de ítem.*

3.1. Documento de especificaciones.

La información necesaria para este documento puede ser calculada directamente de la información contenida tabla de *traducción* de la base de datos. La Tabla A- IX detalla las directivas *Delta* incluidas en el fichero de especificaciones.

DIRECTIVA *DELTA*
NUMBER OF CHARACTERS (número de caracteres del conjunto de datos)
MAXIMUM NUMBER OF STATES (número máximo de estados del conjunto de datos)
MAXIMUM NUMBER OF ITEMS (número máximo de ítems)
NUMBERS OF STATES (para cada carácter, su número de estados)

Tabla A- IX. *KMtoDelta: Algunas directivas Delta.*

El ejemplo de fichero de especificaciones para las familias de la división Gimnospermas es el siguiente:

3.2. Documento de caracteres.

Describe los caracteres y atributos que se consideran válidos en las descripciones. Nuevamente esta información se obtiene de la tabla de traducción.

*CHARACTER LIST
#1. Aspecto/
1. Arbol/
2. Arbusto/
3. Palmera/
#2. Tipo de arbusto/
1. Erecto/
2. Postrado/
#3. Tallo articulado/
1. Si/
2. No/
#4. Forma de la copa/
1. Aparasolada/
2. Piramidal/

Figura A- VIII. *KMtoDelta. Documento de caracteres.*

3.3. Documento de descripciones.

Se obtiene de la tabla de datos correspondiente.

*ITEM DESCRIPTIONS

#Familia Araucariaceae/ 1,1 10,1 17,1 29,2 15,2 16,2 18,1 37,2 20,2 3,2 51,1 54,2 81,2 55,1
#Familia Cephalotaxaceae/ 1,1 10,2 17,2 29,1 15,2 16,2 18,1 37,2 20,2 3,2 51,2 54,2 81,2 55,2
#Familia Cupressaceae/ 1,2/1 10,1 17,2 29,2 15,1 16,1 18,2 28,1/3 37,2 20,2 3,2 51,2/1 54,2 81,1/2 55,1/2
#Familia Cycadaceae/ 1,3 10,2 17,2 29,2 15,1 16,2 18,2 37,2 20,2 3,2 51,2 54,2 81,1 55,2
#Familia Ephedraceae/ 1,2 10,2 17,2 29,2 15,1 16,1 18,2 28,3 37,2 20,1 3,1 51,2 54,2 81,2 55,2
#Familia Ginkgoaceae/ 1,1 10,2 17,2 29,2 15,2 16,2 18,1 28,2 37,1 20,2 3,2 51,2 54,2 81,2 55,2
#Familia Pinaceae/ 1,1 10,1 17,2 29,1 15,2 16,2 18,1 28,1 37,1/2 20,2 3,2 51,1 54,2 81,1 55,1
#Familia Taxaceae/ 1,1/2 10,2 17,2 29,1 15,2 16,2 18,1 37,2 20,2 3,2 51,2 54,1 81,2 55,2
#Familia Taxodiaceae/ 1,1 10,1 17,2/1 29,2/1 15,2 16,2 18,1 28,1 37,1/2 20,2 3,2 51,1 54,2 81,1 55,1

Figura A- IX. *KMtoDelta. Documento de descripciones.*

Apéndice B. Claves para la División Gimnospermas generadas con XKey.

DIVISIÓN GYMNOSPERMAE.

DIVISIÓN GYMNOSPERMAE/ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Semilla con arilo: 1 vez* *Planta: 1 vez* *Forma de la hoja: 2 veces* *Disposición de las hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Resinosa: 1 vez* *Consistencia de la fructificación: 1 vez* *Semillas numerosas: 1 vez*
(0) Disposición de las hojas En espiral.....1 (1) Planta Dioica; Consistencia de la fructificación Carnosa; Semillas numerosas No.....2 (2) Semilla con arilo Si; Resinosa No .....Familia *Taxaceae* (2) Semilla con arilo No; Resinosa Si .....Familia *Cephalotaxaceae* (1) Planta Monoica; Consistencia de la fructificación Leñosa; Semillas numerosas Si.....3 (3) Forma de la hoja Linear (planoaguzada) .....Familia *Taxodiaceae* (3) Forma de la hoja Acicular .....Familia *Pinaceae* (3) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (0) Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (0) Disposición de las hojas Ternadas .....Familia *Cupressaceae* (0) Disposición de las hojas Imbricadas.....4	

DIVISIÓN GYMNOSPERMAE/ENTROPÍA
(4) Forma de la hoja Escamosa .....Familia *Cupressaceae* (4) Forma de la hoja Escamosa curvada hacia el ápice .....Familia *Araucariaceae* (0) Disposición de las hojas En verticilos de cuatro .....Familia *Cupressaceae* (0) Disposición de las hojas En los nudos del tallo .....Familia *Ephedraceae* (0) Disposición de las hojas Hojas agrupadas en la terminación de brotes laterales (braquiblastos) .....Familia *Ginkgoaceae* (0) Disposición de las hojas Formando una corona apical .....Familia *Cycadaceae*

Tabla B- I. *Clave para la división Gymnospermae generada según el criterio de mínima entropía.*

DIVISIÓN GYMNOSPERMAE /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:1*
Atributos incluidos en la clave: *Forma de la hoja: 3 veces* *Resinosa: 1 vez* *Planta: 1 vez* *Semillas numerosas: 1 vez*	*Caracteres confirmadores incluidos:* *Disposición de las hojas: 1 vez*
(0) Resinosa No.....1 (1) Forma de la hoja Linear (planoaguzada); Disposición de las hojas En espiral .....Familia *Taxaceae* (1) Forma de la hoja Escamosa; Disposición de las hojas En los nudos del tallo .....Familia *Ephedraceae* (1) Forma de la hoja En forma de abanico, con escotadura central; Disposición de las hojas Hojas agrupadas en la terminación de brotes laterales (braquiblastos) .....Familia *Ginkgoaceae* (1) Forma de la hoja En forma de palmera; Disposición de las hojas Formando una corona apical .....Familia *Cycadaceae* (0) Resinosa Si.....2 (2) Semillas numerosas Si.....3 (3) Planta Monoica.....4 (4) Forma de la hoja Acicular .....Familia *Pinaceae* (4) Forma de la hoja Linear (planoaguzada) .....Familia *Taxodiaceae* (4) Forma de la hoja Escamosa .....Familia *Cupressaceae* (4) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (3) Planta Dioica .....Familia *Araucariaceae* (2) Semillas numerosas No.....5 (5) Forma de la hoja Acicular .....Familia *Cupressaceae* (5) Forma de la hoja Escamosa .....Familia *Cupressaceae* (5) Forma de la hoja Linear (planoaguzada) .....Familia *Cephalotaxaceae*	

Tabla B- II. *Clave para la división Gymnospermae generada según el criterio de proporción de ganancia.*

DIVISIÓN GYMNOSPERMAE /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 9	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Forma de la hoja: 4 veces* *Tallo articulado: 1 vez* *Hoja: 3 veces* *Aspecto: 2 veces* *Semilla con arilo: 1 vez* *Planta: 2 veces* *Consistencia de la fructificación: 1 vez* *Disposición de las hojas: 1 vez* *Semillas numerosas: 1 vez*	*Caracteres confirmadores incluidos:* *Disposición de las hojas: 3 veces* *Resinosa: 1 vez*

(0) Consistencia de la fructificación Carnosa.....1
(1) Planta Dioica.....2
(2) Aspecto Arbusto.....3
(3) Hoja Persistente.....4
(4) Tallo articulado No.....5
(5) Forma de la hoja Linear (planoaguzada); Disposición de las hojas En espiral
.....Familia *Taxaceae*
(5) Forma de la hoja Acicular; Disposición de las hojas Ternadas
.....Familia *Cupressaceae*
(5) Forma de la hoja Escamosa; Disposición de las hojas Imbricadas
.....Familia *Cupressaceae*
(4) Tallo articulado Si
.....Familia *Ephedraceae*
(3) Hoja Más o menos caediza
.....Familia *Ephedraceae*
(2) Aspecto Árbol.....6
(6) Hoja Persistente.....7
(7) Semilla con arilo No; Resinosa Si.....8
(8) Forma de la hoja Escamosa; Disposición de las hojas Imbricadas
.....Familia *Cupressaceae*
(8) Forma de la hoja Linear (planoaguzada); Disposición de las hojas En espiral
.....Familia *Cephalotaxaceae*
(8) Forma de la hoja Acicular; Disposición de las hojas Ternadas
.....Familia *Cupressaceae*
(7) Semilla con arilo Si; Resinosa No
.....Familia *Taxaceae*
(6) Hoja Caduca
.....Familia *Ginkgoaceae*
(2) Aspecto Palmera
.....Familia *Cycadaceae*
(1) Planta Monoica
.....Familia *Cupressaceae*
(0) Consistencia de la fructificación Leñosa.....9
(9) Aspecto Árbol.....10
(10) Hoja Persistente.....11
(11) Planta Monoica.....12
(12) Semillas numerosas Si.....13
(13) Disposición de las hojas Fasciculadas
.....Familia *Pinaceae*
(13) Disposición de las hojas En espiral.....14
(14) Forma de la hoja Linear (planoaguzada)
.....Familia *Taxodiaceae*
(14) Forma de la hoja Acicular
.....Familia *Pinaceae*
(14) Forma de la hoja Aleznada
.....Familia *Taxodiaceae*
(13) Disposición de las hojas Imbricadas
.....Familia *Cupressaceae*
(12) Semillas numerosas No

DIVISIÓN GYMNOSPERMAE /INDICE DE DIVERSIDAD DE GINI
.....Familia *Cupressaceae* (11) Planta Dioica .....Familia *Araucariaceae* (10) Hoja Caduca.....15 (15) Forma de la hoja Acicular; Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (15) Forma de la hoja Linear (planoaguzada); Disposición de las hojas En espiral .....Familia *Taxodiaceae* (9) Aspecto Arbusto .....Familia *Cupressaceae*

Tabla B- III. *Clave para la división Gymnospermae generada según el criterio de Gini.*

DIVISIÓN GYMNOSPERMAE /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave: 6	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Semilla con arilo: 1 vez* *Planta: 1 vez* *Hoja: 2 veces* *Forma de la hoja: 1 vez* *Disposición de las hojas: 1 vez* *Tallo articulado: 1 vez*	*Caracteres confirmadores incluidos:* *Resinosa: 2 veces* *Consistencia de la fructificación: 1 vez* *Semillas numerosas: 1 vez*
(0) Forma de la hoja Linear (planoaguzada).....1 (1) Hoja Persistente.....2 (2) Planta Dioica; Consistencia de la fructificación Carnosa; Semillas numerosas No.....3 (3) Semilla con arilo Si; Resinosa No .....Familia *Taxaceae* (3) Semilla con arilo No; Resinosa Si .....Familia *Cephalotaxaceae* (2) Planta Monoica; Consistencia de la fructificación Leñosa; Semillas numerosas Si .....Familia *Taxodiaceae* (1) Hoja Caduca .....Familia *Taxodiaceae* (0) Forma de la hoja Acicular.....4 (4) Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (4) Disposición de las hojas Ternadas .....Familia *Cupressaceae* (4) Disposición de las hojas En espiral .....Familia *Pinaceae* (0) Forma de la hoja Escamosa.....5 (5) Hoja Persistente.....6 (6) Tallo articulado No; Resinosa Si .....Familia *Cupressaceae* (6) Tallo articulado Si; Resinosa No .....Familia *Ephedraceae* (5) Hoja Más o menos caediza .....Familia *Ephedraceae* (0) Forma de la hoja En forma de abanico, con escotadura central .....Familia *Ginkgoaceae* (0) Forma de la hoja En forma de palmera .....Familia *Cycadaceae* (0) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (0) Forma de la hoja Escamosa curvada hacia el ápice .....Familia *Araucariaceae*	

Tabla B- IV. *Clave para la división Gymnospermae generada según el criterio de Dallwitz.*

DIVISIÓN GYMNOSPERMAE /CRITERIO EXPERTO	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Semilla con arilo: 1 vez* *Planta: 1 vez* *Forma de la hoja: 1 vez* *Disposición de las hojas: 2 veces*	*Caracteres confirmadores incluidos:* *Resinosa: 1 vez* *Consistencia de la fructificación: 1 vez* *Semillas numerosas: 1 vez*
(0) Forma de la hoja Linear (planoaguzada).....1 (1) Planta Dioica; Consistencia de la fructificación Carnosa; Semillas numerosas No.....2 (2) Semilla con arilo Si; Resinosa No .....Familia *Taxaceae* (2) Semilla con arilo No; Resinosa Si .....Familia *Cephalotaxaceae* (1) Planta Monoica; Consistencia de la fructificación Leñosa; Semillas numerosas Si .....Familia *Taxodiaceae* (0) Forma de la hoja Acicular.....3 (3) Disposición de las hojas Fasciculadas .....Familia *Pinaceae* (3) Disposición de las hojas Ternadas .....Familia *Cupressaceae* (3) Disposición de las hojas En espiral .....Familia *Pinaceae* (0) Forma de la hoja Escamosa.....4 (4) Disposición de las hojas Imbricadas .....Familia *Cupressaceae* (4) Disposición de las hojas En verticilos de cuatro .....Familia *Cupressaceae* (4) Disposición de las hojas En los nudos del tallo .....Familia *Ephedraceae* (0) Forma de la hoja En forma de abanico, con escotadura central .....Familia *Ginkgoaceae* (0) Forma de la hoja En forma de palmera .....Familia *Cycadaceae* (0) Forma de la hoja Aleznada .....Familia *Taxodiaceae* (0) Forma de la hoja Escamosa curvada hacia el ápice .....Familia *Araucariaceae*	

Tabla B- V. *Clave para la división Gymnospermae generada según el criterio del experto.*

FAMILIA CUPRESSACEAE.

FAMILIA *CUPRESSACEAE* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Aspecto: 1 vez* *Tamaño de la piña (largo): 1 vez* *Número de escamas en la piña: 1 vez* *Consistencia de la fructificación: 1 vez*	*Caracteres confirmadores incluidos:* *Escamas de la piña peltadas: 2 veces* *Escamas de la piña planas: 2 veces* *Disposición de las ramillas en un solo plano: 2 veces*
(0) Consistencia de la fructificación Leñosa.....1 (1) Tamaño de la piña (largo) Entre 1 y 2 cm.....2 (2) Aspecto Árbol; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano No	

FAMILIA *CUPRESSACEAE* /ENTROPÍA
.....Género *Cupressus* (2) Aspecto Arbusto; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano Si .....Género *Platycladus* (1) Tamaño de la piña (largo) Hasta 1 cm.....3 (3) Número de escamas en la piña Entre 6 y 8; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano Si .....Género *Chamaecyparis* (3) Número de escamas en la piña Hasta 4; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano No .....Género *Tetraclinis* (1) Tamaño de la piña (largo) Entre 2 y 4 cm .....Género *Calocedrus* (0) Consistencia de la fructificación Carnosa .....Género *Juniperus*

Tabla B- VI. *Clave para la familia Cupressaceae generada según el criterio de mínima entropía.*

FAMILIA *CUPRESSACEAE* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Aspecto: 1 vez* *Tamaño de la piña (largo): 1 vez* *Número de escamas en la piña: 1 vez* *Consistencia de la fructificación: 1 vez*	*Caracteres confirmadores incluidos:* *Escamas de la piña peltadas: 2 veces* *Escamas de la piña planas: 2 veces* *Disposición de las ramillas en un solo plano: 2 veces*
(0) Consistencia de la fructificación Leñosa.....1 (1) Tamaño de la piña (largo) Entre 1 y 2 cm.....2 (2) Aspecto Árbol; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano No .....Género *Cupressus* (2) Aspecto Arbusto; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano Si .....Género *Platycladus* (1) Tamaño de la piña (largo) Hasta 1 cm.....3 (3) Número de escamas en la piña Entre 6 y 8; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano Si .....Género *Chamaecyparis* (3) Número de escamas en la piña Hasta 4; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano No .....Género *Tetraclinis* (1) Tamaño de la piña (largo) Entre 2 y 4 cm .....Género *Calocedrus* (0) Consistencia de la fructificación Carnosa .....Género *Juniperus*	

Tabla B- VII. *Clave para la familia Cupressaceae generada según el criterio de proporción de ganancia*

FAMILIA *CUPRESSACEAE* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 7	*Total de caracteres confirmadores:4*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 1 vez* *Número de escamas en la piña: 3 veces* *Escamas de la piña peltadas: 1 vez* *Consistencia de la fructificación: 2 veces* *Forma de la hoja: 2 veces* *Planta: 2 veces* *Aspecto: 1 vez*	*Caracteres confirmadores incluidos:* *Presencia de apófisis: 2 veces* *Disposición de las ramillas en un solo plano: 3 veces* *Escamas de la piña planas: 1 vez* *Tamaño de la piña (largo): 2 veces*

(0) Aspecto Árbol.....1
(1) Planta Monoica.....2
(2) Forma de la hoja Escamosa.....3
(3) Consistencia de la fructificación Leñosa.....4
(4) Escamas de la piña peltadas Si; Escamas de la piña planas No.....5
(5) Número de escamas en la piña Entre 6 y 8.....6
(6) Tamaño de la piña (largo) Entre 1 y 2 cm; Presencia de apófisis Si; Disposición de las ramillas en un solo plano No
.....Género *Cupressus*
(6) Tamaño de la piña (largo) Hasta 1 cm; Presencia de apófisis No; Disposición de las ramillas en un solo plano Si
.....Género *Chamaecyparis*
(5) Número de escamas en la piña Mas de 8
.....Género *Cupressus*
(4) Escamas de la piña peltadas No; Escamas de la piña planas Si.....7
(7) Número de escamas en la piña Hasta 4; Tamaño de la piña (largo) Hasta 1 cm; Disposición de las ramillas en un solo plano No
.....Género *Tetraclinis*
(7) Número de escamas en la piña Entre 6 y 8; Tamaño de la piña (largo) Entre 2 y 4 cm; Disposición de las ramillas en un solo plano Si
.....Género *Calocedrus*
(3) Consistencia de la fructificación Carnosa
.....Género *Juniperus*
(2) Forma de la hoja Acicular
.....Género *Juniperus*
(1) Planta Dioica
.....Género *Juniperus*
(0) Aspecto Arbusto.....8
(8) Planta Monoica.....9
(9) Forma de la hoja Escamosa.....10
(10) Consistencia de la fructificación Leñosa.....11
(11) Número de escamas en la piña Entre 6 y 8; Tamaño de la piña (largo) Entre 1 y 2 cm; Presencia de apófisis Si; Disposición de las ramillas en un solo plano Si
.....Género *Platycladus*
(11) Número de escamas en la piña Hasta 4; Tamaño de la piña (largo) Hasta 1 cm; Presencia de apófisis No; Disposición de las ramillas en un solo plano No
.....Género *Tetraclinis*
(10) Consistencia de la fructificación Carnosa
.....Género *Juniperus*
(9) Forma de la hoja Acicular
.....Género *Juniperus*
(8) Planta Dioica
.....Género *Juniperus*

Tabla B- VIII. *Clave para la familia Cupressaceae generada según el criterio de Gini.*

FAMILIA *CUPRESSACEAE* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Aspecto: 2 veces* *Tamaño de la piña (largo): 1 vez* *Número de escamas en la piña: 1 vez* *Consistencia de la fructificación: 1 vez*	*Caracteres confirmadores incluidos:* *Escamas de la piña peltadas: 2 veces* *Escamas de la piña planas: 2 veces* *Disposición de las ramillas en un solo plano: 2 veces*
(0) Consistencia de la fructificación Leñosa.....1 (1) Tamaño de la piña (largo) Entre 1 y 2 cm.....2 (2) Aspecto Árbol; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano No .....Género *Cupressus* (2) Aspecto Arbusto; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano Si .....Género *Platycladus* (1) Tamaño de la piña (largo) Hasta 1 cm.....3 (3) Aspecto Árbol.....4 (4) Número de escamas en la piña Entre 6 y 8; Escamas de la piña peltadas Si; Escamas de la piña planas No; Disposición de las ramillas en un solo plano Si .....Género *Chamaecyparis* (4) Número de escamas en la piña Hasta 4; Escamas de la piña peltadas No; Escamas de la piña planas Si; Disposición de las ramillas en un solo plano No .....Género *Tetraclinis* (3) Aspecto Arbusto .....Género *Tetraclinis* (1) Tamaño de la piña (largo) Entre 2 y 4 cm .....Género *Calocedrus* (0) Consistencia de la fructificación Carnosa .....Género *Juniperus*	

Tabla B- IX. *Clave para la familia Cupressaceae generada según el criterio de Dallwitz.*

FAMILIA PINACEAE.

FAMILIA *PINACEAE* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Presencia de braquiblastos: 1 vez* *Hoja: 1 vez* *Disposición de las piñas: 1 vez* *Escamas tectrices de la piña sobresalientes: 1 vez*	*Caracteres confirmadores incluidos:* *Piñas en ramas superiores de la copa: 1 vez* *Escamas de la piña caducas al madurar: 1 vez*
(0) Disposición de las piñas Erectas.....1 (1) Hoja Persistente; Escamas de la piña caducas al madurar Si.....2 (2) Presencia de braquiblastos No; Piñas en ramas superiores de la copa Si .....Género *Abies* (2) Presencia de braquiblastos Si; Piñas en ramas superiores de la copa No .....Género *Cedrus* (1) Hoja Caduca; Escamas de la piña caducas al madurar No .....Género *Larix* (0) Disposición de las piñas Colgantes.....3 (3) Escamas tectrices de la piña sobresalientes Si .....Género *Pseudotsuga* (3) Escamas tectrices de la piña sobresalientes No .....Género *Picea*	

FAMILIA *PINACEAE* /ENTROPÍA
(0) Disposición de las piñas Patentes o erecto-patentes .....Género *Pinus*

Tabla B- X. *Clave para la familia Pinaceae generada según el criterio de mínima entropía.*

FAMILIA *PINACEAE* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:1*
Atributos incluidos en la clave: *Presencia de braquiblastos: 1 vez* *Disposición de las piñas: 1 vez* *Escamas tectrices de la piña sobresalientes: 1 vez* *Hoja: 1 vez*	*Caracteres confirmadores incluidos:* *Piñas en ramas superiores de la copa: 1 vez*
(0) Hoja Persistente.....1 (1) Disposición de las piñas Erectas.....2 (2) Presencia de braquiblastos No; Piñas en ramas superiores de la copa Si .....Género *Abies* (2) Presencia de braquiblastos Si; Piñas en ramas superiores de la copa No .....Género *Cedrus* (1) Disposición de las piñas Colgantes.....3 (3) Escamas tectrices de la piña sobresalientes Si .....Género *Pseudotsuga* (3) Escamas tectrices de la piña sobresalientes No .....Género *Picea* (1) Disposición de las piñas Patentes o erecto-patentes .....Género *Pinus* (0) Hoja Caduca .....Género *Larix*	

Tabla B- XI. *Clave para la familia Pinaceae generada según el criterio de proporción de ganancia.*

FAMILIA *PINACEAE* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 6	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Piñas en ramas superiores de la copa: 1 vez* *Escamas tectrices de la piña sobresalientes: 1 vez* *Presencia de braquiblastos: 1 vez* *Braquiblastos: 1 vez* *Hoja: 1 vez* *Forma de la copa: 1 vez*	*Caracteres confirmadores incluidos:* *Escamas de la piña caducas al madurar: 2 veces* *Disposición de las piñas: 2 veces* *Hojas dispuestas en espiral en los macroblastos: 1 vez*
(0) Forma de la copa Piramidal.....1 (1) Hoja Persistente.....2 (2) Presencia de braquiblastos No.....3 (3) Piñas en ramas superiores de la copa Si; Escamas de la piña caducas al madurar Si; Disposición de las piñas Erectas .....Género *Abies* (3) Piñas en ramas superiores de la copa No; Escamas de la piña caducas al madurar No; Disposición de las piñas Colgantes.....4 (4) Escamas tectrices de la piña sobresalientes Si .....Género *Pseudotsuga* (4) Escamas tectrices de la piña sobresalientes No .....Género *Picea*	

FAMILIA *PINACEAE* /INDICE DE DIVERSIDAD DE GINI
(2) Presencia de braquiblastos Si.....5 (5) Braquiblastos Desarrollados con muchas hojas fasciculadas; Hojas dispuestas en espiral en los macroblastos Si; Escamas de la piña caducas al madurar Si; Disposición de las piñas Erectas .....Género *Cedrus* (5) Braquiblastos Con 2 a 5 hojas; Hojas dispuestas en espiral en los macroblastos No; Escamas de la piña caducas al madurar No; Disposición de las piñas Patentes o erecto-patentes .....Género *Pinus* (1) Hoja Caduca .....Género *Larix* (0) Forma de la copa Aparasolada .....Género *Pinus*

Tabla B- XII. *Clave para la familia Pinaceae generada según el criterio de Gini.*

FAMILIA *PINACEAE* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:5	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Presencia de braquiblastos: 1 vez* *Hoja: 1 vez* *Disposición de las piñas: 1 vez* *Escamas tectrices de la piña sobresalientes: 1 vez* *Forma de la copa: 1 vez*	*Caracteres confirmadores incluidos:* *Piñas en ramas superiores de la copa: 1 vez* *Escamas de la piña caducas al madurar: 1 vez*
(0) Forma de la copa Piramidal.....1 (1) Disposición de las piñas Erectas.....2 (2) Hoja Persistente; Escamas de la piña caducas al madurar Si.....3 (3) Presencia de braquiblastos No; Piñas en ramas superiores de la copa Si .....Género *Abies* (3) Presencia de braquiblastos Si; Piñas en ramas superiores de la copa No .....Género *Cedrus* (2) Hoja Caduca; Escamas de la piña caducas al madurar No .....Género *Larix* (1) Disposición de las piñas Colgantes.....4 (4) Escamas tectrices de la piña sobresalientes Si .....Género *Pseudotsuga* (4) Escamas tectrices de la piña sobresalientes No .....Género *Picea* (1) Disposición de las piñas Patentes o erecto-patentes .....Género *Pinus* (0) Forma de la copa Aparasolada .....Género *Pinus*	

Tabla B- XIII. *Clave para la familia Pinaceae generada según el criterio de Dallwitz.*

FAMILIA TAXODIACEAE.

FAMILIA *TAXODIACEAE* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave:3	*Total de caracteres confirmadores:9*
Atributos incluidos en la clave: *Hoja: 1 vez* *Forma de la hoja: 1 vez* *Tronco de 10 a 12 m de diámetro: 1 vez*	*Caracteres confirmadores incluidos:* *Con raíces aéreas: 1 vez* *Base del tronco: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Envés con dos bandas: 1 vez* *Corteza rojiza al desprenderse en placas: 1 vez* *Corteza fibrosa: 1 vez* *Tamaño de la piña (largo): 1 vez* *Escamas de la piña peltadas: 1 vez*
(0) Hoja Caduca; Con raíces aéreas Si; Base del tronco Ensanchada .....Género *Taxodium* (0) Hoja Persistente; Con raíces aéreas No; Base del tronco No ensanchada.....1 (1) Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Envés con dos bandas Si .....Género *Sequoia* (1) Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Envés con dos bandas No.....2 (2) Tronco de 10 a 12 m de diámetro Si; Corteza rojiza al desprenderse en placas No; Corteza fibrosa No; Tamaño de la piña (largo) Entre 4 y 6 cm; Escamas de la piña peltadas Si .....Género *Sequoiadendron* (2) Tronco de 10 a 12 m de diámetro No; Corteza rojiza al desprenderse en placas Si; Corteza fibrosa Si; Tamaño de la piña (largo) Entre 2 y 4 cm; Escamas de la piña peltadas No .....Género *Cryptomeria*	

Tabla B- XIV. *Clave para la familia Taxodiaceae generada según el criterio de mínima entropía.*

FAMILIA *TAXODIACEAE* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 3	*Total de caracteres confirmadores:9*
Atributos incluidos en la clave: *Hoja: 1 vez* *Forma de la hoja: 1 vez* *Tronco de 10 a 12 m de diámetro: 1 vez*	*Caracteres confirmadores incluidos:* *Con raíces aéreas: 1 vez* *Base del tronco: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Envés con dos bandas: 1 vez* *Corteza rojiza al desprenderse en placas: 1 vez* *Corteza fibrosa: 1 vez* *Tamaño de la piña (largo): 1 vez* *Escamas de la piña peltadas: 1 vez*
(0) Hoja Caduca; Con raíces aéreas Si; Base del tronco Ensanchada .....Género *Taxodium* (0) Hoja Persistente; Con raíces aéreas No; Base del tronco No ensanchada.....1 (1) Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Envés con dos bandas Si .....Género *Sequoia* (1) Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Envés con dos bandas No.....2 (2) Tronco de 10 a 12 m de diámetro Si; Corteza rojiza al desprenderse en placas No; Corteza fibrosa No; Tamaño de la piña (largo) Entre 4 y 6 cm; Escamas de la piña peltadas Si .....Género *Sequoiadendron* (2) Tronco de 10 a 12 m de diámetro No; Corteza rojiza al desprenderse en placas Si; Corteza	

FAMILIA *TAXODIACEAE* /PROPORCIÓN DE GANANCIA
fibrosa Si; Tamaño de la piña (largo) Entre 2 y 4 cm; Escamas de la piña peltadas No .....Género *Cryptomeria*

Tabla B- XV. *Clave para la familia Taxodiaceae generada según el criterio de proporción de ganancia.*

FAMILIA *TAXODIACEAE* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave:4	*Total de caracteres confirmadores:5*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 1 vez* *Hoja: 1 vez* *Tronco de 10 a 12 m de diámetro: 1 vez* *Envés con dos bandas: 1 vez*	*Caracteres confirmadores incluidos:* *Forma de la hoja: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Con raíces aéreas: 1 vez* *Corteza fibrosa: 1 vez*
(0) Envés con dos bandas No.....1 (1) Tronco de 10 a 12 m de diámetro No; Corteza fibrosa Si.....2 (2) Tamaño de la piña (largo) Entre 1 y 2 cm .....Género *Taxodium* (2) Tamaño de la piña (largo) Entre 2 y 4 cm.....3 (3) Hoja Caduca; Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Con raíces aéreas Si .....Género *Taxodium* (3) Hoja Persistente; Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Con raíces aéreas No .....Género *Cryptomeria* (1) Tronco de 10 a 12 m de diámetro Si; Corteza fibrosa No .....Género *Sequoiadendron* (0) Envés con dos bandas Si .....Género *Sequoia*	

Tabla B- XVI. *Clave para la familia Taxodiaceae generada según el criterio de Gini.*

FAMILIA *TAXODIACEAE* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:3	*Total de caracteres confirmadores:8*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 1 vez* *Hoja: 1 vez* *Forma de la hoja: 1 vez*	*Caracteres confirmadores incluidos:* *Envés con dos bandas: 1 vez* *Con raíces aéreas: 1 vez* *Corteza rojiza al desprenderse en placas: 1 vez* *Base del tronco: 1 vez* *Hoja punzante: 1 vez* *Hoja plana: 1 vez* *Escamas de la piña peltadas: 1 vez* *Escamas de la piña recurvadas, con 4-6 esquinas: 1 vez*
(0) Tamaño de la piña (largo) Entre 1 y 2 cm .....Género *Taxodium* (0) Tamaño de la piña (largo) Entre 2 y 4 cm.....1 (1) Forma de la hoja Linear (planoaguzada); Hoja punzante No; Hoja plana Si; Escamas de la piña peltadas Si; Escamas de la piña recurvadas, con 4-6 esquinas No.....2 (2) Hoja Caduca; Envés con dos bandas No; Con raíces aéreas Si; Corteza rojiza al desprenderse en placas No; Base del tronco Ensanchada .....Género *Taxodium* (2) Hoja Persistente; Envés con dos bandas Si; Con raíces aéreas No; Corteza rojiza al desprenderse en placas Si; Base del tronco No ensanchada .....Género *Sequoia*	

FAMILIA *TAXODIACEAE* /CRITERIO DE DALLWITZ
(1) Forma de la hoja Aleznada; Hoja punzante Si; Hoja plana No; Escamas de la piña peltadas No; Escamas de la piña recurvadas, con 4-6 esquinas Si .....Género *Cryptomeria* (0) Tamaño de la piña (largo) Entre 4 y 6 cm .....Género *Sequoiadendron*

Tabla B- XVII. *Clave para la familia Taxodiaceae generada según el criterio de Dallwitz.*

GÉNERO PINUS.

GÉNERO *PINUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 2	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Características de la apófisis: 1 vez* *Color de la corteza (ritidoma): 1 vez*	*Caracteres confirmadores incluidos:* *Piñas brillantes: 1 vez* *Hoja de color: 1 vez* *Con hojas: 1 vez*
(0) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (0) Características de la apófisis Poco prominente.....1 (1) Color de la corteza (ritidoma) Gris-ceniciento; Piñas brillantes Si; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus nigra subsp. salzmannii* (1) Color de la corteza (ritidoma) Pardo-rojizo; Piñas brillantes No; Hoja de color Verde claro; Con hojas Rigidas .....Especie *Pinus sylvestris* (0) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (0) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (0) Características de la apófisis Convexa .....Especie *Pinus pinea* (0) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (0) Características de la apófisis Prominente .....Especie *Pinus canariensis*	

Tabla B- XVIII. *Clave para el género Pinus generada según el criterio de mínima entropía.*

GÉNERO *PINUS* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave:2	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Características de la apófisis: 1 vez* *Color de la corteza (ritidoma): 1 vez*	*Caracteres confirmadores incluidos:* *Piñas brillantes: 1 vez* *Hoja de color: 1 vez* *Con hojas: 1 vez*
(0) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (0) Características de la apófisis Poco prominente.....1 (1) Color de la corteza (ritidoma) Gris-ceniciento; Piñas brillantes Si; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus nigra subsp. salzmannii* (1) Color de la corteza (ritidoma) Pardo-rojizo; Piñas brillantes No; Hoja de color Verde claro; Con hojas Rígidas .....Especie *Pinus sylvestris*	

GÉNERO *PINUS* /PROPORCIÓN DE GANANCIA
(0) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (0) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (0) Características de la apófisis Convexa .....Especie *Pinus pinea* (0) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (0) Características de la apófisis Prominente .....Especie *Pinus canariensis*

Tabla B- XIX. *Clave para el género Pinus generada según el criterio de proporción de ganancia.*

GÉNERO *PINUS* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 10	*Total de caracteres confirmadores:10*
Atributos incluidos en la clave: *Características de la apófisis: 2 veces* *Tamaño: 5 veces* *Tamaño de las hojas (largo): 3 veces* *Tamaño de la piña (largo): 3 veces* *Tamaño de la piña (ancho): 1 vez* *Tamaño de las hojas (ancho): 2 veces* *Piñas brillantes: 1 vez* *Forma de la copa: 1 vez* *Número de hojas por fascículo: 1 vez* *Color de la corteza (ritidoma): 1 vez*	*Caracteres confirmadores incluidos:* *Hoja de color: 6 veces* *Con hojas: 4 veces* *Color de la corteza (ritidoma): 1 vez* *Color de las ramillas: 3 veces* *Piñas con pedúnculo evidente: 3 veces* *Tamaño de las hojas (ancho): 2 veces* *Características de la apófisis: 4 veces* *Semilla alada y persistente: 1 vez* *Con piñón: 1 vez* *Tamaño de la piña (ancho): 1 vez*
(0) Número de hojas por fascículo 2.....1 (1) Forma de la copa Piramidal; Semilla alada y persistente Si; Con piñón No.....2 (2) Piñas brillantes Si.....3 (3) Tamaño de la piña (ancho) Entre 4 y 7.5 cm.....4 (4) Tamaño de la piña (largo) Entre 8 y 12 cm.....5 (5) Tamaño de las hojas (largo) Entre 7 y 15 cm.....6 (6) Tamaño 4000 cm; Color de la corteza (ritidoma) Pardo-rojizo; Color de las ramillas Pardo-rojizas o castañas; Piñas con pedúnculo evidente No; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm.....7 (7) Características de la apófisis Prominente y punzante; Hoja de color Verde oscuro; Con hojas Rígidas .....Especie *Pinus pinaster* (7) Características de la apófisis Muy prominente y punzante; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus radiata* (6) Tamaño 2000 cm; Color de la corteza (ritidoma) Gris-ceniciento; Color de las ramillas Cenicientas; Piñas con pedúnculo evidente Si; Tamaño de las hojas (ancho) Hasta 0.1 cm .....Especie *Pinus halepensis* (5) Tamaño de las hojas (largo) Entre 15 y 20 cm .....Especie *Pinus pinaster* (5) Tamaño de las hojas (largo) Entre 20 y 25 cm .....Especie *Pinus pinaster* (5) Tamaño de las hojas (largo) Entre 5 y 7 cm .....Especie *Pinus halepensis* (4) Tamaño de la piña (largo) Entre 12 y 15 cm.....8 (8) Tamaño de las hojas (largo) Entre 7 y 15 cm.....9 (9) Características de la apófisis Prominente y punzante; Hoja de color Verde oscuro; Con hojas	

GÉNERO *PINUS* /INDICE DE DIVERSIDAD DE GINI

Rígidas
.....Especie *Pinus pinaster*
(9) Características de la apófisis Muy prominente y punzante; Hoja de color Verde intenso; Con hojas Flexibles
.....Especie *Pinus radiata*
(8) Tamaño de las hojas (largo) Entre 15 y 20 cm
.....Especie *Pinus pinaster*
(8) Tamaño de las hojas (largo) Entre 20 y 25 cm
.....Especie *Pinus pinaster*
(4) Tamaño de la piña (largo) Entre 15 y 18 cm
.....Especie *Pinus pinaster*
(4) Tamaño de la piña (largo) Mas de 18 cm
.....Especie *Pinus pinaster*
(4) Tamaño de la piña (largo) Entre 6 y 8 cm
.....Especie *Pinus halepensis*
(3) Tamaño de la piña (ancho) Entre 2 y 4 cm.....10
(10) Tamaño de las hojas (largo) Entre 5 y 7 cm.....11
(11) Tamaño de las hojas (ancho) Hasta 0.1 cm.....12
(12) Tamaño 4000 cm; Color de las ramillas Pardo-rojizas o castañas; Piñas con pedúnculo evidente No; Características de la apófisis Poco prominente; Hoja de color Verde intenso
.....Especie *Pinus nigra subsp. salzmannii*
(12) Tamaño 2000 cm; Color de las ramillas Cenicientas; Piñas con pedúnculo evidente Si; Características de la apófisis Poco convexa; Hoja de color Verde claro
.....Especie *Pinus halepensis*
(11) Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm.....13
(13) Tamaño de la piña (largo) Entre 4 y 6 cm.....14
(14) Tamaño 4000 cm; Características de la apófisis Poco prominente; Hoja de color Verde intenso; Con hojas Flexibles
.....Especie *Pinus nigra subsp. salzmannii*
(14) Tamaño 2500 cm; Características de la apófisis Muy prominente, ganchuda; Hoja de color Verde oscuro; Con hojas Rígidas
.....Especie *Pinus uncinata*
(13) Tamaño de la piña (largo) Entre 6 y 8 cm
.....Especie *Pinus uncinata*
(10) Tamaño de las hojas (largo) Entre 7 y 15 cm.....15
(15) Tamaño de las hojas (ancho) Hasta 0.1 cm.....16
(16) Tamaño 4000 cm; Color de las ramillas Pardo-rojizas o castañas; Piñas con pedúnculo evidente No; Características de la apófisis Poco prominente; Hoja de color Verde intenso
.....Especie *Pinus nigra subsp. salzmannii*
(16) Tamaño 2000 cm; Color de las ramillas Cenicientas; Piñas con pedúnculo evidente Si; Características de la apófisis Poco convexa; Hoja de color Verde claro
.....Especie *Pinus halepensis*
(15) Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm
.....Especie *Pinus nigra subsp. salzmannii*
(10) Tamaño de las hojas (largo) Entre 3 y 5 cm
.....Especie *Pinus uncinata*
(2) Piñas brillantes No
.....Especie *Pinus sylvestris*
(1) Forma de la copa Aparasolada; Semilla alada y persistente No; Con piñón Si
.....Especie *Pinus pinea*
(0) Número de hojas por fascículo 3.....17
(17) Color de la corteza (ritidoma) Gris-ceniciento; Tamaño de la piña (ancho) Entre 2 y 4 cm; Hoja de color Verde oscuro; Con hojas Rígidas
.....Especie *Pinus uncinata*
(17) Color de la corteza (ritidoma) Pardo-rojizo; Tamaño de la piña (ancho) Entre 4 y 7.5 cm; Hoja de color Verde intenso; Con hojas Flexibles.....18
(18) Tamaño de la piña (largo) Entre 8 y 12 cm
.....Especie *Pinus radiata*
(18) Tamaño de la piña (largo) Entre 12 y 15 cm.....19

GÉNERO *PINUS* /INDICE DE DIVERSIDAD DE GINI
(19) Tamaño 6000 cm; Características de la apófisis Prominente; Tamaño de las hojas (ancho) Hasta 0.1 cm .....Especie *Pinus canariensis* (19) Tamaño 4000 cm; Características de la apófisis Muy prominente y punzante; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm .....Especie *Pinus radiata* (18) Tamaño de la piña (largo) Entre 15 y 18 cm .....Especie *Pinus canariensis*

Tabla B- XX. *Clave para el género Pinus generada según el criterio de Gini.*

GÉNERO *PINUS* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave: 7	*Total de caracteres confirmadores:9*
Atributos incluidos en la clave: *Tamaño de la piña (ancho): 1 vez* *Tamaño: 2 veces* *Con hojas: 1 vez* *Color de la corteza (ritidoma): 1 vez* *Forma de la copa: 1 vez* *Tamaño de la piña (largo): 1 vez* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Color de la corteza (ritidoma): 1 vez* *Características de la apófisis: 2 veces* *Piñas brillantes: 1 vez* *Hoja de color: 3 veces* *Color de las ramillas: 1 vez* *Piñas con pedúnculo evidente: 1 vez* *Tamaño de la piña (ancho): 1 vez* *Semilla alada y persistente: 1 vez* *Con piñón: 1 vez*
(0) Con hojas Rígidas.....1 (1) Tamaño de la piña (ancho) Entre 4 y 7.5 cm .....Especie *Pinus pinaster* (1) Tamaño de la piña (ancho) Entre 2 y 4 cm.....2 (2) Tamaño 4000 cm; Color de la corteza (ritidoma) Pardo-rojizo; Características de la apófisis Poco prominente; Piñas brillantes No; Hoja de color Verde claro .....Especie *Pinus sylvestris* (2) Tamaño 2500 cm; Color de la corteza (ritidoma) Gris-ceniciento; Características de la apófisis Muy prominente, ganchuda; Piñas brillantes Si; Hoja de color Verde oscuro .....Especie *Pinus uncinata* (0) Con hojas Flexibles.....3 (3) Color de la corteza (ritidoma) Gris-ceniciento.....4 (4) Tamaño 4000 cm; Color de las ramillas Pardo-rojizas o castañas; Piñas con pedúnculo evidente No; Características de la apófisis Poco prominente; Hoja de color Verde intenso .....Especie *Pinus nigra subsp. salzmannii* (4) Tamaño 2000 cm; Color de las ramillas Cenicientas; Piñas con pedúnculo evidente Si; Características de la apófisis Poco convexa; Hoja de color Verde claro .....Especie *Pinus halepensis* (3) Color de la corteza (ritidoma) Pardo-rojizo.....5 (5) Forma de la copa Aparasolada; Tamaño de la piña (ancho) Mas de 7.5 cm; Hoja de color Verde claro; Semilla alada y persistente No; Con piñón Si .....Especie *Pinus pinea* (5) Forma de la copa Piramidal; Tamaño de la piña (ancho) Entre 4 y 7.5 cm; Hoja de color Verde intenso; Semilla alada y persistente Si; Con piñón No.....6 (6) Tamaño de la piña (largo) Entre 8 y 12 cm .....Especie *Pinus radiata* (6) Tamaño de la piña (largo) Entre 12 y 15 cm.....7 (7) Tamaño de las hojas (largo) Entre 7 y 15 cm .....Especie *Pinus radiata*	

GÉNERO *PINUS* /CRITERIO DE DALLWITZ
(7) Tamaño de las hojas (largo) Entre 15 y 20 cm .....Especie *Pinus canariensis* (7) Tamaño de las hojas (largo) Entre 20 y 25 cm .....Especie *Pinus canariensis* (6) Tamaño de la piña (largo) Entre 15 y 18 cm .....Especie *Pinus canariensis*

Tabla B- XXI. *Clave para el género Pinus generada según el criterio de Dallwitz.*

GÉNERO *PINUS* /CRITERIO EXPERTO1	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:5*
Atributos incluidos en la clave: *Características de la apófisis: 2 veces* *Con hojas: 2 veces* *Número de hojas por fascículo: 1 vez* *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Color de la corteza (ritidoma): 1 vez* *Tamaño de la piña (ancho): 1 vez* *Hoja de color: 1 vez* *Características de la apófisis: 1 vez* *Tamaño de las hojas (ancho): 1 vez*
(0) Número de hojas por fascículo 2.....1 (1) Con hojas Rígidas.....2 (2) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (2) Características de la apófisis Poco prominente .....Especie *Pinus sylvestris* (2) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (1) Con hojas Flexibles.....3 (3) Características de la apófisis Poco prominente .....Especie *Pinus nigra subsp. salzmannii* (3) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (3) Características de la apófisis Convexa .....Especie *Pinus pinea* (3) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (0) Número de hojas por fascículo 3.....4 (4) Con hojas Rígidas; Color de la corteza (ritidoma) Gris-ceniciento; Tamaño de la piña (ancho) Entre 2 y 4 cm; Hoja de color Verde oscuro .....Especie *Pinus uncinata* (4) Con hojas Flexibles; Color de la corteza (ritidoma) Pardo-rojizo; Tamaño de la piña (ancho) Entre 4 y 7.5 cm; Hoja de color Verde intenso.....5 (5) Tamaño 4000 cm; Características de la apófisis Muy prominente y punzante; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm .....Especie *Pinus radiata* (5) Tamaño 6000 cm; Características de la apófisis Prominente; Tamaño de las hojas (ancho) Hasta 0.1 cm .....Especie *Pinus canariensis*	

Tabla B- XXII. *Clave para el género Pinus generada según el criterio del Experto.*

GÉNERO *PINUS* /CRITERIO EXPERTO2	
Total de atributos diferentes incluidos en la clave: 5	*Total de caracteres confirmadores:6*
Atributos incluidos en la clave: *Características de la apófisis: 1 vez* *Color de la corteza (ritidoma): 1 vez* *Con piñón: 1 vez* *Número de hojas por fascículo: 1 vez* *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Piñas brillantes: 1 vez* *Hoja de color: 1 vez* *Con hojas: 1 vez* *Forma de la copa: 1 vez* *Semilla alada y persistente: 1 vez* *Características de la apófisis: 1 vez*
(0) Número de hojas por fascículo 2.....1 (1) Con piñón No; Forma de la copa Piramidal; Semilla alada y persistente Si.....2 (2) Características de la apófisis Prominente y punzante .....Especie *Pinus pinaster* (2) Características de la apófisis Poco prominente.....3 (3) Color de la corteza (ritidoma) Gris-ceniciento; Piñas brillantes Si; Hoja de color Verde intenso; Con hojas Flexibles .....Especie *Pinus nigra subsp. salzmannii* (3) Color de la corteza (ritidoma) Pardo-rojizo; Piñas brillantes No; Hoja de color Verde claro; Con hojas Rígidas .....Especie *Pinus sylvestris* (2) Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (2) Características de la apófisis Poco convexa .....Especie *Pinus halepensis* (2) Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (1) Con piñón Si; Forma de la copa Aparasolada; Semilla alada y persistente No .....Especie *Pinus pinea* (0) Número de hojas por fascículo 3.....4 (4) Tamaño 2500 cm; Características de la apófisis Muy prominente, ganchuda .....Especie *Pinus uncinata* (4) Tamaño 4000 cm; Características de la apófisis Muy prominente y punzante .....Especie *Pinus radiata* (4) Tamaño 6000 cm; Características de la apófisis Prominente .....Especie *Pinus canariensis*	

Tabla B- XXIII. *Clave para el género Pinus generada según el criterio del Experto.*

GÉNERO JUNIPERUS.

GÉNERO *JUNIPERUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:7*
Atributos incluidos en la clave: *Número de franjas blancas en el haz: 1 vez* *Tamaño: 1 vez* *Hojas rectas: 1 vez* *Tamaño de la arcestida: 1 vez*	*Caracteres confirmadores incluidos:* *Color de la arcestida: 1 vez* *Arcestida pruinosa: 1 vez* *Tamaño de la arcestida: 1 vez* *Número de semillas de la arcestida: 1 vez* *Hojas incurvas: 1 vez* *Tamaño de las hojas (ancho): 1 vez* *En dunas y arenales: 1 vez*
(0) Tamaño 1500 cm.....1 (1) Número de franjas blancas en el haz 1; Color de la arcestida Negro-azulado; Arcestida	

GÉNERO *JUNIPERUS* /ENTROPÍA
pruinosa Si; Tamaño de la arcestida Entre 0.6 y 1 cm; Número de semillas de la arcestida 3 .....Especie *Juniperus communis subsp. communis* (1) Número de franjas blancas en el haz 2; Color de la arcestida Castaño; Arcestida pruinosa No; Tamaño de la arcestida Más de 1 cm; Número de semillas de la arcestida 1-3 .....Especie *Juniperus oxycedrus subsp. badia* (0) Tamaño 250 cm.....2 (2) Hojas rectas Si; Hojas incurvas No .....Especie *Juniperus communis subsp. hemisphaerica* (2) Hojas rectas No; Hojas incurvas Si .....Especie *Juniperus communis subsp. alpina* (0) Tamaño 400 cm .....Especie *Juniperus oxycedrus subsp. oxycedrus* (0) Tamaño 300 cm .....Especie *Juniperus oxycedrus subsp. macrocarpa* (0) Tamaño 200 cm .....Especie *Juniperus navicularis* (0) Tamaño 800 cm.....3 (3) Tamaño de la arcestida Entre 0.6 y 1 cm; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm; En dunas y arenales No .....Especie *Juniperus phoenicea subsp. phoenicea* (3) Tamaño de la arcestida Más de 1 cm; Tamaño de las hojas (ancho) Hasta 0.1 cm; En dunas y arenales Si .....Especie *Juniperus phoenicea subsp. turbinata* (0) Tamaño 2000 cm .....Especie *Juniperus thurifera* (0) Tamaño 100 cm .....Especie *Juniperus sabina*

Tabla B- XXIV. *Clave para el género Juniperus generada según el criterio de mínima entropía.*

GÉNERO *JUNIPERUS* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 5	*Total de caracteres confirmadores:6*
Atributos incluidos en la clave: *Tamaño: 2 veces* *Arcestida pruinosa: 1 vez* *Planta: 1 vez* *Tamaño de la arcestida: 1 vez* *Hojas incurvas: 1 vez*	*Caracteres confirmadores incluidos:* *Número de franjas blancas en el haz: 1 vez* *Disposición de las arcestidas: 1 vez* *Número de semillas de la arcestida: 1 vez* *Forma de la hoja: 1 vez* *Tamaño de las hojas (ancho): 1 vez* *En dunas y arenales: 1 vez*
(0) Hojas incurvas No.....1 (1) Arcestida pruinosa Si.....2 (2) Tamaño 1500 cm .....Especie *Juniperus communis subsp. communis* (2) Tamaño 250 cm .....Especie *Juniperus communis subsp. hemisphaerica* (2) Tamaño 300 cm .....Especie *Juniperus oxycedrus subsp. macrocarpa* (2) Tamaño 2000 cm .....Especie *Juniperus thurifera* (2) Tamaño 100 cm .....Especie *Juniperus sabina* (1) Arcestida pruinosa No.....3 (3) Planta Dioica; Número de franjas blancas en el haz 2; Disposición de las arcestidas Axilares; Número de semillas de la arcestida 1-3; Forma de la hoja Acicular.....4	

GÉNERO *JUNIPERUS* /PROPORCIÓN DE GANANCIA
(4) Tamaño 400 cm .....Especie *Juniperus oxycedrus subsp. oxycedrus* (4) Tamaño 1500 cm .....Especie *Juniperus oxycedrus subsp. badia* (4) Tamaño 200 cm .....Especie *Juniperus navicularis* (3) Planta Monoica; Número de franjas blancas en el haz 0; Disposición de las arcestidas Terminales; Número de semillas de la arcestida 3-9; Forma de la hoja Escamosa.....5 (5) Tamaño de la arcestida Entre 0.6 y 1 cm; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm; En dunas y arenales No .....Especie *Juniperus phoenicea subsp. phoenicea* (5) Tamaño de la arcestida Más de 1 cm; Tamaño de las hojas (ancho) Hasta 0.1 cm; En dunas y arenales Si .....Especie *Juniperus phoenicea subsp. turbinata* (0) Hojas incurvas Si .....Especie *Juniperus communis subsp. alpina*

Tabla B- XXV. *Clave para el género Juniperus generada según el criterio de proporción de ganancia.*

GÉNERO *JUNIPERUS* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 10	*Total de caracteres confirmadores:15*
Atributos incluidos en la clave: *Tamaño: 8 veces* *Disposición de las arcestidas: 1 vez* *Tamaño de las hojas (largo): 3 veces* *Tamaño de la arcestida: 4 veces* *Altitud: 4 veces* *Número de franjas blancas en el haz: 2 veces* *Aspecto: 1 vez* *Tipo de arbusto: 4 veces* *Distribución de las hojas: 1 vez* *Hojas incurvas: 1 vez*	*Caracteres confirmadores incluidos:* *Número de franjas blancas en el haz: 3 veces* *Color de la arcestida: 8 veces* *Arcestida pruinosa: 6 veces* *Número de semillas de la arcestida: 6 veces* *Forma de la hoja: 2 veces* *Hoja punzante: 2 veces* *Hojas rectas: 2 veces* *Distribución de las hojas: 1 vez* *Planta: 6 veces* *Disposición de las arcestidas: 3 veces* *Tamaño de las hojas (largo): 1 vez* *En dunas y arenales: 3 veces* *Tamaño de las hojas (ancho): 2 veces* *Tamaño: 1 vez* *Tamaño de la arcestida: 1 vez*
(0) Aspecto Árbol.....1 (1) Altitud Hasta 1000 m.....2 (2) Tamaño de la arcestida Entre 0.6 y 1 cm.....3 (3) Disposición de las arcestidas Axilares; Forma de la hoja Acicular; Hoja punzante Si; Hojas rectas Si; Distribución de las hojas Distanciada.....4 (4) Tamaño 1500 cm; Número de franjas blancas en el haz 1; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 3 .....Especie *Juniperus communis subsp. communis* (4) Tamaño 400 cm; Número de franjas blancas en el haz 2; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 1-3 .....Especie *Juniperus oxycedrus subsp. oxycedrus* (3) Disposición de las arcestidas Terminales; Forma de la hoja Escamosa; Hoja punzante No; Hojas rectas No; Distribución de las hojas Densa.....5 (5) Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....6	

GÉNERO *JUNIPERUS* /INDICE DE DIVERSIDAD DE GINI

(6) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9
.....Especie *Juniperus phoenicea subsp. phoenicea*
(6) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4
.....Especie *Juniperus thurifera*
(5) Tamaño de las hojas (largo) Entre 1 y 2 cm
.....Especie *Juniperus phoenicea subsp. phoenicea*
(5) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm
.....Especie *Juniperus thurifera*
(2) Tamaño de la arcestida Más de 1 cm
.....Especie *Juniperus oxycedrus subsp. badia*
(1) Altitud Entre 1000 y 2000 m.....7
(7) Número de franjas blancas en el haz 1; Disposición de las arcestidas Axilares; Forma de la hoja Acicular; Hoja punzante Si; Hojas rectas Si
.....Especie *Juniperus communis subsp. communis*
(7) Número de franjas blancas en el haz 0; Disposición de las arcestidas Terminales; Forma de la hoja Escamosa; Hoja punzante No; Hojas rectas No.....8
(8) Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....9
(9) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9
.....Especie *Juniperus phoenicea subsp. phoenicea*
(9) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4
.....Especie *Juniperus thurifera*
(8) Tamaño de las hojas (largo) Entre 1 y 2 cm
.....Especie *Juniperus phoenicea subsp. phoenicea*
(8) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm
.....Especie *Juniperus thurifera*
(0) Aspecto Arbusto.....10
(10) Hojas incurvas No.....11
(11) Distribución de las hojas Distanciada.....12
(12) Tipo de arbusto Erecto.....13
(13) Tamaño de la arcestida Entre 0.6 y 1 cm; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm.....14
(14) Altitud Hasta 1000 m.....15
(15) Número de franjas blancas en el haz 1; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 3; Tamaño de las hojas (largo) Entre 1 y 2 cm
.....Especie *Juniperus communis subsp. communis*
(15) Número de franjas blancas en el haz 2; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 1-3; Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....16
(16) Tamaño 400 cm; En dunas y arenales No
.....Especie *Juniperus oxycedrus subsp. oxycedrus*
(16) Tamaño 200 cm; En dunas y arenales Si
.....Especie *Juniperus navicularis*
(14) Altitud Entre 1000 y 2000 m
.....Especie *Juniperus communis subsp. communis*
(13) Tamaño de la arcestida Más de 1 cm; Tamaño de las hojas (ancho) Mas de 0.2 cm
.....Especie *Juniperus oxycedrus subsp. macrocarpa*
(12) Tipo de arbusto Postrado
.....Especie *Juniperus oxycedrus subsp. macrocarpa*
(11) Distribución de las hojas Densa.....17
(17) Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....18
(18) Tipo de arbusto Postrado.....19
(19) Tamaño 250 cm; Planta Dioica; Número de franjas blancas en el haz 1; Disposición de las arcestidas Axilares; Color de la arcestida Negro-azulado
.....Especie *Juniperus communis subsp. hemisphaerica*
(19) Tamaño 800 cm; Planta Monoica; Número de franjas blancas en el haz 0; Disposición de las arcestidas Terminales; Color de la arcestida Castaño
.....Especie *Juniperus phoenicea subsp. turbinata*

GÉNERO *JUNIPERUS* /INDICE DE DIVERSIDAD DE GINI
(18) Tipo de arbusto Erecto.....20 (20) Tamaño de la arcestida Entre 0.6 y 1 cm.....21 (21) Altitud Hasta 1000 m.....22 (22) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9 .....Especie *Juniperus phoenicea subsp. phoenicea* (22) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4 .....Especie *Juniperus thurifera* (21) Altitud Entre 1000 y 2000 m.....23 (23) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9 .....Especie *Juniperus phoenicea subsp. phoenicea* (23) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4 .....Especie *Juniperus thurifera* (20) Tamaño de la arcestida Más de 1 cm .....Especie *Juniperus phoenicea subsp. turbinata* (17) Tamaño de las hojas (largo) Entre 1 y 2 cm.....24 (24) Tipo de arbusto Postrado.....25 (25) Tamaño 250 cm; Planta Dioica; Número de franjas blancas en el haz 1; Disposición de las arcestidas Axilares; Color de la arcestida Negro-azulado .....Especie *Juniperus communis subsp. hemisphaerica* (25) Tamaño 800 cm; Planta Monoica; Número de franjas blancas en el haz 0; Disposición de las arcestidas Terminales; Color de la arcestida Castaño .....Especie *Juniperus phoenicea subsp. turbinata* (24) Tipo de arbusto Erecto.....26 (26) Tamaño de la arcestida Entre 0.6 y 1 cm; Tamaño de las hojas (ancho) Entre 0.1 y 0.2 cm; En dunas y arenales No .....Especie *Juniperus phoenicea subsp. phoenicea* (26) Tamaño de la arcestida Más de 1 cm; Tamaño de las hojas (ancho) Hasta 0.1 cm; En dunas y arenales Si .....Especie *Juniperus phoenicea subsp. turbinata* (17) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm.....27 (27) Altitud Hasta 1000 m .....Especie *Juniperus thurifera* (27) Altitud Entre 1000 y 2000 m.....28 (28) Tipo de arbusto Erecto; Tamaño 2000 cm; Tamaño de la arcestida Entre 0.6 y 1 cm; En dunas y arenales Si .....Especie *Juniperus thurifera* (28) Tipo de arbusto Postrado; Tamaño 100 cm; Tamaño de la arcestida Hasta 0.6 cm; En dunas y arenales No .....Especie *Juniperus sabina* (27) Altitud Mas de 2000 m .....Especie *Juniperus sabina* (10) Hojas incurvas Si .....Especie *Juniperus communis subsp. alpina*

Tabla B- XXVI. *Clave para el género Juniperus generada según el criterio de Gini.*

GÉNERO *JUNIPERUS* /CRITERIO EXPERTO1	
Total de atributos diferentes incluidos en la clave: 6	*Total de caracteres confirmadores:5*
Atributos incluidos en la clave: *Tipo de arbusto: 2 veces* *Hojas rectas: 1 vez* *Número de semillas de la arcestida: 1 vez* *En dunas y arenales: 1 vez* *Tamaño de la arcestida: 2 veces* *Aspecto: 1 vez*	*Caracteres confirmadores incluidos:* *Distribución de las hojas: 1 vez* *Hojas incurvas: 1 vez* *Arcestida pruinosa: 1 vez* *En dunas y arenales: 3 veces* *Tamaño de la arcestida: 1 vez*
(0) Número de semillas de la arcestida 3.....1 (1) Tipo de arbusto Erecto; Distribución de las hojas Distanciada .....Especie *Juniperus communis subsp. communis* (1) Tipo de arbusto Postrado; Distribución de las hojas Densa.....2 (2) Hojas rectas Si; Hojas incurvas No .....Especie *Juniperus communis subsp. hemisphaerica* (2) Hojas rectas No; Hojas incurvas Si .....Especie *Juniperus communis subsp. alpina* (0) Número de semillas de la arcestida 1-3.....3 (3) Tamaño de la arcestida Entre 0.6 y 1 cm.....4 (4) En dunas y arenales No .....Especie *Juniperus oxycedrus subsp. oxycedrus* (4) En dunas y arenales Si .....Especie *Juniperus navicularis* (3) Tamaño de la arcestida Más de 1 cm.....5 (5) Aspecto Árbol; Arcestida pruinosa No; En dunas y arenales No .....Especie *Juniperus oxycedrus subsp. badia* (5) Aspecto Arbusto; Arcestida pruinosa Si; En dunas y arenales Si .....Especie *Juniperus oxycedrus subsp. macrocarpa* (0) Número de semillas de la arcestida 3-9.....6 (6) Tamaño de la arcestida Entre 0.6 y 1 cm; En dunas y arenales No .....Especie *Juniperus phoenicea subsp. phoenicea* (6) Tamaño de la arcestida Más de 1 cm; En dunas y arenales Si .....Especie *Juniperus phoenicea subsp. turbinata* (0) Número de semillas de la arcestida 2-4.....7 (7) Tipo de arbusto Erecto; Tamaño de la arcestida Entre 0.6 y 1 cm; En dunas y arenales Si .....Especie *Juniperus thurifera* (7) Tipo de arbusto Postrado; Tamaño de la arcestida Hasta 0.6 cm; En dunas y arenales No .....Especie *Juniperus sabina*	

Tabla B- XXVII. *Clave para el género Juniperus generada según el criterio del Experto.*

GÉNERO *JUNIPERUS* /CRITERIO EXPERTO2	
Total de atributos diferentes incluidos en la clave: 8	*Total de caracteres confirmadores:8*
Atributos incluidos en la clave: *Tipo de arbusto: 1 vez* *Hojas rectas: 1 vez* *Número de franjas blancas en el haz: 1 vez* *En dunas y arenales: 1 vez* *Tamaño de la arcestida: 2 veces* *Aspecto: 1 vez* *Forma de la hoja: 1 vez* *Planta: 1 vez*	*Caracteres confirmadores incluidos:* *Distribución de las hojas: 1 vez* *Hojas incurvas: 1 vez* *Color de la arcestida: 2 veces* *Número de semillas de la arcestida: 2 veces* *Arcestida pruinosa: 2 veces* *En dunas y arenales: 2 veces* *Disposición de las arcestidas: 1 vez* *Hoja punzante: 1 vez*
(0) Forma de la hoja Acicular; Disposición de las arcestidas Axilares; Hoja punzante Si.....1	

GÉNERO *JUNIPERUS* /CRITERIO EXPERTO2
(1) Número de franjas blancas en el haz 1; Color de la arcestida Negro-azulado; Número de semillas de la arcestida 3.....2
(2) Tipo de arbusto Erecto; Distribución de las hojas Distanciada
.....Especie *Juniperus communis subsp. communis*
(2) Tipo de arbusto Postrado; Distribución de las hojas Densa.....3
(3) Hojas rectas Si; Hojas incurvas No
.....Especie *Juniperus communis subsp. hemisphaerica*
(3) Hojas rectas No; Hojas incurvas Si
.....Especie *Juniperus communis subsp. alpina*
(1) Número de franjas blancas en el haz 2; Color de la arcestida Castaño; Número de semillas de la arcestida 1-3.....4
(4) Tamaño de la arcestida Entre 0.6 y 1 cm.....5
(5) En dunas y arenales No
.....Especie *Juniperus oxycedrus subsp. oxycedrus*
(5) En dunas y arenales Si
.....Especie *Juniperus navicularis*
(4) Tamaño de la arcestida Más de 1 cm.....6
(6) Aspecto Árbol; Arcestida pruinosa No; En dunas y arenales No
.....Especie *Juniperus oxycedrus subsp. badia*
(6) Aspecto Arbusto; Arcestida pruinosa Si; En dunas y arenales Si
.....Especie *Juniperus oxycedrus subsp. macrocarpa*
(0) Forma de la hoja Escamosa; Disposición de las arcestidas Terminales; Hoja punzante No.....7
(7) Tamaño de la arcestida Entre 0.6 y 1 cm.....8
(8) Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9; En dunas y arenales No
.....Especie *Juniperus phoenicea subsp. phoenicea*
(8) Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4; En dunas y arenales Si
.....Especie *Juniperus thurifera*
(7) Tamaño de la arcestida Más de 1 cm
.....Especie *Juniperus phoenicea subsp. turbinata*
(7) Tamaño de la arcestida Hasta 0.6 cm
.....Especie *Juniperus sabina*

Tabla B- XXVIII. *Clave para el género Juniperus generada según el criterio del Experto.*

GÉNERO *JUNIPERUS* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:7	*Total de caracteres confirmadores:9*
Atributos incluidos en la clave: *Número de franjas blancas en el haz: 2 veces* *Tamaño: 4 veces* *Distribución de las hojas: 1 vez* *Hojas rectas: 1 vez* *Tamaño de las hojas (largo): 2 veces* *Aspecto: 1 vez* *Altitud: 1 vez*	*Caracteres confirmadores incluidos:* *Color de la arcestida: 3 veces* *Arcestida pruinosa: 3 veces* *Tamaño de la arcestida: 1 vez* *Número de semillas de la arcestida: 3 veces* *Disposición de las arcestidas: 1 vez* *Forma de la hoja: 1 vez* *Hoja punzante: 1 vez* *Hojas incurvas: 1 vez* *Planta: 2 veces*
(0) Distribución de las hojas Distanciada.....1 (1) Tamaño 1500 cm.....2 (2) Número de franjas blancas en el haz 1; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Tamaño de la arcestida Entre 0.6 y 1 cm; Número de semillas de la arcestida 3 .....Especie *Juniperus communis subsp. communis* (2) Número de franjas blancas en el haz 2; Color de la arcestida Castaño; Arcestida pruinosa No; Tamaño de la arcestida Más de 1 cm; Número de semillas de la arcestida 1-3	

GÉNERO *JUNIPERUS* /CRITERIO DE DALLWITZ
.....Especie *Juniperus oxycedrus subsp. badia*
(1) Tamaño 400 cm
.....Especie *Juniperus oxycedrus subsp. oxycedrus*
(1) Tamaño 300 cm
.....Especie *Juniperus oxycedrus subsp. macrocarpa*
(1) Tamaño 200 cm
.....Especie *Juniperus navicularis*
(0) Distribución de las hojas Densa.....3
(3) Hojas rectas Si
.....Especie *Juniperus communis subsp. hemisphaerica*
(3) Hojas rectas No.....4
(4) Número de franjas blancas en el haz 1; Disposición de las arcestidas Axilares; Forma de la hoja Acicular; Hoja punzante Si; Hojas incurvas Si
.....Especie *Juniperus communis subsp. alpina*
(4) Número de franjas blancas en el haz 0; Disposición de las arcestidas Terminales; Forma de la hoja Escamosa; Hoja punzante No; Hojas incurvas No.....5
(5) Aspecto Árbol.....6
(6) Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....7
(7) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9
.....Especie *Juniperus phoenicea subsp. phoenicea*
(7) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4
.....Especie *Juniperus thurifera*
(6) Tamaño de las hojas (largo) Entre 1 y 2 cm
.....Especie *Juniperus phoenicea subsp. phoenicea*
(6) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm
.....Especie *Juniperus thurifera*
(5) Aspecto Arbusto.....8
(8) Altitud Hasta 1000 m.....9
(9) Tamaño de las hojas (largo) Entre 0.2 y 1 cm.....10
(10) Tamaño 800 cm; Planta Monoica; Color de la arcestida Castaño; Arcestida pruinosa No; Número de semillas de la arcestida 3-9
.....Especie *Juniperus phoenicea subsp. phoenicea*
(10) Tamaño 2000 cm; Planta Dioica; Color de la arcestida Negro-azulado; Arcestida pruinosa Si; Número de semillas de la arcestida 2-4
.....Especie *Juniperus thurifera*
(9) Tamaño de las hojas (largo) Entre 1 y 2 cm
.....Especie *Juniperus phoenicea subsp. phoenicea*
(9) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm
.....Especie *Juniperus thurifera*
(8) Altitud Entre 1000 y 2000 m.....11
(11) Tamaño 800 cm
.....Especie *Juniperus phoenicea subsp. phoenicea*
(11) Tamaño 2000 cm
.....Especie *Juniperus thurifera*
(11) Tamaño 100 cm
.....Especie *Juniperus sabina*
(8) Altitud Nivel del mar
.....Especie *Juniperus phoenicea subsp. turbinata*
(8) Altitud Mas de 2000 m
.....Especie *Juniperus sabina*

Tabla B- XXIX. *Clave para el género Juniperus generada según el criterio de Dallwitz.*

GÉNERO CUPRESSUS.

GÉNERO *CUPRESSUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 3	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Color de la piña joven: 1 vez* *Disposición de las ramas: 1 vez* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Tamaño de las hojas (largo): 1 vez* *Ápice de la hoja agudo: 1 vez* *Hoja de color: 1 vez*
(0) Disposición de las ramas Fastigiadas o ascendentes.....1 (1) Color de la piña joven Verde; Tamaño de las hojas (largo) Hasta 0.1 cm; Ápice de la hoja agudo No; Hoja de color Verde .....Especie *Cupressus sempervirens* (1) Color de la piña joven Glauco; Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm; Ápice de la hoja agudo Si; Hoja de color Glauco .....Especie *Cupressus arizónica* (0) Disposición de las ramas Erecto-patentes.....2 (2) Tamaño de las hojas (largo) Hasta 0.1 cm .....Especie *Cupressus sempervirens* (2) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm .....Especie *Cupressus macrocarpa* (0) Disposición de las ramas Patentes pero péndulas en el ápice .....Especie *Cupressus lusitánica*	

Tabla B- XXX. *Clave para el género Cupressus generada según el criterio de mínima entropía.*

GÉNERO *CUPRESSUS* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave:2	*Total de caracteres confirmadores:0*
Atributos incluidos en la clave: *Tamaño de las hojas (largo): 1 vez* *Disposición de las ramas: 1 vez*	*Caracteres confirmadores incluidos:*
(0) Tamaño de las hojas (largo) Hasta 0.1 cm .....Especie *Cupressus sempervirens* (0) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm.....1 (1) Disposición de las ramas Fastigiadas o ascendentes .....Especie *Cupressus arizónica* (1) Disposición de las ramas Erecto-patentes .....Especie *Cupressus macrocarpa* (1) Disposición de las ramas Patentes pero péndulas en el ápice .....Especie *Cupressus lusitánica*	

Tabla B- XXXI. *Clave para el género Cupressus generada según el criterio de proporción de ganancia.*

GÉNERO *CUPRESSUS* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:5*
Atributos incluidos en la clave: *Color de la piña joven: 1 vez* *Disposición de las ramas: 1 vez* *Tamaño de las hojas (largo): 1 vez* *Tamaño de la piña (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Tamaño de las hojas (largo): 1 vez* *Ápice de la hoja agudo: 1 vez* *Hoja de color: 1 vez* *Escamas de la piña obtusamente mucronadas: 1 vez* *Escamas de la piña cuspidadas: 1 vez*
(0) Tamaño de la piña (largo) Entre 2 y 4 cm; Escamas de la piña obtusamente mucronadas Si; Escamas de la piña cuspidadas No.....1 (1) Disposición de las ramas Fastigiadas o ascendentes.....2 (2) Color de la piña joven Verde; Tamaño de las hojas (largo) Hasta 0.1 cm; Ápice de la hoja agudo No; Hoja de color Verde .....Especie *Cupressus sempervirens* (2) Color de la piña joven Glauco; Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm; Ápice de la hoja agudo Si; Hoja de color Glauco .....Especie *Cupressus arizónica* (1) Disposición de las ramas Erecto-patentes.....3 (3) Tamaño de las hojas (largo) Hasta 0.1 cm .....Especie *Cupressus sempervirens* (3) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm .....Especie *Cupressus macrocarpa* (0) Tamaño de la piña (largo) Entre 1 y 2 cm; Escamas de la piña obtusamente mucronadas No; Escamas de la piña cuspidadas Si .....Especie *Cupressus lusitánica*	

Tabla B- XXXII. *Clave para el género Cupressus generada según el criterio de Gini.*

GÉNERO *CUPRESSUS* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:3	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Color de la piña joven: 1 vez* *Disposición de las ramas: 1 vez* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Tamaño de las hojas (largo): 1 vez* *Ápice de la hoja agudo: 1 vez* *Hoja de color: 1 vez*
(0) Disposición de las ramas Fastigiadas o ascendentes.....1 (1) Color de la piña joven Verde; Tamaño de las hojas (largo) Hasta 0.1 cm; Ápice de la hoja agudo No; Hoja de color Verde .....Especie *Cupressus sempervirens* (1) Color de la piña joven Glauco; Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm; Ápice de la hoja agudo Si; Hoja de color Glauco .....Especie *Cupressus arizónica* (0) Disposición de las ramas Erecto-patentes.....2 (2) Tamaño de las hojas (largo) Hasta 0.1 cm .....Especie *Cupressus sempervirens* (2) Tamaño de las hojas (largo) Entre 0.1 y 0.2 cm .....Especie *Cupressus macrocarpa* (0) Disposición de las ramas Patentes pero péndulas en el ápice .....Especie *Cupressus lusitánica*	

Tabla B- XXXIII. *Clave para el género Cupressus generada según el criterio de Dallwitz.*

GÉNERO EPHEDRA.

GÉNERO EPHEDRA /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:1*
Atributos incluidos en la clave: *Grosor de las ramillas: 1 vez*	*Caracteres confirmadores incluidos:* *Color de las ramillas: 1 vez*
(0) Grosor de las ramillas Gruesas (0.15-0.22 cm); Color de las ramillas Cenicientas .....Especie *Ephedra fragilis* (0) Grosor de las ramillas Delgadas (0.07-0.1 cm); Color de las ramillas Verde-amarillentas .....Especie *Ephedra distachya* (0) Grosor de las ramillas Delgadas (0.04-0.07 cm); Color de las ramillas Castaño oscuro .....Especie *Ephedra nebrodensis*	

Tabla B- XXXIV. *Clave para el género Ephedra generada según el criterio de mínima entropía.*

FAMILIA *TAXODIACEAE* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 2	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño: 1 vez* *Grosor de las ramillas: 1 vez*	*Caracteres confirmadores incluidos:* *Ramas fácilmente desarticulables: 1 vez* *Color de las ramillas: 1 vez*
(0) Tamaño 400 cm; Ramas fácilmente desarticulables Si .....Especie *Ephedra fragilis* (0) Tamaño 100 cm; Ramas fácilmente desarticulables No.....1 (1) Grosor de las ramillas Delgadas (0.07-0.1 cm); Color de las ramillas Verde-amarillentas .....Especie *Ephedra distachya* (1) Grosor de las ramillas Delgadas (0.04-0.07 cm); Color de las ramillas Castaño oscuro .....Especie *Ephedra nebrodensis*	

Tabla B- XXXV. *Clave para el género Ephedra generada según el criterio de proporción de ganancia.*

GÉNERO EPHEDRA /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave: 4	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño: 1 vez* *Grosor de las ramillas: 1 vez* *Tamaño del sincarpo (largo): 1 vez* *Color del sincarpo: 1 vez*	*Caracteres confirmadores incluidos:* *Ramas fácilmente desarticulables: 1 vez* *Color de las ramillas: 1 vez*
(0) Color del sincarpo Rojizo.....1 (1) Tamaño 400 cm; Ramas fácilmente desarticulables Si .....Especie *Ephedra fragilis* (1) Tamaño 100 cm; Ramas fácilmente desarticulables No.....2 (2) Tamaño del sincarpo (largo) 0.5-0.7 cm.....3 (3) Grosor de las ramillas Delgadas (0.07-0.1 cm); Color de las ramillas Verde-amarillentas .....Especie *Ephedra distachya* (3) Grosor de las ramillas Delgadas (0.04-0.07 cm); Color de las ramillas Castaño oscuro .....Especie *Ephedra nebrodensis* (2) Tamaño del sincarpo (largo) 0.3-0.5 cm .....Especie *Ephedra nebrodensis* (0) Color del sincarpo Amarillento .....Especie *Ephedra nebrodensis*	

Tabla B- XXXVI. *Clave para el género Ephedra generada según el criterio de Gini.*

GÉNERO EPHEDRA /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:1*
Atributos incluidos en la clave: *Grosor de las ramillas: 1 vez*	*Caracteres confirmadores incluidos:* *Color de las ramillas: 1 vez*
(0) Grosor de las ramillas Gruesas (0.15-0.22 cm); Color de las ramillas Cenicientas .....Especie *Ephedra fragilis* (0) Grosor de las ramillas Delgadas (0.07-0.1 cm); Color de las ramillas Verde-amarillentas .....Especie *Ephedra distachya* (0) Grosor de las ramillas Delgadas (0.04-0.07 cm); Color de las ramillas Castaño oscuro .....Especie *Ephedra nebrodensis*	

Tabla B- XXXVII. *Clave para el género Ephedra generada según el criterio de Dallwitz.*

GÉNERO ABIES.

GÉNERO *ABIES* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave:1	*Total de caracteres confirmadores:4*
Atributos incluidos en la clave: *Con hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Hojas discoloras: 1 vez* *Disposición de las escamas tectrices: 1 vez* *Color de las ramas: 1 vez* *Ramas: 1 vez*
(0) Con hojas Flexibles; Hojas discoloras Si; Disposición de las escamas tectrices Exertas; Color de las ramas Cenicientas; Ramas Pubescentes .....Especie *Abies alba* (0) Con hojas Rígidas; Hojas discoloras No; Disposición de las escamas tectrices Inclusas; Color de las ramas Castaño-rojizas; Ramas Glabras .....Especie *Abies pinsapo*	

Tabla B- XXXVIII. *Clave para el género Abies generada según el criterio de mínima entropía.*

GÉNERO *ABIES* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:4*
Atributos incluidos en la clave: *Con hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Hojas discoloras: 1 vez* *Disposición de las escamas tectrices: 1 vez* *Color de las ramas: 1 vez* *Ramas: 1 vez*
(0) Con hojas Flexibles; Hojas discoloras Si; Disposición de las escamas tectrices Exertas; Color de las ramas Cenicientas; Ramas Pubescentes .....Especie *Abies alba* (0) Con hojas Rígidas; Hojas discoloras No; Disposición de las escamas tectrices Inclusas; Color de las ramas Castaño-rojizas; Ramas Glabras .....Especie *Abies pinsapo*	

Tabla B- XXXIX. *Clave para el género Abies generada según el criterio de proporción de ganancia.*

GÉNERO *ABIES* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave:1	*Total de caracteres confirmadores:4*
Atributos incluidos en la clave: *Con hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Hojas discoloras: 1 vez* *Disposición de las escamas tectrices: 1 vez* *Color de las ramas: 1 vez* *Ramas: 1 vez*
(0) Con hojas Flexibles; Hojas discoloras Si; Disposición de las escamas tectrices Exertas; Color de las ramas Cenicientas; Ramas Pubescentes .....Especie *Abies alba* (0) Con hojas Rígidas; Hojas discoloras No; Disposición de las escamas tectrices Inclusas; Color de las ramas Castaño-rojizas; Ramas Glabras .....Especie *Abies pinsapo*	

Tabla B- XL. *Clave para el género Abies generada según el criterio de Gini.*

GÉNERO *ABIES* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:1	*Total de caracteres confirmadores:4*
Atributos incluidos en la clave: *Con hojas: 1 vez*	*Caracteres confirmadores incluidos:* *Hojas discoloras: 1 vez* *Disposición de las escamas tectrices: 1 vez* *Color de las ramas: 1 vez* *Ramas: 1 vez*
(0) Con hojas Flexibles; Hojas discoloras Si; Disposición de las escamas tectrices Exertas; Color de las ramas Cenicientas; Ramas Pubescentes .....Especie *Abies alba* (0) Con hojas Rígidas; Hojas discoloras No; Disposición de las escamas tectrices Inclusas; Color de las ramas Castaño-rojizas; Ramas Glabras .....Especie *Abies pinsapo*	

Tabla B- XLI. *Clave para el género Abies generada según el criterio de Dallwitz.*

GÉNERO CYCAS.

GÉNERO CYCAS /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Hojas escurridas sobre el raquis: 1 vez* *Forma de las hojuelas: 1 vez*
(0) Tamaño de las hojas (largo) Hasta 100 cm; Hojas escurridas sobre el raquis No; Forma de las hojuelas Con el borde revuelto .....Especie *Cycas revoluta* (0) Tamaño de las hojas (largo) Mas de 100 cm; Hojas escurridas sobre el raquis Si; Forma de las hojuelas Acintadas y planas .....Especie *Cycas circinalis*	

Tabla B- XLII. *Clave para el género Cycas generada según el criterio de mínima entropía.*

GÉNERO CYCAS /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave:	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Hojas escurridas sobre el raquis: 1 vez* *Forma de las hojuelas: 1 vez*
(0) Tamaño de las hojas (largo) Hasta 100 cm; Hojas escurridas sobre el raquis No; Forma de las hojuelas Con el borde revuelto .....Especie *Cycas revoluta* (0) Tamaño de las hojas (largo) Mas de 100 cm; Hojas escurridas sobre el raquis Si; Forma de las hojuelas Acintadas y planas .....Especie *Cycas circinalis*	

Tabla B- XLIII. *Clave para el género Cycas generada según el criterio de proporción de ganancia.*

GÉNERO CYCAS /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave:1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Hojas escurridas sobre el raquis: 1 vez* *Forma de las hojuelas: 1 vez*
(0) Tamaño de las hojas (largo) Hasta 100 cm; Hojas escurridas sobre el raquis No; Forma de las hojuelas Con el borde revuelto .....Especie *Cycas revoluta* (0) Tamaño de las hojas (largo) Mas de 100 cm; Hojas escurridas sobre el raquis Si; Forma de las hojuelas Acintadas y planas .....Especie *Cycas circinalis*	

Tabla B- XLIV. *Clave para el género Cycas generada según el criterio de Gini.*

GÉNERO CYCAS /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Hojas escurridas sobre el raquis: 1 vez* *Forma de las hojuelas: 1 vez*
(0) Tamaño de las hojas (largo) Hasta 100 cm; Hojas escurridas sobre el raquis No; Forma de las hojuelas Con el borde revuelto .....Especie *Cycas revoluta* (0) Tamaño de las hojas (largo) Mas de 100 cm; Hojas escurridas sobre el raquis Si; Forma de las hojuelas Acintadas y planas .....Especie *Cycas circinalis*	

Tabla B- XLV. *Clave para el género Cycas generada según el criterio de Dallwitz.*

GÉNERO CEDRUS.

GÉNERO *CEDRUS* /ENTROPÍA	
Total de atributos diferentes incluidos en la clave: 1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Guía del árbol: 1 vez* *Ramas colgantes: 1 vez*
(0) Tamaño 6000 cm; Guía del árbol Recurvada; Ramas colgantes Si .....Especie *Cedrus deodara* (0) Tamaño 5000 cm; Guía del árbol No recurvada; Ramas colgantes No .....Especie *Cedrus atlántica*	

Tabla B- XLVI. *Clave para el género Cedrus generada según el criterio de mínima entropía.*

GÉNERO *CEDRUS* /PROPORCIÓN DE GANANCIA	
Total de atributos diferentes incluidos en la clave:1	*Total de caracteres confirmadores:2*
Atributos incluidos en la clave: *Tamaño: 1 vez*	*Caracteres confirmadores incluidos:* *Guía del árbol: 1 vez* *Ramas colgantes: 1 vez*
(0) Tamaño 6000 cm; Guía del árbol Recurvada; Ramas colgantes Si .....Especie *Cedrus deodara* (0) Tamaño 5000 cm; Guía del árbol No recurvada; Ramas colgantes No .....Especie *Cedrus atlántica*	

Tabla B- XLVII. *Clave para el género Cedrus generada según el criterio de proporción de ganancia*

GÉNERO *CEDRUS* /INDICE DE DIVERSIDAD DE GINI	
Total de atributos diferentes incluidos en la clave:2	*Total de caracteres confirmadores:3*
Atributos incluidos en la clave: *Tamaño: 2 veces* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:* *Guía del árbol: 2 veces* *Ramas colgantes: 2 veces* *Hoja de color: 2 veces*
(0) Tamaño de las hojas (largo) Entre 2 y 2.5 cm.....1 (1) Tamaño 6000 cm; Guía del árbol Recurvada; Ramas colgantes Si; Hoja de color Verde .....Especie *Cedrus deodara* (1) Tamaño 5000 cm; Guía del árbol No recurvada; Ramas colgantes No; Hoja de color Verde-glauco .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 2.5 y 3 cm.....2 (2) Tamaño 6000 cm; Guía del árbol Recurvada; Ramas colgantes Si; Hoja de color Verde .....Especie *Cedrus deodara* (2) Tamaño 5000 cm; Guía del árbol No recurvada; Ramas colgantes No; Hoja de color Verde-glauco .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 3 y 5 cm .....Especie *Cedrus deodara* (0) Tamaño de las hojas (largo) Entre 1 y 2 cm .....Especie *Cedrus atlántica*	

Tabla B- XLVIII. *Clave para el género Cedrus generada según el criterio de Gini.*

GÉNERO *CEDRUS* /CRITERIO DE DALLWITZ	
Total de atributos diferentes incluidos en la clave:2	*Total de caracteres confirmadores:0*
Atributos incluidos en la clave: *Tamaño de la piña (largo): 2 veces* *Tamaño de las hojas (largo): 1 vez*	*Caracteres confirmadores incluidos:*
(0) Tamaño de las hojas (largo) Entre 2 y 2.5 cm.....1 (1) Tamaño de la piña (largo) Entre 8 y 12 cm .....Especie *Cedrus deodara* (1) Tamaño de la piña (largo) Entre 4 y 6 cm .....Especie *Cedrus atlántica* (1) Tamaño de la piña (largo) Entre 6 y 8 cm .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 2.5 y 3 cm.....2 (2) Tamaño de la piña (largo) Entre 8 y 12 cm .....Especie *Cedrus deodara* (2) Tamaño de la piña (largo) Entre 4 y 6 cm .....Especie *Cedrus atlántica* (2) Tamaño de la piña (largo) Entre 6 y 8 cm .....Especie *Cedrus atlántica* (0) Tamaño de las hojas (largo) Entre 3 y 5 cm .....Especie *Cedrus deodara* (0) Tamaño de las hojas (largo) Entre 1 y 2 cm .....Especie *Cedrus atlántica*	

Tabla B- XLIX. *Clave para el género Cedrus generada según el criterio de Dallwitz.*

Apéndice C. Introducción al lenguaje XML.

4. ¿Qué es XML?

XML significa *eXtensible markup language*, o **lenguaje de anotación extensible**. Ya conocemos el lenguaje HTML (*hypertext markup language*), lenguaje de anotación para páginas web que permite navegación tipo hipertexto; sin embargo, XML no es sólo un lenguaje, es una forma de especificar lenguajes, de ahí lo de extensible. Por lo tanto, XML no es un lenguaje para hacer mejores páginas web, sino un lenguaje para información auto-descrita.

5. Historia de XML.

XML se inició como un subconjunto de SGML (S*tructured Generalized Markup Language*), un estándar ISO para documentos estructurados sumamente complejo para servir documentos en la web. XML es algo así como SGML simplificado.

Este lenguaje tiene gran número de aplicaciones. La mayor parte de los portales y sitios de noticias ya están basados en XML, porque permite estructurar la información y posteriormente aplicar de forma sencilla transformaciones para presentarlo. Lo habitual es que la información almacenada en una base de datos se convierta a XML y luego se transforme para servirla al cliente.

6. XML bien formado.

Como lenguaje de anotación, las sentencias en XML consisten en una serie de etiquetas (llamadas **elementos**) con una serie de modificadores (llamados **atributos**). Las etiquetas pueden estar anidadas unas dentro de otras, pero toda etiqueta que se abra se tiene que cerrar, y siempre en el mismo orden. En caso de que un elemento no tenga pareja (por no tener ningún contenido dentro), se le denomina elemento vacío y se indica con un "/" al final. Los elementos se agrupan en documentos, tales como el del Ejemplo C- I que describe las estancias de una casa.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <micasa>
   - <habitacion id="comedor">
       <mueble>aparador</mueble>
       <mueble>sofá</mueble>
       <puerta a="balcón" />
    </habitacion>
</micasa>
```

Ejemplo C- I. *Ejemplo básico de XML.*

Todos los documentos XML deben estar **bien formados**, eso que significa que se debe cumplir lo siguiente:

- Si no se utiliza DTD, el documento debe comenzar con una declaración de documento *standalone*, tal como la que se pone en la primera línea del ejemplo.
- Todas las etiquetas deben estar equilibradas: esto es, todos los elementos que contengan datos de tipo carácter deben tener etiquetas de principio y fin.
- Todos los valores de los atributos deben ir entrecomillados.

- Cualquier elemento vacío (por ejemplo, aquellos que no tienen etiqueta final como "*<IMG>*", "*<HR>*", y "*
" y otros de HTML) deben terminar con "/>*" o convertirse en "no vacíos" añadiéndoles una etiqueta de fin.
- No debe haber etiquetas aisladas ("<" ó "*&*") en el texto (p.e. debe escribirse como "*<*" y "*&*"), y la secuencia "*]]>*" debe escribirse como "*]]>*" si no ocurre esto como final de una sección marcada como CDATA.
- Los elementos deben anidar dentro de sí sus propiedades.
- Los ficheros bien-formados sin DTD pueden utilizar atributos en sus elementos, pero éstos deben ser todos del tipo CDATA, por defecto. El tipo CDATA (*character DATA*) son caracteres.
- Los nombres de las etiquetas pueden ser alfanuméricos, comenzando con una letra, e incluyendo los caracteres "-" y ":", aunque este último tiene un significado especial.

En un documento XML, aparte de elementos y atributos, puede haber otros elementos como

- **Entidades**. Representan símbolos "atómicos" que habitualmente deben ser entendidos por el navegador. Las entidades van encerradas entre los símbolos "*&*" y ";". LaTabla C- Imuestra algunos ejemplos de entidades XML.

ENTIDAD	CARACTER
&amp;	&
&lt;	<
&gt;	>
&apos;	'
&quot;	"

Tabla C- I. *Entidades XML.*

- **Comentarios**, que se procesan de forma diferente al texto, y que, tal como en HTML, van precedidos por "*<!--* " y acaban con "*-->*".
- **Secciones CDATA**. Sirven para extraer del documento XML una sección que va a ser interpretada tal cual, sin hacer ninguna modificación. Puede servir, por ejemplo, para incluir HTML "mal-formado" dentro de un documento XML. El Ejemplo C- II incluye todos los elementos descritos.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
      - <!-- Descripción de los elementos de una casa soñada  -->
- <micasa>
   - <habitacion id="comedor">
      <mueble>aparador</mueble>
      <mueble>sofá "de época"</mueble>
    </habitacion>
   - <habitacion id="cocina">
      - <mueble>
              - <![CDATA[
               <p>En la pared de la derecha hay un frigorífico <p>Y en
               la de la izquierda, sólo mugre
          ]]>
      </mueble>
      <mueble>fregadero</mueble>
   </habitacion>
</micasa>
```

Ejemplo C- II. *Ejemplo de documento XML completo.*

7. XML *namespaces.*

Si todo el mundo definiera sus etiquetas, un documento acabaría siendo un caos de diferentes etiquetas procedentes de diferentes sitios, y, lo que es peor, de etiquetas con el mismo nombre que, en realidad, significan cosas diferentes. El concepto de **espacios de nombres** (*namespaces*) permite dividir el conjunto de todos los nombres posibles, de forma que se pueda definir a qué zona de ese espacio corresponde una etiqueta. De este modo, etiquetas con el mismo nombre, pero definidas por dos autores diferentes, pueden diferenciarse en el espacio de nombres. El espacio de nombres no es esencial en todos los documentos, pero resulta útil cuando se usan etiquetas de diferente procedencia (por ejemplo, etiquetas nuevas dentro de un documento XML), o etiquetas que se quieren procesar de forma diferente. El espacio de nombres de una etiqueta se indica con un prefijo y "*:: <namespace:etiqueta*" (ver Ejemplo C- III).

```
- <mc:micasa xmlns:mc="http://www.geneura.org/micasa">
   - <mc:habitacion mc:id="comedor">
      <mc:mueble>aparador</mc:mueble>
      <mc:mueble>sofá "de época"</mc:mueble>
   </mc:habitacion>
</mc:micasa>
```

Ejemplo C- III. *Ejemplo de documento XML con namespaces.*

En este documento, se utiliza la primera línea para declarar el prefijo del espacio de nombres mediante el atributo "*xmlns*" (XML *namespace*). En este caso, hemos elegido el prefijo "*mc*". El espacio de nombres tiene que tener asignado un URI (*Universal Resource Identification*), que es un identificador único en el documento.

Un documento XML puede tener tantos espacios de nombres como se quieran declarar, y se pueden mezclar elementos de diferentes espacios de nombres, e incluso sin ningún espacio

8. XML y diccionarios de datos.

En algunos casos, es necesario validar que un documento XML es correcto, es decir, que las etiquetas que se usan son correctas y que están anidadas de la forma adecuada. Para ello se pueden utilizar dos herramientas:

- **DTD, o *data type dictionnary***. El DTD del Ejemplo C- IV indica es que una habitación tiene uno o varios muebles, y una o varias puertas (que se indica con "+"); a su vez, "*micasa*" puede tener una o más habitaciones, y cada uno de los elementos pueden tener los atributos que se indican con la sentencia *ATTLIST*. Como se puede ver, no se trata de XML, aunque se parezca.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT mc:habitacion (mc:mueble+)>
<!ATTLIST mc:habitacion
    mc:id CDATA #REQUIRED
>
<!ELEMENT mc:micasa (mc:habitacion)>
<!ATTLIST mc:micasa
    xmlns:mc CDATA #REQUIRED
>
<!ELEMENT mc:mueble (#PCDATA)>
```

Ejemplo C- IV. *Ejemplo de DTD.*

- **XSchema**, el equivalente a un DTD, pero en XML. Un XSchema describe la sintaxis correcta de un documento XML. El esquema del Ejemplo C- V es el equivalente al DTD del ejemplo anterior.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--W3C Schema generated by XMLSPY v2004 rel. 2 U (http://www.xmlspy.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
    <xs:complexType name="mc:habitacionType">
        <xs:sequence>
            <xs:element name="mc:mueble" ref="mc:mueble" minOccurs="1"
maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:attribute name="mc:id" type="xs:string" use="required"/>
    </xs:complexType>
    <xs:element name="mc:micasa">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="mc:habitacion" type="mc:habitacionType"/>
            </xs:sequence>
            <xs:attribute name="xmlns:mc" type="xs:anyURI" use="required"/>
        </xs:complexType>
    </xs:element>
    <xs:element name="mc:mueble" type="xs:string"/>
</xs:schema>
```

Ejemplo C- V. *Ejemplo de esquema XML.*

9. Lectura y manipulación de un documento XML.

El W3C [*World Wide Web Consortium*, http://www.w3.org] ha especificado dos normas para trabajar con documentos XML. Esta especificación permite que cualquier lenguaje de programación pueda navegar a través de un documento XML y manipularlo.

La importancia de estas normas es inmensa. Si cada lenguaje de programación implementase su vía especial para navegar a través de los documentos XML, sería caótico: ciertos lenguajes solo permitirían realizar ciertas operaciones y no habría consistencia alguna. Las normas creadas por el W3C aportan una interfaz independiente del lenguaje de programación en todos los documentos XML bien formados. Puesto que la interfaz está entre el documento XML y una aplicación es conocida como una API (*Application Program Interface,* interfaz de programación de la aplicación). Estas API describen una jerarquía de objetos, con métodos y atributos que simplifican las tareas relativas al recorrido y acceso a las partes del documento. Estos dos mecanismos se

denominan **SAX** (*Simple API for XML Parsing*) y **DOM** (*Document Object Model*).

- SAX se utiliza para hacer un recorrido secuencial de los elementos del documento XLM. Por este motivo, utiliza mucha menos memoria que DOM, es más rápido porque no invierte tiempo construyendo el árbol y el código basado en este modelo y es escalable a cualquier cantidad arbitraria de información[16].
- DOM implica la creación de un árbol en memoria que contiene todo el documento XML. Un nodo DOM de nivel 1 contiene 9 punteros a otros nodos. Cada puntero ocupa 4 bytes lo que hace un total de 36 bytes de punteros por nodo. Además de esta, contiene otros tipos de información como el nombre y el tipo del nodo.

[16] Los documentos SDD pueden ser excesivamente grandes. De hecho, el programa *DAtoSDD* ha generado documentos XML, relacionados con proyectos reales, de 16MB. Los documentos XML son documentos de texto, no es difícil imaginar que representar 16MB de texto en una estructura de árbol con sus correspondientes enlaces en memoria puede dejar a nuestra máquina sin memoria. Esto conduce a optar por el modelo SAX para realizar la lectura / escritura del documento.

Apéndice D. Pruebas experimentales realizadas al sistema GREEN.

Nombre del carácter	Valor del carácter	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
Experimento		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Aspecto	Arbusto					X												X	X								X		X		
	Arbol		X	X	X		X	X	X	X	X	X	X	X	X	X	X			X				X	X	X		X		X	
	Palmera	X																													
Tipo de arbusto	Postrado					X																					X				
	Erecto				X																			X							
Tallo articulado	Si																	X	X												
	No		X																					X			X	X	X		
Forma de la copa	Aparasolada																								X						
	Piramidal		X	X				X		X	X													X							
Guia del arbol	No recurvada																							X	X						
	Recurvada																														
Tamaño	100 cm																														
	1500 cm																														
	800 cm																														
	2000 cm																														
	4000 cm				X																										
	2500 cm																														
	5000 cm																							X							
	3000 cm																														
	400 cm																														
	6000 cm																														

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	200 cm																														
	250 cm																														
	300 cm																														
Tronco de 10 a 12 m de diametro	Si																														
	No																							X	X						
Color de la corteza (ritidoma)	Gris-ceniciento																														
	Pardo-rojizo																								X						
Corteza rojiza al desprenderse en placas	Si																								X						
	No																														
Resinosa	Si		X									X	X	X	X	X	X													X	
	No	X																													
Tamaño de las hojas (ancho)	Entre 0.1 y 0.2 cm									X	X										X	X		X							
	Hasta 0.1 cm							X	X											X					X						
	Mas de 0.2 cm																						X								
Tamaño de las hojas (largo)	Entre 15 y 20 cm														X																
	Entre 20 y 25 cm																														
	Entre 3 y 5 cm															X															
	Entre 2.5 y 3 cm								X											X											

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	Entre 7 y 15 cm											X		X											X						
	Entre 5 y 7 cm												X	X			X														
	Entre 1 y 2 cm							X		X	X												X	X							
	Mas de 100 cm	X																													
	Hasta 0.1 cm						X																								
	Entre 0.1 y 0.2 cm																														
	Entre 0.2 y 1 cm					X															X	X									
	Entre 2 y 2.5 cm										X																				
	Hasta 100 cm																														
Numero de hojas por fasciculo	2											X	X				X								X						
	3													X																	
Distribucion de las hojas	Distanciada																														
	Densa																			X	X	X				X	X	X	X	X	
Hojas alternas o en espiral	Si	X						X	X	X	X										X			X							
	No																					X									
Hojas discoloras	No							X		X										X	X										
	Si					X					X																				
Enves con dos bandas	No					X				X														X							
	Si										X												X								
Numero de	2																														

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
franjas blancas en el haz	1					X																									
	0																														
Hoja plana	No																														
	Si							X			X									X			X	X		X					
Hojas incurvas	No																														
	Si																				X										
Hojas rectas	Si					X				X										X				X							
	No																														
Hoja punzante	No				X			X		X	X									X			X	X							
	Si					X			X			X		X	X						X	X									
Con hojas	Rigidas									X												X	X								
	Flexibles										X									X				X							
Apice de la hoja agudo	No							X				X																			
	Si					X														X		X	X	X							
Hojas escurridas sobre el raquis	No																			X											
	Si																							X							
Forma de las hojuelas	Acintadas y planas																														
	Con el borde revuelto	X																													
Disposicion de las ramas	Erecto-patentes																														
	Fastigiadas o ascendentes																													X	

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	Patentes pero pendulas en el apice																														
Ramas facilmente desarticulables	No																														
	Si			X														X	X												
Color de las ramas	Cenicientas			X																		X	X								
	Castaño-rojizas											X																			
Ramas colgantes	Si																														
	No																														
Ramas	Glabras																	X	X	X		X	X								
	Pubescentes																														
Presencia de braquiblastos	Si								X			X	X	X	X	X	X			X					X						
	No									X	X										X	X		X							
Braquiblastos	Desarrollados con muchas hojas fasciculadas								X											X											
	Con 2 a 5 hojas											X	X	X	X	X	X								X						
Con raices aereas	No																	X	X	X	X	X	X	X							
	Si																														
Grosor de las ramillas	Delgadas (0.04-0.07 cm)																														X

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	Delgadas (0.07-0.1 cm)																														
	Gruesas (0.15-0.22 cm)																	X	X												
Color de las ramillas	Pardo-rojizas o castañas																														
	Verde-amarillentas		X															X													
	Cenicientas			X																											
	Castaño oscuro																		X												X
Disposicion de las ramillas en un solo plano	Si		X																												
	No			X	X													X	X	X							X	X	X	X	
Consistencia de la fructificacion	Leñosa		X	X			X	X				X	X	X	X	X	X	X		X	X		X							X	
	Carnosa				X	X												X						X			X	X	X		
Tamaño del sincarpo (largo)	0.7-0.9 cm																														
	0.5-0.7 cm																														
	0.3-0.5 cm																	X													
Color del sincarpo	Rojizo																	X													X
	Amarillento																														
Semilla con arilo	No																										X	X			
	Si																							X							
Semillas numerosas	Si		X					X				X			X		X			X	X	X									
	No						X																X	X							

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
Disposicion de las piñas	Patentes o erecto-patentes												X																		
	Erectas									X	X																				
	Colgantes							X																							
Piñas con pedunculo evidente	Si												X							X											
	No														X	X	X						X								
Piñas en ramas superiores de la copa	No																														
	Si																														
Tamaño de la piña (ancho)	Mas de 7.5 cm																														
	Entre 4 y 7.5 cm							X							X																
	Entre 2 y 4 cm															X	X			X			X								
Tamaño de la piña (largo)	Hasta 1 cm		X	X			X																								
	Entre 8 y 12 cm							X																							
	Entre 6 y 8 cm												X																		
	Mas de 18 cm																														
	Entre 4 y 6 cm																														
	Entre 2 y 4 cm															X	X			X										X	
	Entre 15 y 18 cm																														
	Entre 1 y 2 cm		X	X																		X									

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	Entre 12 y 15 cm														X																
Color de la piña joven	Glauco			X																											
	Verde																													X	
Piñas brillantes	Si																X														
	No			X									X		X	X															
Numero de escamas en la piña	Entre 6 y 8		X				X																								
	Hasta 4			X																											
	Mas de 8							X				X	X		X		X			X		X									
Escamas de la piña peltadas	Si																					X								X	
	No																														
Escamas de la piña cuspidadas	Si						X																								
	No																			X											
Escamas de la piña planas	Si			X				X												X											
	No																														
Escamas de la piña obtusamente mucronadas	Si																														X
	No																			X											
Escamas de la piña recurvadas, con 4-6 esquinas	No																														
	Si																														
Escamas tectrices de la piña sobresalientes	No							X																							
	Si																														

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
Escamas de la piña caducas al madurar	No							X												X										
	Si																														
Disposicion de las escamas tectrices	Exertas																														
	Inclusas							X																							
Presencia de apofisis	No																			X										
	Si														X	X	X														
Caracteristicas de la apofisis	Prominente y punzante														X																
	Poco prominente											X	X			X															
	Poco convexa																														
	Convexa											X																			
	Muy prominente, ganchuda																X														
	Muy prominente y punzante																X														
	Prominente																														
Semilla alada y persistente	Si			X				X																							
	No																														
Con piñon	No		X				X	X							X		X						X								
	Si																														
Color de la	Negro-azulado				X	X																					X	X			

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
arcestida	Castaño																												X		
Arcestida pruinosa	No																														
	Si				X	X																						X			
Tamaño de la arcestida	Hasta 0.6 cm																														
	Entre 0.6 y 1 cm				X	X																					X	X			
	Más de 1 cm				X																								X		
Disposicion de las arcestidas	Terminales				X																						X	X	X		
	Axilares					X																									
Numero de semillas de la arcestida	1-3					X																									
	2-4				X																										
	3																														
	3-9																														
Planta	Dioica	X				X																						X			
	Monoica				X		X			X	X	X		X	X	X	X						X					X		X	
Altitud	Entre 1000 y 2000 m				X																	X									
	Mas de 2000 m																														
	Nivel del mar			X																	X										
	Hasta 1000 m				X													X													
En dunas y arenales	No																														
	Si																											X			

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
Base del tronco	No ensanchada																							X							
	Ensanchada																														
Corteza fibrosa	Si																														
	No																					X									
Forma de la hoja	Linear (planoaguzada)							X		X	X												X	X							
	Acicular					X		X	X	X	X	X	X	X	X	X	X														
	Escamosa curvada hacia el ápice																				X										
	Escamosa		X	X	X		X																				X	X	X	X	
	En forma de palmera	X																													
	Aleznada																					X									
	En forma de abanico, con escotadura central																									X					
Disposicion de las hojas	Hojas agrupadas en la terminación de brotes laterales (braquiblastos)																														
	Imbricadas		X	X	X		X														X	X					X	X	X	X	

Nombre del carácter	**Valor del carácter**	*Cycas revoluta*	*Platycladus orientalis*	*Tetraclinis articulata*	*Juniperus Thuriera*	*Juniperus communis*	*Chamaeciparis lawsoniana*	*Picea abies*	*Cedrus atlantica*	*Abies pinsapo*	*Abies alba*	*Pinus nigra*	*Pinus halepensis*	*Pinus canariensis*	*Pinus pinaster*	*Pinus sylvestris*	*Pinus uncinata*	*Ephedra scoparia*	*Ephedra fragilis*	*Larix decidua*	*Araucaria eterophylla*	*Sequoiadendron giganteum*	*Sequoia serpemvirens*	*Taxus baccata*	*Pinus pinea*	*Ginkgo biloba*	*Juniperus sabina*	*Juniperus thurifera*	*Juniperus phoenicea subs.*	*Cupressus serpemvirens*	*Ephedra nebrodensis*
	En los nudos del tallo																														
	En espiral	X						X		X	X												X	X							
	Fasciculadas								X			X	X	X	X	X	X			X											
	Formando una corona apical																														
	En verticilos de cuatro																														
	Ternadas					X																									
Hoja	Persistente	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X				X	X	X	X			X	X	X	X	
	Caduca																	X		X											
	Más o menos caediza																														
Hojas dispuestas en espiral en los macroblastos	No																						X								
	Si																			X											
Hoja de color	Verde-glauco						X		X																						
	Verde oscuro	X												X						X	X										
	Verde intenso																							X							
	Verde claro		X	X	X	X	X			X		X				X						X									
	Verde														X															X	
	Glauco			X																											

Glosarios.

Glosario de términos botánicos

-B-

BASE DE DATOS TAXONÓMICA: Base de datos en la que el criterio de organización es estrictamente taxonómico.

BIODIVERSIDAD: Variabilidad de organismos vivos en la Tierra. Incluye, tanto los ecosistemas terrestres, aéreos, marinos, acuáticos y otros complejos biológicos, como la diversidad dentro de cada especie, entre las especies y entre los ecosistemas.

BIOLOGÍA EVOLUTIVA: Estudio de los parentescos e historia de los seres vivos.

-C-

CLADISMO: Sistema de clasificación basado en los postulados de la filogenia.

CLASIFICACIÓN: Es la elaboración de un sistema lógico que agrupe a las plantas que presentan caracteres comunes. Las especies similares se agrupan en géneros, estos en familias, órdenes, etc.

CODON: Triplete de nucleótidos del DNA que codifica un aminoácido.

COLECCIONES BIOLÓGICAS: Conjunto de muestras de origen biológico perfectamente conservados, identificados y ordenadas que constituyen un registro permanente de la biodiversidad.

-E-

ESPECIE: Conjunto de individuos semejantes y aislados genéticamente de otros grupos próximos, que se pueden cruzar entre sí dando descendencia fértil. Es la categoría básica sobre la que se construyen las clasificaciones.

-F-

FILOGENIA: Estudio de la historia evolutiva de los seres vivos.

-G-

GIMNOSPERMAS: División del reino vegetal que comprende las plantas con semillas primitivas, o sea, con los óvulos desnudos insertos en hojas carpelares que no se han soldado en un ovario.

-I-

IDENTIFICACIÓN, DETERMINACIÓN: Reconocimiento de los caracteres de la planta, a la que se le aplica un nombre que ha sido dado con anterioridad a una planta similar.

INDIVIDUO: Cada ser organizado respecto de la especie a que pertenece.

-M-

MICORRIZAS: Simbiosis entre un hongo y las raíces de las plantas superiores, de la cual salen ambos componentes beneficiados.

-S-

SINONIMIA: Nombre incorrecto para un taxon. Generalmente es un nombre legítimo y válidamente publicado pero posterior al nombre correcto.

SISTEMÁTICA: Ciencia que se ocupa de la clasificación de los seres vivos. Se suele usar como equivalente a Taxonomía.

SUBESPECIE: Conjunto de individuos separados del resto de la especie por un conjunto de caracteres heredables, y que están aislados en el tiempo o en el espacio.

-T-

TAXA: Conjunto de categorías que componen la clasificación de los seres vivos.

TAXON: Cada una de las categorías o subdivisiones de la clasificación de los seres vivos, que se ordenan según jerarquías, por ejemplo: especie, género, familia, etc.

TAXONOMÍA: Ciencia que se ocupa de la descripción, denominación y clasificación de los seres vivos.

-V-

VARIABILIDAD INTRA-TAXON: Conjunto de caracteres que pueden diferir entre individuos o poblaciones pertenecientes a una misma categoría taxonómica; por ejemplo, diferencias apreciadas en el seno de un mismo género o de una especie**.**

VARIABILIDAD INTER-TAXON: Subdivisión de una categoría taxonómica en rangos inferiores; por ejemplo, una especie en subespeces, variedades o formas.

VARIEDAD: Semejante a la subespecie, pero los caracteres que definen la variación no son heredables de forma constante.

Glosario de términos informáticos

-A-

ADQUISICIÓN Y ELICITACIÓN DEL CONOCIMIENTO: El proceso de adquisición del conocimiento de un sistema experto se compone de dos fases, la elicitación del conocimiento y la representación del conocimiento. A través de la elicitación, el ingeniero del conocimiento obtiene y depura el conocimiento experto a partir de varias fuentes de información [González & Dankel, 1993].

APRENDIZAJE AUTOMÁTICO: Subcampo de la Inteligencia Artificial que se ocupa de aquellos programas capaces de aprender a partir de la experiencia.

-B-

BASE DE CONOCIMIENTO: Módulo de un sistema experto que contiene el conocimiento sobre el dominio de aplicación.

BASE DE DATOS DISTRIBUIDA: Base de datos instalada en un entorno de red en el que sus componentes residen en más de un sistema. También permite el acceso, modificación y actualización de los datos que contiene, desde cualquiera de los sistemas conectados, de manera sincronizada.

BASE DE DATOS FEDERADA: Un sistema de base de datos federadas consta de componentes que son autónomos aún siendo participantes de una federación. Permiten la distribución parcial y controlada de sus datos. No hay un control centralizado, pues cada componente mantiene el control de sus datos.

-C-

CGI: *Common Getaway Interface*. Interfaz de intercambio de datos estándar en Internet a través del cual se organiza el envío de recepción de datos entre navegadores y programas residentes en servidores web.

COMPLETITUD: Propiedad de una base de conocimiento que garantiza que a cada objetivo conduce al menos una regla (no hay objetivos perdidos) [Suwa *et al.*, 1985]. Otros enfoques también incluyen la detección de atributos sin referenciar, valores ilegales de atributos y condiciones inalcanzables. [Nguyen *et al.*, 1987].

CONSISTENCIA: Propiedad de una base de conocimiento que garantiza que cada regla es coherente con el resto.

-D-

DISCRETIZAR: Dividir los valores de un atributo continuo en un conjunto de intervalos adyacentes.

-E-

ENCADENAMIENTO DE REGLAS: Cuando las premisas de algunas reglas coinciden con las conclusiones de otras, se produce el encadenamiento de reglas.

-F-

FACTOR DE CERTEZA: Medida, generalmente comprendida entre –1 y 1 que refleja el nivel de creencia en una hipótesis dada la información disponible.

-G-

GIS: *Geographical Information System*. Sistema de Información Geográfica que busca referenciar una base de datos con la cartografía de un territorio, ligando ambos conceptos.

-K-

KDD, DATA-MINING: *Knowledge Discovery in Databases*. Proceso de descubrimiento de conocimiento en grandes volúmenes de datos, por ejemplo, bases de datos. Si bien el nombre con el que apareció esta área de investigación fue el de KDD, en la actualidad este nombre ha sido sustituido por el de *Data Mining*.

-M-

MEMORIA DE TRABAJO: Módulo de un sistema experto que contiene los hechos descubiertos durante la sesión de consulta al sistema.

MOTOR DE INFERENCIA: Módulo de un sistema experto que combina los hechos de la memoria de trabajo con el conocimiento de la base de conocimiento para inferir conclusiones sobre el problema.

-N-

NP-COMPLETO: Se dice que un problema es NP-completo cuando no se puede resolver con un algoritmo de tiempo polinomial. De forma muy general, se trata de problemas que no se pueden resolver de forma exacta en un tiempo de ejecución razonable.

-P-

PROPORCIÓN DE GANANCIA

RAZONAMIENTO HACIA ATRÁS (BACKWARD CHAININING): Este método de razonamiento que va de una posible solución hacia atrás en las premisas para determinar si los datos admiten esta solución.

-R-

RAZONAMIENTO HACIA DELANTE (FORWARD CHAINING): Proceso de solución comienza recogiendo información que será utilizada para inferir conclusiones lógicas. Este tipo de razonamiento es modelado en un sistema experto mediante el encadenamiento de reglas hacia delante.

REFORZADOR DE CONSISTENCIA: Módulo o conjunto de módulos del sistema encargados de la validación y mantenimiento de la consistencia de la base de conocimiento.

REGLA: Método de representación del conocimiento que conecta uno o más antecedentes contenidos en la parte IF con uno o más consecuentes contenidos en la parte THEN expresando una relación causa-efecto. Una regla puede tener varias premisas unidas con conectores AND u OR o una combinación de ambas. Lo mismo sucede con la conclusión.

REGLA DE DIVISIÓN: Regla heurística utilizada como criterio para ramificar un nodo de un árbol de decisión.

REGLAS CIRCULARES: Conjunto de reglas que especifican una secuencia de razonamiento circular.

REGLAS CON CONDICIONES IF INNECESARIAS: Dos reglas tienen una condición IF innecesaria tienen la misma conclusión y antecedentes, salvo uno que es contradictorio.

REGLAS CONFLICTIVAS: Conjunto de reglas con premisas idénticas y conclusiones contradictorias.

REGLA INALCANZABLE: Regla cuyos antecedentes nunca se pueden satisfacer.

REGLA PERDIDA: Regla cuya conclusión no es alcanzable por ninguna regla.

REGLAS REDUNDANTES: Conjunto de reglas que tiene exactamente los mismos antecedentes, la misma conclusión y el mismo factor de certeza.

REGLA SIN SALIDA: Una regla se considera sin salida cuando su conclusión no es ni una conclusión intermedia ni final.

REGLAS SUBSUMIDAS: Una regla se considera subsumida por otra si tiene un antecedente más extenso y la misma conclusión.

-S-

SHELL: Es un sistema experto genérico, sin base de conocimiento, de modo que puede particularizarse para cualquier dominio de aplicación. Las *shells* comerciales incorporan opciones adicionales como diversos métodos de representación del conocimiento, de tratamiento de incertidumbre y razonamiento [Grzymala-Busse, 1991].

SISTEMA CLIENTE/SERVIDOR: Modelo lógico de una forma de proceso cooperativo, independiente de plataformas hardware y sistemas

operativos. El concepto se refiere más a una filosofía que a un conjunto determinado de productos. Generalmente, el modelo se refiere a un puesto de trabajo o cliente que accede mediante una combinación de hardware y software a los recursos situados en un ordenador denominado servidor.

SISTEMA EXPERTO: Programas de Inteligencia Artificial que consiguen una capacidad similar a la de un experto en la resolución de problemas mediante la reproducción de un cuerpo de conocimiento.

-X-

XML: Acrónimo de *eXtensible Markup Language*, o lenguaje de anotación extensible. Realmente XML es un lenguaje de meta-marcado que se puede utilizar para la especificar lenguajes de marcado.

Bibliografía.

[1] Bartley, M.; Cross, N. 2000. "*Navikey 2.0"*. [En linea]. [Consulta: 03/06/2003]. <http://www.huh.harvard.edu/databases/legacy/navikey/index.html>.

[2] Batchelor, W. D., R. W. McClendon, D. B. Adams, and J. W. Jones. 1989. Evaluation of SMARTSOY: An Expert Simulation System for Insect Pest Management. *Agricultural Systems* 31 (1): 67-81.

[3] Berendsohn, W. G. 1997. A taxonomic information model for botanical databases: The IOPI model. *Taxon*, 46 (2): 283-309.

[4] Berendsohn, W. G. 2001. Biodiversity Informatics. In: *Contributions to Global Change Research: A Report by the German National Committee on Global Change Research*, Bonn, pp 89-94.

[5] Bisby, F. A. 1998. Putting names to things and keeping track: the Species 2000 programme for a coordinate catalogue of life. In: *Information*

technology, plant pathology & biodiversity. Bridge, Jeffries, Morse & Scott. Oxon, New York.

[6] Boddy, L.; Morris, C. W.; Morgan, A. 1998. Development of artificial neural networks for identification. In: *Information technology, plant pathology and biodiversity*. Bridge, P., Jeffries, P., Morse D.R. & Scott, P.R. (Eds.). Wallingford, US & New York, USA, CAB International.

[7] Breiman, L; Friedman, J.H.; Olshen, R.A.; Stone, C.J. 1984. *Classification and Regression Trees.* Wadsworth, California, USA.

[8] Buchanan, B. G.; Shortlife E. H. 1984. *Rule-Based Expert Systems*. Reading, MA: Addison-Wesley.

[9] Cabrero-Carnosa, M.; Castro-Pereiro, M.; Graña-Ramos, M.; Hernández-Pereira, E.; Moret-Bonillo, V.; Martín-Egaña, M.; Verea-Hernando, H. 2003. An intelligent System for the detection and interpretation of sleep apneas. *Expert Systems with Applications* 24, 335-349.

[10] Castroviejo, S. 2002. Riqueza florística de la Península Ibérica e Islas Baleares. El proyecto "Flora Ibérica". En: Pineda, F.D. *et al*. (2002). *La Diversidad Biológica de España,* 167-174. Madrid, Pearson Educación, s. a. 432 pp.

[11] CBIT (Centre for Biological Information Technology, University of Queensland). 1994. *"LucID version 2.1"*. [En línea]. [Consulta: 25/05/2003]. <http://www.lucidcentral.com/default.htm>.

[12] Conruyt, N.; Grosser, D.; Ralambondrainy, H. 1997. IKBS: An Iterative Knowledge Base System for improving description, classification and identification of biological objects. In: *Proceedings of the Indo-French Workshop on Symbolic Data Analysis and its Applications*, París (IFWSDAA'97).

[13] Conruyt, N.; Grosser, D. 1999 a. Managing complex knowledge in natural sciences. In: *Proceedings of the International Conference on Case-Based Reasoning*, Berlin (ICCBR'99).

[14] Conruyt, N.; Grosser, D. 1999 b. Tree Based classification approach for dealing with complex knowledge in natural sciences. *Proceedings of Machine Learning and Applications*, Chania - Greece, (ACAI'99).

[15] Dallwitz, M. J. 1974. A flexible computer program for generating identification keys. *Systematic Zoology* 1, 50-57.

[16] Dallwitz, M. J. 1992. A comparison of matrix-based Taxonomic identification systems with rule-based systems. In: *Proceedings of IFAC Workshop on Expert Systems in Agriculture*, pp. 215-218.

[17] Dallwitz, M. J. 2000 a. "*A Comparison of Interactive Identification Programs"*. [En línea]. [Consulta: 23/05/2003]. <http://biodiversity.uno.edu/delta/www/comparison.htm>.

[18] Dallwitz, M. J. 2000 b. "*Principles of interactive keys"*. [En línea]. [Consulta: 23/05/2003]. <http://biodiversity.uno.edu/delta/www/comparison.htm>.

[19] Dallwitz, M. J.; Paine, T.A.; Zurcher, E.J. 2000. "*User's guide to the DELTA system: A General System for Processing Taxonomic Descriptions, 4.12 edition"*. CSIRO Division of Entomology, Canberra. [En línea]. [Consulta: 25/05/2003]. < http://biodiversity.uno.edu/delta/>.

[20] Daoliang, Li; Zetian, Fu; Yanqing, Duan. 2002. Fihs-Expert: a web-based expert system for fish disease diagnosis. *Expert Systems with Applications* 23, 311-320.

[21] Dempster, A. 1967. Upper and Lower Probabilities Induced by a Multi-valued Mapping. *Annals of Mathematical Statistics,* vol 38, no. 2, pp. 325-399.

[22] Diederich, J.; Fortuner R. 1996. Endorsement of observations in identification. In: *Fifth IEEE International Conference on Fuzzy Systems*, 8-11 September 1996, New Orleans, LO, USA, pp. 175-179.

[23] Diederich, J.; Fortuner, R.; Milton, J. 2000. Genisys and computer-assisted identification of nematodes. *Nematology*, 2(1) 17-30.

[24] Dodd, J. C.; Rosendahl, S. 1996. The BEG Expert System – a multimedia identification system for arbuscular mycorrhizal fungi. *Mycorrhiza*, 6:275-278.

[25] Domingos, P. 1998. Occams's Two Razors: The Sharp and The Blunt. Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining (KDD-98), August 27-31, New York City, USA pp. 34-37.

[26] Domingos, P. 1999. the Role of Occams's Raxor in Knowledge Disvorvery. *Data Mining and Knowledge Discovery* Volume 3, pp 409-425.

[27] Donoghue, M.J.; Eriksson, T.; Piel, W.; Rice, K.; Sanderson M. 1996. *TreeBase. A database of phylogenetic knowledge.* [En línea]. [Consulta: 23/06/2004]. <http://herbaria.harvard.edu/treebase/ >.

[28] Duda, R; Hart, P. E.; Nilsson, N. J.; Reboh, R.; Slocum, J. & Sutheland, G. 1978. Development of the PROSPECTOR Consultation System for Mineral Exploration. *SRI Report*. Stanford Research Institute, Menlo Park, CA.

[29] Duncan, T.; Meacham, C. A. 1986 Multiple-entry-keys for the identification of angiosperm families using a microcomputer. *Taxon* 35, 492-494.

[30] Durkin, J. 1994. *Expert Systems. Design and Development.* Prentice Hall International Editions. Macmillan Publishing Company, New Jersey.

[31] Farris, J. S. 1988. "*Hennig86. Version 1.5"*. Port Jefferson Station: New York

[32] Fermanian, T.; Michalski, R. S. 1989. Weeder: An Advisory System for the Identification of Grasses in Turf. *Agronomy Journal* ,81(2), pp 313-316.

[33] Fermanian, T.; Michalski, R. S. 1992. Agriassistant: A new generation tool for agricultural advisory systems. *Expert Systems in Developing Countries, practice and promise*. Chapter 5. Westview Press.

[34] Forgy, C. L. 1982. Rete: A fast Algorithm for the Many Pattern/Many Object Pattern Problem. *Artificial Intelligence* 19, 17-37.

[35] Forrest, A.; Walsh, L. 2000. "*Diagnosis of Oral Ulceration"*. The University of Queensland. [En línea]. [Consulta: 25/05/2003]. <http://www.lucidcentral.com/keys/cpitt/public/ulcers/default.htm>.

[36] Fortuner, R. 1989. *Nematode identification and expert system technology.* New York USA, Plenum Publishing Corp.

[37] Friedman, H. 1998. Java Expert System *Shell* (Jess), *Technical Report. #SAND98-8206*, Sandia National Labs, Livermore, CA.

[38] Golding, S. 2000. "*Key to Mineral*s". Knowledge Books and Software. [En línea]. [Consulta: 25/05/2003]. <http://www.kbs.com.au/prod_show.asp?pID=20>.

[39] Gonzalez, A.J.; Dankel, D.D. 1993. *The Engineering of Knowledge-Based-Systems. Theory and Practice.* Prentice Hall, Englewood Cliffs, New Jersey.

[40] Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27, 857-871.

[41] Grove, R.; Hulse, A. 1999. An Internet – Based Expert System for Reptile Identification. In: *The First International Conference on The Practical Application of Java*, London-UK, pp165-173. [En línea]. [Consulta: 25/05/2003]. <http://grove.cs.jmu.edu/parih>.

[42] Grove, R. 2000. Internet-Based Expert Systems. *Expert Systems*, 17/3:129-135.

[43] Grzymala-Busse, J.W. 1991. *Managing Uncertainty in Expert Systems*. Ed. Kluwer Academic Publishers.

[44] Guala, G. F. 1999. "*Grasses of Florida: Interactive Key"*. [En línea]. [Consulta: 03/03/2003]. < http://www.virtualherbarium.org/grass/navikey/navikey.html>.

[45] Hagedorn, G. 1995-2003. *"Delta AccessVersion 1.81 for Access 97"*. [En línea]. [Consulta: 25/05/2003]. <http://www.diversitycampus.net/Workbench/download.html>.

[46] Harrison, P. R.; Kovalchik, J.G. 1998. Expert Systems and uncertainty. In: Liebowitz, J. (Ed.), Handbook of Applied Expert Systems. CRC Press, pp 1-11.

[47] Huson, D.H.; Wetzel, R. 1994. SplitsTree, version 1.01. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics* 14(1):68--73, 1998.

[48] Ignizio, J. P. 1991. *Introduction to Expert Systems. The Development and Implementation of Rule-Based Expert Systems.* Mc Graw-Hill, Inc.

[49] Intelsys Inc. 2000-2001. "*XID Authoring System 3.0 (Demo Version)"*. [En línea]. [Consulta: 3/06/2003]. <http://www.xidservices.com>.

[50] IPNI. 1999. "*International Plant Names Index"*. [En línea]. [Consulta: 23/05/2003]. <http://www.ipni.org>.

[51] Josie Lynn Catindig. 2001. "*Insects found in rice"*. International Rice Research Institute. [En línea]. [Consulta: 25/05/2003]. <http://www.irri.org/publications/catalog/cdrom.asp>.

[52] Lamberti, F.; Ciancio, A. 1994. The relationships between species within the *Xiphinema americanum* group. *Bulletin IEPP* 24, 475-484.

[53] Li Da Xu; Ning Liang; Qiong Gao. 2001. An integrated knowledge-based system for grasslands ecosystems. *Knowledge Based Systems* 14, 271-280.

[54] López González, G.; Do Amaral Franco, J.; 1986. *Gymnospermae*. En: Castroviejo, S. *et al.*, *Flora Ibérica* Vol I, 161-195. Madrid, Real Jardín Botánico, CSIC.

[55] López González, G. 2001. *Los árboles y arbustos de la Península Ibérica e islas Baleares*, Tomo I, 861pp. Madrid, Ediciones Multiprensa.

[56] Maddison, W. P.; Maddison, D. R. 1992. "*MacClade: Analysis of Phylogeny and Character Evolution. Version 3.3*". 98pp. (Sinauer Associates: Sunderland, Massachusetts).

[57] Maddison, D. R.; Swofford, D. L.; Maddison, W.P. 1997. NEXUS: An extensible File Format for Systematic Information. *Systematic Biology*, 46.

[58] Maddison, W. P.; Maddison, D.R. 2003. "*Mesquite: a modular system for evolutionary analysis. Version 0.995*". [En línea]. [Consulta 7/06/2003]. <http://mesquiteproject.org>.

[59] Mahaman, B. D.; Harizanis, P.; Filis, I.; Antonopoulou, E.; Yialouris, C. P.; Sideridis, A. B. 2002. A diagnostic expert system for honeybee pests. *Computers and electronics in agriculture*, 36: 17-31.

[60] Mahaman, B.; Passam, H. C.; Sideridis, A. B. *et al.,* 2003. DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in Solanaceous crop systems. *Agricultural Systems*, 76 (3): 1119-1135.

[61] Mamdani, E.H.; Efstathiou, H. 1985. Higer-order Logics for Handling Uncertainty in Expert Systems. *International Journal of Man-Machine Studies,* 22(3): 283-293.

[62] Marcot, B. 1987. Testing your Knowledge Base. *AI Expert*, vol 2 no. 8, pp 42-47.

[63] Matsatsinis, N. F.; Doumpos, M.; Zopounidis, C. 1997. Knowledge acquisition and representation for expert systems in the field of financial analysis. *Expert Systems with Applications*, 12(2), 247-262.

[64] Meacham, C. A. 1986-1996. *"Meka version 3.0"*. [En línea]. [Consulta: 3/06/2003]. <http://ucjeps.berkeley.edu/meacham/meka/>.

[65] Morales, C.; Quesada, C.; Baena, L. 2001. *Guías de la Naturaleza. Árboles y Arbustos*. Diputación de Granada. Granada.

[66] Morris, R. A; Stevenson, R. D. 2003. *Electronic Field Guide: An Object-Oriented WWW Database to Identify Species and Record Ecological Observations*. University of Massachussets, Boston. [En línea]. [Consulta: 19/09/2003]. <http://www.cs.umb.edu/efg>.

[67] Morse, L. E. 1970. Computer aids to plant identification. *American Journal of Botany*, 57 (6)754-&.

[68] Morse, L. E. 1971. Specimen identification and key construction with time-sharing computers. *Taxon,* 20:269–282.

[69] Nguyen, T. A; Perkins, W.A; Laffey, T.J.; Pecora, D. 1987. Knowledge base verification.*AI Magzine* 8, 69-75.

[70] O'Keefe, Robert M., Osman Balci, and Eric P. Smith. 1987. Validating Expert System Performance. *IEEE Expert* (winter): 81-89.

[71] Page R.D.M. 1993. *COMPONENT, version 2.0. "*. [En línea]. [Consulta:23/06/2004].< http://taxonomy.zoology.gla.ac.uk/rod/cpw.html.

[72] Payne, R. W. 1975. Genkey: a program for constructing diagnostic keys. *Biological Identification with computers*. Pp 65-72. Academic Press: London.

[73] Pankhurst, R. J. 1970. A computer program for generating diagnostic keys. *Computer Journal,* 13 (2), 145-151.

[74] Pankhurst, R. J. 1991. *Practical taxonomic Computing.* Cambridge University Press.

[75] Pankhurst, R. J.; Pullan M. 1994. *"Pandora User Guide version 3.1"*. [En línea].
[Consulta:8/06/2003].<http://www.ibiblio.org/pub/academic/biology/ecology+evolution/software/pandora>.

[76] Peterson, A. T.; Ortega-Huerta, M. A.; Bartley, J.; Sanchez-Cordero, V.; Soberon, J.; Buddemeier, R. H.; Stockwell, D. R. B. 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature* 416 (6881): 626-629.

[77] Pullan, M. R.; Watson, M. F.; Kennedy, J. B.; Raguenaud, C.; Hyam, R. 2000.The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon*, 49 (1): 55-75.

[78] Quinlan, J.R. 1986. Induction on Decision Trees. *Machine Learning*, 1, 1986, pp. 81-106.

[79] Quinlan, J.R. 1993. C4.5: *Programs for Machine Learning*. Morgan Kaufmann, ISBN 1-55860-238-0.

[80] Reinke, R. E. 1984. Knowledge acquisition and refinement tools for the ADVISE Meta Expert System. *M. S. Thesis ISG 844*, UIUCDCSF84921, Urbana, IL: Department of Computer Science, University of Illinois.

[81] Robertson, A.; Noren, J. G. 2001. Knowledge-based system for structured examination, diagnosis and therapy in treatment of traumatised teeth. *Dental Traumatology*, 17(1):5-9.

[82] Shaffer, G. 1976. *Mathematical Theory of Evidence.* Princeton, NJ: Princeton University Press.

[83] Schalk, P. H.; Heijman, R. P. 1996. *ETI's Linnaeus II taxonomic software: A new tool for interactive education*. UniServe.Science News, University of Sydney, Vol.3. [En línea]. [Consulta 7/5/2003]. <http://www.eti.uva.nl/Home/Articles.html>.

[84] Schalk, P. H.; Troost, D. G. 1999. Computer tools for accessing biodiversity information *Nature and Resources*, 35 (3): 31-38.

[85] Shortlife, E.; Buchanan, B. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23, 351-379.

[86] Shortlife, E. H. 1976. *Computer-Based Medical Consultations: Mycin*. new York, NY: American Elsevier.

[87] Silsby, J.; Trueman, J. 2002. "*Dragonflies of the World: Interactive Identification to Subfamilies"*. CSIRO Publishing. [En línea]. [Consulta: 25/05/2003]. <http://www.publish.csiro.au/books/bookpage.cfm?PID=2917 >.

[88] Stockwell, D. R. B. 1997. *Overview of Computational Biodiversity Research.* [En línea].[Consulta: 3/10/2002]. <http://biodi.sdsc.edu/Doc/BIS/overview.html>.

[89] Spiegel M. R. 2000. *Estadística*. McGraw-Hill/Interamericana de España, S.A.U.

[90] Suwa, M.; Scott, A. C.; Shortlife E. F. 1985. Completeness and consistency in a rule-based system. Rule Based Expert Systems. In *The MYCIN Experiments of the Stanford Heuristic Programming Project*, B. G. Buchanan, E.F. Shortlife (eds.), Addison-Wesley, 159-144.

[91] Swofford, D. L. 1991. *PAUP: phylogenetic analysis using parsimony.* Version 3.1. Illinois Natural history Survey: Champaign.

[92] TDWG working group:Structure of Descriptive Data (SDD). 2003. *SDD Part 0: Introduction and Primer to the SDD Standard*. [En línea]. [Consulta 23/3/2004]. <http://160.45.63.11/Projects/TDWG-SDD/Primer/index.htm>.

[93] UBio. 2003. X:ID, Version. [En línea]. [Consulta 30/7/2003]. <http://ubio.org/offerings/applications/key/index.html>.

[94] University of Queensland. 1999. Insects to Order. *Knowledge Books and Software*. [En línea]. [Consulta: 25/05/2003]. <http://www.lucidcentral.com/keys/cpitt/public/Insects/html/intro.htm>.

[95] University of Toronto. Department of Botany; the University of Toronto Libraries; the Royal Ontario Museum. 1996 a. *Pollyclave a multi-entry identification key version 1.6*. [En línea]. [Consulta 6/06/2003]. <http://prod.library.utoronto.ca/pollyclave/index.html>.

[96] University of Toronto. Department of Botany; the University of Toronto Libraries; the Royal Ontario Museum. 1996 b. *Phalaenopsis species*. [En línea]. [Consulta 6/06/2003]. <http://prod.library.utoronto.ca/polyclave/orchids/phalhome.htm >.

[97] Vranes, S.; Stanojevic, M.; Stevanovic, V.; Lucin, M. 1996. INVEX: Investment advisory expert system. *Expert Sytems*, 13(2), 105-119.

[98] Wang, X.; Chen, B.; Qian, G.; Ye, F. 2000. On the optimization of fuzzy decision tress. *Fuzzy Sets and Systems*, 112, Elsevier Science B.V., pp. 117-125.

[99] Watson, L; Dallwitz, M.J.; Johnston, C.R. 1986. Grass genera of the world: 728 detailed descriptions from an automated database. *Australian Journal of Botany*. 34, 223-30.

[100] Zadeh, L. A. 1965. *Fuzzy Sets*. Information and Control, vol. 8, no. 3, pp. 338-353..