CAPÍTULO 27

Nuevas tecnologías y aplicaciones de bases de datos

A lo largo de este libro, hemos abordado una serie de cuestiones relacionadas con el modelado, diseño y funciones de las bases de datos así como con las cuestiones relativas a la estructura interna y rendimiento de los sistemas de gestión de bases de datos. En el capítulo anterior, consideramos algunas tendencias de la tecnología de gestión de bases de datos como, por ejemplo, los almacenes de datos, que proporcionan bases de datos de gran tamaño para dar soporte a la toma de decisión.

En este capítulo, nos centraremos en dos categorías de avances más recientes en el campo de las bases de datos: (1) las nuevas tecnologías de bases de datos, y (2) los dominios de aplicación principales. La primera se ocupa de la creación de nuevos entornos así como de la funcionalidad de los SGBD, con el fin de que se puedan manejar una serie de nuevas aplicaciones entre las que se incluyen el acceso universal a las bases de datos en la *World Wide Web* y las bases de datos en Internet; las bases de datos multimedia que proporcionan una funcionalidad adicional para sostener el almacenamiento y procesamiento de información multimedia; así como bases de datos móviles y bases de datos conectadas intermitentemente que permiten al usuario final extraer y trasladar partes de una base de datos mientras se encuentra móvil en un campo. De la Sección 27.1 a la 27.3 haremos una breve introducción y examinaremos las cuestiones y enfoques para solucionar los problemas específicos que surgen en estas tres nuevas tecnologías.

Posteriormente, examinaremos tres dominios de aplicación que desde siempre se han basado en el procesamiento manual de los sistemas de ficheros o en soluciones de sistemas hechos a medida. En la Sección 27.4, se analizan los sistemas de información geográfica, que se ocupan sólo de datos geográficos o de datos espaciales junto con datos no espaciales, como es el cómputo de censos. En la Sección 27.5 se examinan las bases de datos biológicas, en concreto, las que contienen datos genéticos en diferentes organismos, incluyendo los datos sobre el genoma humano. La Sección 27.6 aborda las bibliotecas digitales, grandes almacenes de datos digitales en múltiples medios. Esta es una aplicación que conlleva el reto de agrupar los conjuntos de documentos más desestructurados, variados y sin relación alguna entre sí, junto con otros elementos de biblioteca y hacer que

estén disponibles para una búsqueda y recuperación eficientes dentro de un sistema. Una característica común de todas estas aplicaciones es la naturaleza específica del dominio de los datos en cada dominio de aplicación concreto. Además, todas ellas se caracterizan por su naturaleza «estática», una situación en la que el usuario final sólo puede recuperar datos de la base de datos; la actualización con información nueva está limitada a los expertos en dominios de bases de datos que son los que supervisan y analizan los nuevos datos que se introducen.

27.1. Bases de datos en la World Wide Web

La *World Wide Web* (WWW), (conocida popularmente como «la Web»), se creó originariamente en Suiza en CERN[1] a principios de los años noventa como un sistema de servicios de información hipermedia a gran escala destinado a que los científicos biológicos compartieran información.[2] En la actualidad, esta tecnología permite que todo aquél que tenga acceso a Internet tenga acceso universal a esta información compartida y la Web contiene cientos de millones de páginas Web que se hallan al alcance de millones de usuarios.

En la tecnología Web, una arquitectura básica de tipo cliente-servidor subyace a todas las actividades. La información se almacena en computadores designados como servidores Web en ficheros compartidos accesibles al público. La información está codificada empleando el ***HyperText Markup Language*** (**HTML** o Lenguaje de marcas de hipertexto). Una serie de herramientas permite que los usuarios creen páginas Web formateadas con etiquetas HTML, que se combinan libremente con contenido multimedia, que van desde gráficos a audio e incluso vídeo. Una página contiene numerosos **hipervínculos** intercalados, literalmente un vínculo que permite que un usuario «navegue» o se mueva de una página a otra a través de Internet. Esta capacidad ofrece unas posibilidades tremendas a los usuarios finales a la hora de buscar y navegar por información relacionada entre sí (con frecuencia a través de distintos continentes).

La información contenida en la Web se organiza conforme a un ***Uniform Resource Locator*** (**URL** o Localizador Uniforme de Recursos), algo similar a una dirección que proporciona el nombre del camino completo de un fichero. El nombre del camino consta de una secuencia de nombres de máquina y directorios separados por barras oblicuas («/») y que termina con el nombre de un fichero. Por ejemplo, el índice de este libro tiene la siguiente dirección URL:

http://www.awl.com/cseng/authors/elmasri/Dbase3e/Dbase3e.html

Un URL siempre comienza con un ***hypertext transport protocol*** (**http** o protocolo de transporte de hipertexto), que es el protocolo empleado por los **navegadores Web**, un programa que se comunica con el servidor Web, y viceversa. Los navegadores Web interpretan y presentan a los usuarios los documentos HTML. Entre los navegadores famosos se encuentran el Internet Explorer de Microsoft y el Netscape Navigator. Un conjunto de documentos HTML y otros ficheros accesibles a través del URL de un servidor Web recibe el nombre de ***Web site*** (sitio Web). En la dirección URL anterior, «www.awl.com» es el nombre del sitio Web de Addison Wesley Publishing.

---

[1] CERN son las siglas del término francés «Conseil European pour la Recherche Nucleaire» o Consejo Europeo para la Investigación Nuclear.

[2] Esta idea de tan asombroso éxito se atribuye a Berners-Lee y su equipo; *véase* Berners-Lee *et al.* (1992, 1994).

27.1.1. Acceso a las bases de datos en la *World Wide Web*

La tecnología actual ha ido transformándose rápidamente de páginas Web estáticas a dinámicas, en las que el contenido puede estar en un estado constante de cambio. El servidor Web emplea una interfaz estándar denominado ***Common Gateway Interface*** (**CGI** o interfaz de pasarela común) para que actúe como **capa intermedia** (*middleware*), es decir, la capa de software adicional, que se encuentra entre la interfaz que ve el usuario y el SGBD subyacente, que facilita el acceso a bases de datos heterogéneas. La capa intermedia CGI ejecuta programas externos o *scripts* para obtener la información dinámica, y devuelve la información al servidor en HTML, la cual es enviada de nuevo al navegador.

A medida que la Web ha ido experimentando sus transformaciones más recientes, ha sido necesario permitir que los usuarios accedan no sólo a los sistemas de ficheros sino a bases de datos y a los SGBD para mantener el procesamiento de consultas y la creación de informes, entre otros. Los enfoques actuales pueden dividirse en dos categorías:

1. *Acceso mediante scripts CGI*: puede obligarse al servidor de la base de datos a que se relacione con el servidor Web por medio del CGI. La Figura 27.1 muestra un esquema de la arquitectura de acceso a la base de datos en la Web mediante *scripts* CGI, que están escritos en lenguajes como PERL, Tcl, o C. La principal desventaja de este enfoque es que para cada petición de un usuario, el servidor Web debe iniciar un nuevo proceso CGI: cada proceso realiza una nueva conexión con el SGBD y el servidor Web debe esperar hasta que se le envíen los resultados. No se consigue un mayor rendimiento si se agrupan las peticiones de usuarios múltiples; además, el diseñador debe mantener los *scripts* únicamente en los subdirectorios CGI-bin, lo que le hace vulnerable a una posible violación de su seguridad. Asimismo, el hecho de que el CGI no tenga un lenguaje propio sino que requiera que los diseñadores de bases de datos aprendan los lenguajes PERL o Tcl, constituye una desventaja. La manejabilidad de los *scripts* es otro problema si los *scripts* se dispersan por todas partes.

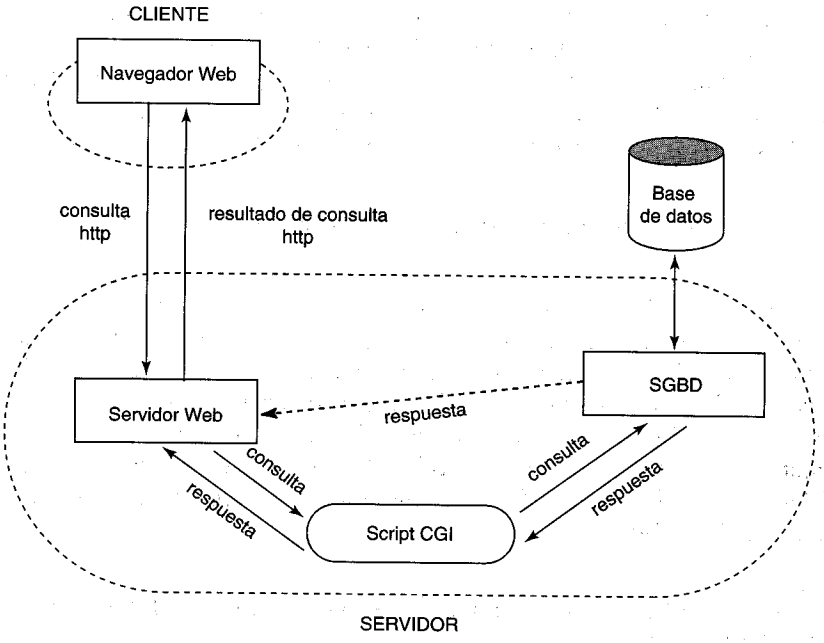

Figura 27.1. Aceso a una base de datos en la Web mediante *Scripts* CGI.

2. *Acceso mediante JDBC*: JDBC es un conjunto de clases Java creadas por Sun Microsystems para permitir el acceso a bases de datos relacionales mediante la ejecución de sentencias SQL. Es una forma de conectarse con bases de datos sin ningún proceso adicional para la petición de cada cliente. Obsérvese que JDBC es un nombre registrado por Sun; *no* corresponde a las siglas de *Java Data Base connectivity* como muchos creen. JDBC tiene la capacidad de conectarse a una base de datos, enviar sentencias SQL a una base de datos y recuperar los resultados de una consulta empleando las clases de Java Connection, Statement y ResultSet respectivamente. Con la anunciada independencia de la plataforma Java, una aplicación puede ejecutarse en cualquier navegador compatible con Java, que descarga el código Java del servidor y lo ejecuta en el navegador del cliente. El código Java es transparente para el SGBD; los *drivers* de JDBC para SGBD individuales en la parte del servidor tienen la labor de interactuar con ese SGBD. Si el *driver* de JDBC se encuentra en la parte del cliente, la aplicación se ejecuta en el cliente y el *driver* comunica sus peticiones al SGBD directamente. Para las peticiones en SQL estándar, se puede acceder a muchos SGBD relacionales de esta manera. El inconveniente de emplear JDBC es la necesidad de ejecutar Java a través de máquinas virtuales con su impacto inherente sobre la eficiencia. El puente JDBC a la *Object Database Connectivity* (ODBC) sigue siendo otra forma de llegar a los SGBD relacionales.

Además del CGI, otros vendedores de servidores Web están lanzando sus propios productos de capa intermedia (*middleware*) destinados a proporcionar una conectividad de base de datos múltiple. Entre estos se incluyen Internet Server API (ISAPI) de Microsoft y Netscape API (NSAPI) de Netscape. En la próxima sección, describiremos la opción de acceso a la Web que ofrece Informix. Otros vendedores de SGBD ya cuentan, o contarán con prestaciones similares para facilitar el acceso a bases de datos en la Web.

27.1.2. La opción de integración Web de INFORMIX

Informix ha hecho frente a las limitaciones del CGI y a las incompatibilidades de CGI, NSAPI, e ISAPI mediante la creación de la *Web Integration Option* (WIO u Opción de Integración Web). WIO elimina la necesidad de utilizar scripts. Los desarrolladores emplean herramientas para crear páginas HTML inteligentes denominadas *Application Pages* (*App Pages* o Páginas de Aplicaciones) directamente dentro de la base de datos. Éstas ejecutan sentencias SQL dinámicamente, formatean los resultados en HTML y devuelven la página Web obtenida a los usuarios finales. La arquitectura esquemática se muestra en la Figura 27.2. La WIO emplea el ***Web Driver***, un proceso CGI sencillo al que se acude cuando se recibe una petición de URL en el servidor Web. Se genera un único identificador de sesión para cada petición pero la aplicación WIO es continua y *no* finaliza después de cada petición.

Cuando la aplicación WIO recibe una petición del Web driver, ésta se conecta a la base de datos y ejecuta Web Explode, una función que ejecuta consultas dentro de las páginas Web y formatea los resultados en forma de página Web que regresa al navegador por medio del Web driver.

Las extensiones de etiquetas HTML de Informix permiten que los autores de páginas Web creen aplicaciones que puedan construir plantillas de páginas Web de forma dinámica a partir del Servidor Dinámico de Informix y las presenten a los usuarios finales. La WIO también permite a los usuarios crear sus propias etiquetas personalizadas para realizar tareas especializadas. De este modo, se pueden diseñar aplicaciones potentes sin tener que recurrir a la creación de programas o scripts. Hay otra característica de WIO que ayuda a las aplicaciones orientadas a transacciones proporcionando una interfaz de programación de aplicaciones (API) que ofrece un conjunto de servi-

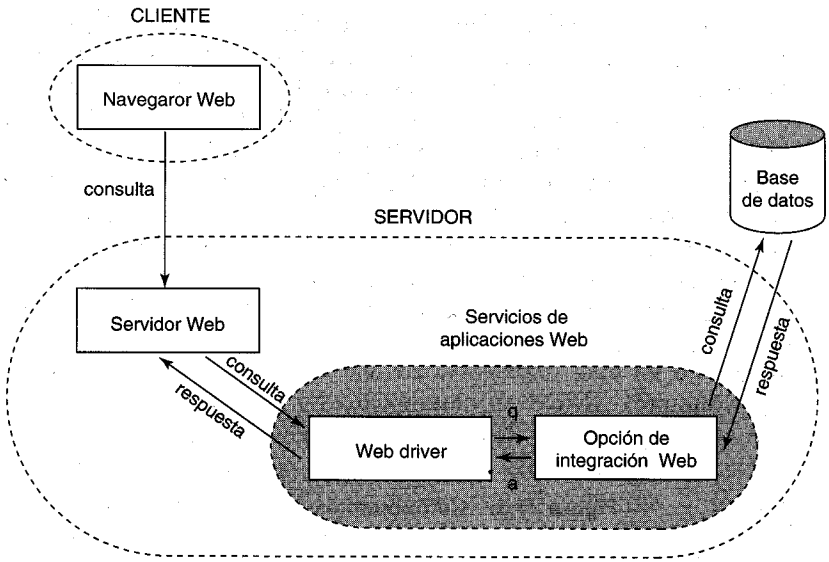

Figura 27.2. Implementación CGI de la opción de integración Web en Informix.

cios básicos como son la gestión de la conexión y la sesión, que pueden incorporarse a la aplicación Web.

La WIO sirve de soporte a aplicaciones creadas en C, C++ y Java. Esta flexibilidad permite a los desarroladores transportar aplicaciones existentes a la Web o crear nuevas aplicaciones en estos lenguajes. La WIO se integra en el software del servidor Web y emplea el mecanismo de seguridad original del Servidor Dinámico de Informix. La arquitectura abierta de la WIO permite el empleo de diversos navegadores y servidores Web.

27.1.3. El servidor Web de ORACLE

ORACLE soporta el acceso desde la Web a bases de datos empleando los componentes que se muestran en la Figura 27.3. El cliente solicita al servidor Web ficheros que se denominan «estáticos» o «dinámicos». Los ficheros estáticos tienen un contenido fijo mientras que los ficheros dinámicos pueden tener un contenido que incluye los resultados de las consultas realizadas a la base de datos. Hay un demonio HTTP (un proceso que se ejecuta continuamente) llamado *Web Listener* que opera en el servidor y atiende las peticiones efectuadas por los clientes. Se obtiene un fichero estático (documento) del sistema de ficheros del servidor y se visualiza en el navegador Web del cliente. El *Web listener* pasa una petición de página dinámica a un *Web request broker* (WRB), que se encarga de despacharla a uno de sus múltiples subprocesos (*threads*), el asociado con el cartucho correspondiente. Los **cartuchos** son módulos de software (que ya fueron mencionados anteriormente en la Sección 13.2.6) que llevan a cabo funciones específicas sobre tipos de datos específicos; estos pueden comunicarse entre sí. Actualmente, existen cartuchos para PL/SQL, Java y Live HTML; también pueden obtenerse cartuchos personalizados.

El servidor Web se halla plenamente integrado con PL/SQL, lo que lo hace eficiente y escalable. Los cartuchos le confieren una flexibilidad adicional, lo que posibilita que se pueda trabajar con otros lenguajes y paquetes de software. Se puede emplear una capa de *sockets* seguros más avanzada que proporciona una comunicación segura a través de Internet. La herramienta de desa-

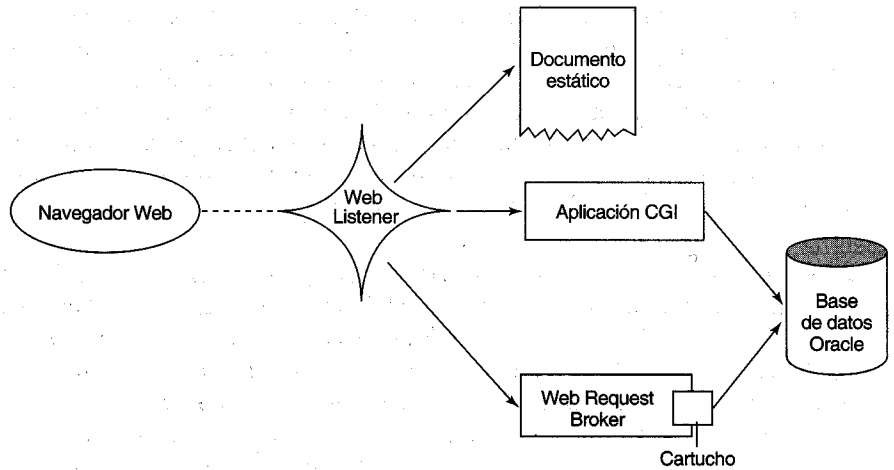

Figura 27.3. Componentes del servidor Web ORACLE.

rrollo Designer 2000 (*véase* la Sección 16.1) cuenta con un generador Web que posibilita que las aplicaciones anteriores creadas para las LAN se transfieran a entornos de Internet e Intranet.

27.1.4. Problemas por resolver con las bases de datos Web

La Web constituye un factor importante a la hora de planificar los entornos informáticos empresariales, tanto para proporcionar acceso desde el exterior a los sistemas de la empresa como para poner a disposición de clientes y proveedores información con fines comerciales y publicitarios. A su vez, debido a requisitos de seguridad, los empleados de algunas organizaciones están autorizados únicamente a operar dentro de **intranets** (subredes a las que no se puede acceder libremente desde el mundo exterior). Entre las aplicaciones destacadas de intranet y la WWW se encuentran las bases de datos que sirven de soporte a escaparates electrónicos, catálogos de productos y repuestos, directorios (guías telefónicas) y agendas, quioscos y librerías. Es probable que el **comercio electrónico** (la adquisición de productos y servicios electrónicamente a través de Internet) se convierta en una de las principales aplicaciones que sustenten estas bases de datos.

Serán muchos los futuros retos de la gestión de bases de datos en la Web, entre los que se encuentran los siguientes:

- La tecnología Web necesita integrarse con la tecnología de objetos. Actualmente, la web puede verse como un sistema de objetos distribuidos, con páginas HTML que funcionan como objetos identificados por el URL.
- La funcionalidad de HTML es demasiado simple para soportar requisitos de aplicaciones complejas. Como ya vimos, la Opción de Integración Web de Informix (WIO) añade más etiquetas al HTML. Por lo general, se necesitarán prestaciones adicionales para (1) hacer que los clientes Web funcionen como la parte visible de la aplicación, integrando datos procedentes de múltiples bases de datos heterogéneas; (2) hacer que los clientes Web presenten diferentes vistas de los mismos datos a diferentes usuarios; y (3) hacer a los clientes Web «inteligentes» proporcionándoles una funcionalidad de minería de datos adicional (*véase* la Sección 26.2).
- El contenido de la página Web puede hacerse más dinámico si se le añade un mayor «comportamiento» de objeto (*véase* el Capítulo 11 para una descripción del modelado de objetos).

A este respecto (1) se puede hacer interactuar a los objetos del cliente y del servidor (páginas HTML); (2) las páginas Web pueden tratarse como conjuntos de objetos programables; y (3) el código del cliente puede acceder a estos objetos y manipularlos dinámicamente.

- Servir de soporte para un gran número de clientes junto con unos tiempos de respuesta razonables para consultas muy extensas (varias decenas de gigabytes de tamaño) constituirán los principales retos de las bases de datos Web. Serán los servidores Web y los SGBD subyacentes los que tendrán que enfrentarse a estos retos.

Se están realizando esfuerzos considerables para hacer frente a estas limitaciones de la tecnología actual de estructuración de datos, especialmente por parte del *World Wide Web Consortium* (W3C). El W3C está diseñando un Modelo de Objetos Web. Asimismo, propone un ***Extensible Markup Language*** (**XML** o lenguaje de marcas extensible) para un intercambio estructurado de documentos en la Web. XML define un subconjunto de **SGML** (***Standard Generalized Markup Language*** o Lenguaje de Marcas Generalizado Estándar), el cual permite la adaptación de lenguajes de marcas con etiquetas específicas para una aplicación. XML está ganando terreno rápidamente debido a su extensibilidad a la hora de definir etiquetas nuevas. El ***Document Object Model*** (**DOM** o Modelo de Documentos Orientado a Objetos) del W3C define un API orientado a objetos para documentos HTML o XML presentados por un cliente Web. El W3C también está definiendo estándares para el modelado de metadatos destinados a describir los recursos de Internet.

La tecnología para modelar información empleando los estándares mencionados anteriormente y para encontrar información en la Web está experimentando una gran evolución. En términos generales, los servidores Web tienen que ganar solidez como tecnología que resulte fiable para manejar bases de datos en el ámbito de producción y que soporten aplicaciones 24 × 7 (24 horas al día, 7 días a la semana). La seguridad sigue siendo un problema crucial cuando se trata de servir de soporte a aplicaciones que conllevan bases de datos financieras y médicas. Además, la transferencia desde entornos de aplicaciones de bases de datos ya existentes a los existentes en la Web requerirá un soporte adecuado que posibilitará que los usuarios continúen con su modo de operación actual, así como una costosa infraestructura para llevar a cabo la migración de datos entre sistemas sin que se introduzcan incoherencias. La funcionalidad tradicional de las bases de datos de consulta y procesamiento de transacciones deberá sufrir las modificaciones precisas para servir de soporte a aplicaciones basadas en la Web. Una de estas áreas la constituyen las bases de datos móviles, que abordaremos en la Sección 27.3.

27.1.5. Bibliografía seleccionada para las bases de datos de la World Wide Web

La idea de la World Wide Web fue propuesta por Berners-Lee (1992, 1994) y su grupo CERN en Ginebra. La Opción de Integración Web (WIO) de Informix se describe en Informix (1998a). Manola (1998) examina la creación de estándares nuevos como XML y el modelo de documentos orientado a objetos para la integración de la tecnología Web con la tecnología de objetos. Mendelzon (1997) describe conceptos para el procesamiento de consultas en la Web; Gravano y García-Molina (1997) proponen la noción de clasificar las respuestas a consultas realizadas de forma libre en la Web; Atzeni *et al.* proponen un modelo de datos denominado ARANEUS así como dos lenguajes para la realización de consultas, la creación de vistas hipertextuales y la derivación de datos. Fraternali (1999) ofrece un estudio de los enfoques destinados a servir de soporte a las aplicaciones web con una gran cantidad de datos. El área del comercio electrónico obtendrá numerosas ventajas y compartirá muchas de las dificultades del desarrollo de bases de datos en la Web; Dogac propor-

ciona una buena perspectiva general de todo ello (1998). Existen varios informes oficiales en los sitios Web de los vendedores de bases de datos en los que se afirma que sus productos ofrecen acceso Web a bases de datos (por ejemplo, www.oracle.com, www.informix.com). El www consortium (W3C) cuenta con un sitio Web en el que se puede encontrar información actualizada sobre XML: www.w3.org/XML/; también se describen protocolos en www.w3.org/protocols/. Para obtener más detalles sobre la especificación API JDBC, se puede consultar la dirección www.java.sun.com/products/.

27.2. Bases de datos multimedia

En los próximos años, se espera que los sistemas de información multimedia dominen nuestra vida diaria. Nuestras casas contarán con una instalación de banda ancha para acceder a aplicaciones multimedia interactivas. Nuestros televisores de alta definición y terminales de computador tendrán acceso a un gran número de bases de datos, incluyendo bibliotecas digitales (*véase* la Sección 27.6), las cuales distribuirán cantidades ingentes de contenido multimedia procedente de fuentes múltiples.

27.2.1. La naturaleza de los datos multimedia y de las aplicaciones

En la Sección 23.3, abordamos las cuestiones del modelado avanzado relacionadas con los datos multimedia. Asimismo, en el Capítulo 13 examinamos el procesamiento de tipos múltiples de datos en el contexto de los SGBD objeto relacionales (SGBDOR). Los SGBD los han ido incorporando constantemente a los tipos de datos que soportan. Hoy en día, se pueden encontrar los siguientes tipos de datos multimedia en los sistemas actuales:

- *Texto*: puede estar formateado o sin formatear. Para facilitar el análisis sintáctico de los documentos, se están empleando estándares como SGML y variaciones como HTML.
- *Gráficos*: entre los ejemplos se incluyen dibujos e ilustraciones que se codifican mediante estándares descriptivos (por ejemplo, CGM, PICT, *postcript*).
- *Imágenes*: incluyen dibujos, fotografías, entre otros, codificadas en formatos estándar como *bitmap* (mapa de bits), JPEG, y MPEG. La compresión se realiza en JPEG y MPEG. Estas imágenes no se subdividen en componentes. Por lo tanto, no es sencillo efectuar una consulta por contenido (por ejemplo, buscar todas las imágenes que contengan círculos).
- *Animaciones*: secuencias temporales de imágenes o datos gráficos.
- *Vídeo*: un conjunto de datos fotográficos secuenciados que se presentan a una velocidad especificada, por ejemplo 30 cuadros por segundo.
- *Audio estructurado*: una secuencia de componentes de audio que constan de nota, tono, duración, etc.
- *Audio*: muestra de datos generada de a partir de grabaciones de sonido constituida por una cadena de bits en formato digital. Las grabaciones analógicas se convierten normalmente a formato digital antes de almacenarlas.
- *Datos multimedia compuestos o combinados*: una combinación de tipos de datos multimedia, como son audio y vídeo, que pueden combinarse físicamente produciendo un nuevo formato de almacenamiento o combinarse lógicamente al tiempo que mantienen los tipos y formatos originales. Asimismo, los datos compuestos contienen información de control adicional que describen la forma en la que se debería proporcionar la información.

Naturaleza de las aplicaciones multimedia. Los datos multimedia pueden almacenarse, enviarse y utilizarse de formas muy diferentes. Las aplicaciones pueden clasificarse de acuerdo con las siguientes características de gestión de datos:

- *Aplicaciones de almacenamiento*: se almacena una gran cantidad de datos multimedia así como de metadatos a fin de que éstos puedan ser recuperados. Un SGBD puede contar con un almacén central que contenga datos multimedia y puede organizarse en una jerarquía de niveles de almacenamiento (discos locales, discos y cintas terciarias, discos ópticos, etc). Entre los ejemplos se incluyen almacenes de imágenes de satélite, dibujos y diseños de ingeniería, fotografías del espacio e imágenes de radiología escaneadas.
- *Aplicaciones de presentación*: un gran número de aplicaciones conlleva el envío de datos multimedia que se hallan sujetos a restricciones temporales. Los datos de audio y vídeo se envían de esta forma. En estas aplicaciones, para tener condiciones óptimas de visión o de escucha se requiere que el SGBD envíe los datos a velocidades determinadas, que superen un determinado valor umbral, para ofrecer «calidad de servicio». Los datos se consumen a medida que éstos se envían, a diferencia de las aplicaciones de almacenamiento, donde se pueden procesar más tarde (por ejemplo, el correo electrónico multimedia). Por ejemplo, la simple proyección multimedia de datos de vídeo, requiere un sistema que simule la funcionalidad de un aparato de vídeo. Las presentaciones multimedia complejas e interactivas conllevan indicaciones de organización para controlar el orden de recuperación de los componentes en serie o en paralelo. Los entornos interactivos deben contar con posibilidades como son el análisis de edición en tiempo real o la anotación de datos de vídeo y audio.
- *Trabajo colaborativo empleando información multimedia*: se trata de una nueva categoría de aplicaciones en las que los ingenieros pueden ejecutar una compleja labor de diseño mediante la combinación de dibujos, adaptando los temas a las restricciones de diseño, y creando documentación nueva, modificar notificaciones, etc. Las redes de asistencia médica inteligentes así como la telemedicina requerirán la colaboración entre los facultativos, analizando datos e información multimedia de los pacientes en tiempo real a medida que se va generando.

Todas estas áreas de aplicación presentan enormes retos para el diseño de los sistemas de bases de datos multimedia.

27.2.2. Cuestiones de gestión de datos

Las aplicaciones multimedia que trabajan con miles de imágenes, documentos, segmentos de audio y vídeo, y datos de texto libre, precisan del modelado adecuado de la estructura y contenido de los datos y, posteriormente, del diseño de esquemas de bases de datos apropiados para el almacenamiento y recuperación de la información multimedia. Los sistemas de información multimedia resultan muy complejos y abarcan un amplio conjunto de cuestiones, entre las que se encuentran las siguientes:

- *Modelado*: esta área cuenta con el potencial para aplicar al problema técnicas de bases de datos frente a las de recuperación de información. Existen problemas a la hora de tratar objetos complejos (*véase* el Capítulo 11) integrados por una amplia variedad de tipos de datos: numéricos, de texto, gráficos (imágenes generadas por ordenador), imágenes gráficas animadas, series audio, y secuencias de vídeo. Los documentos constituyen un área especializada y merecen una consideración especial.
- *Diseño*: el diseño conceptual, lógico y físico de las bases de datos multimedia aún no ha sido tratado del todo y sigue siendo un área de investigación activa. El proceso de diseño puede

fundamentarse en la metodología general descrita en el Capítulo 16, pero las cuestiones de rendimiento y ajuste en cada nivel resultan mucho más complejas.

- *Almacenamiento*: el almacenamiento de datos multimedia en dispositivos estándar similares a discos presenta problemas de representación, compresión y transformación a las jerarquías de los dispositivos, archivado y empleo de búferes durante la operación de entrada/salida. Es probable que una forma en la que la mayoría de los vendedores de productos multimedia aborden esta cuestión sea adheriéndose a estándares como JPEG o MPEG. En los SGBD, un soporte «BLOB» (*Binary Large Object* u Objeto Grande Binario) permite que se almacenen y se recuperen mapas de bits carentes de tipo. Se requerirá software estandarizado para tratar la sincronización y compresión/descompresión, a lo que se unirán problemas de indexación, que continúan en investigación.
- *Consultas y recuperación*: la forma que tienen las «bases de datos» de recuperar información está basada en los lenguajes de consulta y en las estructuras de índices internas. La «recuperación de información» depende estrictamente de palabras claves o términos de índices predefinidos. En lo que se refiere a imágenes, datos de vídeo y audio, esto da lugar a numerosas cuestiones, entre las que se encuentran la formulación eficiente de consultas, la ejecución y optimización de consultas. Es necesario modificar las técnicas de optimización estándar que estudiamos en el Capítulo 18 para tratar los tipos de datos multimedia.
- *Rendimiento*: en cuanto a las aplicaciones multimedia que contienen únicamente documentos y texto, las restricciones de rendimiento vienen determinadas subjetivamente por el usuario. Para aquellas aplicaciones que incluyen reproducción de video o sincronización de audio-vídeo, se imponen limitaciones físicas. Por ejemplo, el vídeo debe reproducirse a una velocidad constante de 60 cuadros por segundo. Las técnicas para la optimización de consultas pueden calcular el tiempo de respuesta estimado antes de evaluar la consulta. El empleo de procesamiento de datos paralelo puede paliar algunos de estos problemas, pero actualmente estos esfuerzos se hallan sujetos a una mayor experimentación.

Estas cuestiones han dado lugar a una serie de problemas de investigación aún por resolver. A continuación, examinaremos algunos de los problemas más representativos.

27.2.3. Problemas de investigación por resolver

Bases de datos frente a perspectivas de recuperación de información. El modelado del contenido de los datos no ha constituido aún una cuestión a tratar en los modelos y sistemas de bases de datos debido a que los datos tienen una estructura rígida y el significado de una ocurrencia de datos puede inferirse del esquema. Por el contrario, la recuperación de información (RI) se ocupa principalmente del modelado del contenido de los documentos de texto (mediante el empleo de palabras clave, índices de frases, redes semánticas, frecuencias de palabras, codificación *soundex*, etc) para los que generalmente no se tiene en cuenta la estructura. Mediante el modelado del contenido, el sistema puede establecer si un documento es relevante para una consulta examinando los descriptores de contenido del documento. Consideremos, por ejemplo, un parte de accidente de una compañía de seguros como un objeto multimedia: éste incluye imágenes del accidente, formularios de seguros estructurados, grabaciones audio de las partes implicadas en el accidente, el informe escrito del representante de la compañía de seguros, y demás información. ¿Qué modelo de datos se debería emplear para representar información multimedia de este tipo? ¿Dé qué modo se deberían formular las consultas sobre estos datos? Por lo tanto, la ejecución eficiente constituye una cuestión compleja, y la heterogeneidad semántica y la complejidad de representación de la información multimedia dan lugar a un gran número de problemas nuevos.

Requisitos del modelado y recuperación de datos multimedia/hipermedia. Para conseguir todo el poder expresivo del modelado de datos multimedia, el sistema debería contar con un constructor general que permita al usuario especificar los enlaces entres dos nodos arbitrarios. Los enlaces hipermedia o hiperenlaces tienen una serie de características diferentes:

- Los enlaces pueden especificarse con o sin información asociada, y pueden tener descripciones de gran tamaño asociadas a ellos.
- Los enlaces pueden tener su origen a partir de un punto específico dentro de un nodo o desde todo el nodo.
- Los enlaces pueden ser direccionales, o no direccionales cuando pueden atravesarse en cualquier sentido.

La capacidad de enlace del modelo de datos debería tener en cuenta todas estas variaciones. Cuando se requiere una recuperación de datos multimedia basada en el contenido, el mecanismo de consulta debería tener acceso a los enlaces y a la información asociada al enlace. El sistema debería ofrecer prestaciones para la definición de vistas de todos los enlaces, privados y públicos. Se puede obtener una información contextual valiosa de la información estructural. Los enlaces hipermedia generados automáticamente no revelan nada nuevo sobre los dos nodos, y tendrían una trascendencia diferente frente a los enlaces hipermedia generados manualmente. Los medios para la creación y empleo de este tipo de enlaces, así como para el desarrollo y uso de lenguajes de consulta de navegación para utilizar dichos enlaces, constituyen características importantes de cualquier sistema que permiten el empleo eficaz de información multimedia. Esta área es importante para las bases de datos interrelacionadas en la WWW. (*Véase* la Sección 27.1 para un análisis de las bases de datos en la WWW.)

Indexación de imágenes. Existen dos métodos para la indexación de imágenes: (1) identificar los objetos automáticamente mediante técnicas de procesamiento de imágenes, y (2) asignar términos y frases de índice mediante la indexación manual. Un problema importante que presenta el empleo de técnicas de procesamiento de imágenes en la indexación de imágenes es el de la escalabilidad. La técnica actual permite únicamente la indexación en imágenes de patrones simples. La complejidad aumenta con el número de características reconocibles. Otro problema importante tiene que ver con la complejidad de la consulta. Pueden emplearse reglas y mecanismos de inferencia, como se vio en el Capítulo 25, para derivar hechos de nivel superior a partir de características simples de imágenes. Esto permite la realización de consultas de mayor nivel como «encontrar edificios de hoteles que tengan vestíbulos abiertos y permitan una intensidad solar máxima en el área de recepción» en una aplicación arquitectónica.

El método de recuperación de información aplicado a la indexación de imágenes está basado en uno de estos tres esquemas de indexación:

1. *Sistemas de clasificación*: clasifica las imágenes jerárquicamente en categorías predeterminadas. En este método, tanto el diseñador de índices como el usuario deberán tener un buen conocimiento de las categorías disponibles. No pueden capturarse detalles más precisos de una imagen compleja ni de las relaciones entre los objetos de una imagen.
2. *Sistemas basados en palabras clave*: emplean un vocabulario de indexación similar al empleado en la indexación de documentos textuales. Se pueden obtener hechos simples representados en la imagen (como «región cubierta de hielo») y hechos que provengan de una interpretación de alto nivel por parte humana (como hielo permanente, nevada reciente y casquete polar).
3. *Sistemas de relaciones entidad-atributo*: se identifican todos los objetos de la imagen así como las relaciones entre los objetos y los atributos de los objetos.

En el caso de los documentos de texto, un diseñador de índices puede elegir las palabras clave entre el conjunto de palabras disponibles en el documento que ha de indexarse. Esto no es posible en el caso de los datos visuales y de vídeo.

Problemas en la recuperación de textos. La recuperación de textos siempre ha constituido una cuestión clave en las aplicaciones comerciales y en los sistemas bibliotecarios, y aunque se ha realizado una gran labor para solucionar algunos de estos problemas, aún existe una necesidad constante de mejora, especialmente en lo que se refiere a las cuestiones siguientes:

- *Indexación de frases*: se pueden realizar mejoras considerables si se asignan descriptores de frases a los documentos (frente a los términos de índice que constan de una única palabra) y si éstos se emplean en las consultas, siempre y cuando dichos descriptores sean buenos indicadores del contenido del documento y de la información precisada.
- *Empleo de diccionarios*: una razón de la deficiente memoria de los sistemas actuales es que el vocabulario del usuario difiere del vocabulario empleado para indexar los documentos. Una solución es la de utilizar un diccionario para ampliar la consulta del usuario con términos relacionados. Entonces, el problema radica en encontrar un diccionario del área de interés.
- *Resolución de la ambigüedad*: Una de las razones de la baja precisión (la proporción del número de elementos relevantes recuperados respecto al número total de elementos recuperados) de los sistemas de recuperación de información de textos es que las palabras tienen significados múltiples. Una forma de resolver la ambigüedad es la de emplear un diccionario *on line*; otra es la de cotejar los contextos en los que aparecen las dos palabras.

Durante las tres primeras décadas del desarrollo de SGBD (aproximadamente de 1965 a 1995), se prestó una especial atención a la gestión de datos numéricos comerciales e industriales. Es probable que durante las siguientes décadas, la información textual no numérica predomine sobre el contenido de las bases de datos. Como consecuencia de ello, se incorporarán a los SGBD una serie de funcionalidades que conllevarán la comparación, conceptualización, comprensión, indexación y resumen de documentos. En la Sección 13.2.6, estudiamos el empleo de *datablades* de texto para el SGBD INFORMIX. Los sistemas de información multimedia prometen la unificación de disciplinas que con anterioridad constituían áreas distintas: la recuperación de información y la gestión de bases de datos.

27.2.4. Aplicaciones de bases de datos multimedia

Es de esperar que las aplicaciones a gran escala de bases de datos multimedia abarquen un gran número de disciplinas y mejoren las posibilidades ya existentes. Algunas de las aplicaciones im ortantes que se verán incluidas son:

- *Gestión de documentos y registros*: un gran número de industrias y empresas llevan registros muy detallados así como diversos documentos. Los datos pueden incluir diseños de ingeniería y datos de fabricación, historiales médicos de pacientes, material de publicación y expedientes de reclamaciones de indemnización de seguros.
- *Difusión de conocimientos*: la modalidad multimedia, que es un medio muy eficaz de difusión de conocimientos, experimentará un crecimiento extraordinario en libros electrónicos, catálogos, manuales, enciclopedias y almacenes de información sobre numerosos temas.
- *Educación y formación*: se puede diseñar material de enseñanza para diferentes públicos (desde preescolares a operadores de equipo o a profesionales) a partir de fuentes multimedia. Se espera que las bibliotecas digitales ejerzan una gran influencia sobre la forma en la que

los futuros estudiantes e investigadores, al igual que otros usuarios, accederán a enormes almacenes de material educativo. (*Véase* la Sección 27.6 en torno a las bibliotecas digitales.)

- *Marketing, publicidad, comercio minorista, entretenimiento y viajes*: no existen prácticamente límites al empleo de la información multimedia en estas aplicaciones, desde presentaciones comerciales sorprendentes hasta visitas virtuales de ciudades y galerías de arte. La industria del cine ya ha mostrado el poder de los efectos especiales a la hora de crear animaciones y animales, alienígenas y efectos especiales diseñados sintéticamente. El empleo de objetos prediseñados almacenados en bases de datos multimedia ampliará el abanico de estas aplicaciones.
- *Control y supervisión en tiempo real*: Unida a una tecnología de base de datos activa (*véase* la Sección 23.1), la presentación multimedia de la información puede resultar un medio muy eficaz de supervisar y controlar tareas complejas como son las operaciones de fabricación, centrales nucleares, pacientes en unidades de cuidados intensivos y sistemas de transporte.

Sistemas comerciales para la gestión de la información multimedia. No existen SGBD diseñados con el único propósito de gestionar datos multimedia y por lo tanto, no hay ninguno que tenga el ámbito de funcionalidad necesaria para servir plenamente de soporte para todas las aplicaciones de gestión de la información multimedia que hemos descrito anteriormente. Sin embargo, en la actualidad existen diversos SGBD que sirven de soporte a tipos de datos multimedia; entre ellos se incluyen Informix Dynamic Server, DB2 Universal database (UDB) de IBM, Oracle 8.0 (*véase* el Capítulo 10), CA-JASMINE, Sybase, ODB II. Todos estos SGBD soportan objetos, lo que resulta esencial para modelar una serie de objetos complejos multimedia. Uno de los principales problemas que presentan estos sistemas es que los «*blades*, cartuchos y *extenders*» destinados a manejar datos multimedia se diseñan de un modo muy ad hoc. Se proporciona funcionalidad sin que se preste demasiada atención a la escalabilidad y al rendimiento. Existen productos que operan ya sea de manera independiente o en conjunción con sistemas de otros vendedores que permiten la recuperación de datos de imágenes por contenido. Entre estos se incluyen Virage, Excalibur y QBIC de IBM. Es necesario que las operaciones en multimedia estén estandarizadas. El MPEG-7 y otros estándares están abordando algunas de estas cuestiones.

27.2.5. Bibliografía seleccionada sobre bases de datos multimedia

La gestión de bases de datos multimedia se está convirtiendo en un área de intensa investigación sobre la que ya se están realizando varios proyectos industriales. Grosky (1994, 1997) tiene dos excelentes trabajos en torno a este tema. Pazandak y Srivastava (1995) ofrecen una valoración de los sistemas de bases de datos relacionados con los requisitos de las bases de datos multimedia. Grosky *et al.* (1997) recoge contribuciones de artículos entre los que se incluye un estudio de Jagadish (1997) sobre la indexación y recuperación basada en el contenido. Asimismo, Faloutsos *et al.* (1994) estudian un sistema para la consulta de imágenes por contenido. Li *et al.* (1998) presentan el modelado de imágenes en el que una imagen se considera un objeto complejo estructurado y jerárquico con propiedades semánticas y visuales. Nwosu *et al.* (1996) y Subramanian y Jajodia (1997) han escrito libros sobre este tema. Pueden consultarse las siguientes referencias en la WWW si se desea más información:

CA-JASMINE (SGBDO multimedia): http://www.cai.com/products/jasmine.htm
ODB II (SGBDO multimedia):
http://www.datamation.com/plugIn/inserts/FOSSI/ODB2/odb2main.html

Excalibur technologies: http://www.excalib.com
Virage, Inc (Recuperación de imágenes basada en el contenido): http://www.virage.com
Producto QBIC de IBM (*Query by Image Content* ó consulta por contenido de imagen): http://www.ibm.com

27.3. Bases de datos móviles

Los avances recientes en tecnología sin cables han dado lugar a la **informática móvil**, una nueva dimensión en la comunicación y procesamiento de datos. El entorno de la informática móvil proporcionará aspectos útiles de la tecnología sin cables a las aplicaciones de bases de datos . La plataforma de informática móvil permite a los usuarios establecer comunicación con otros usuarios y gestionar su trabajo mientras mantienen su movilidad. Esta característica resulta especialmente útil para las organizaciones que se encuentran dispersas geográficamente. Como ejemplos típicos se podrían incluir la policía de tráfico, conductores de taxi, y servicios de información meteorológica así como aplicaciones para informes sobre mercados financieros e información de corredores de bolsa. Sin embargo, existe una serie de problemas de hardware y software que deben ser resueltos antes de que se puedan utilizar plenamente las posibilidades de la informática móvil. Algunos de los problemas de software (que pueden incluir gestión de datos, gestión de transacciones y recuperación de bases de datos), tienen su origen en los sistemas de bases de datos distribuidas. No obstante, en la informática móvil estos problemas son más difíciles de resolver, debido principalmente al escaso ancho de banda de los canales de comunicación sin cables, la relativamente escasa autonomía del suministro eléctrico (batería) y de las unidades móviles, y las localizaciones cambiantes de la información requerida (a veces en la caché, otras en el aire, y otras en el servidor). Además, la informática móvil tiene sus propios retos arquitectónicos.

27.3.1. Arquitectura informática móvil

La arquitectura general de una plataforma móvil se ilustra en la Figura 27.4. Se trata de una arquitectura distribuida donde una serie de ordenadores, a los que generalmente se denomina **Computadores Fijos (CF)** y **Estaciones Base (EB),** están interconectados por medio de una red fija (cableada) de alta velocidad. Los ordenadores fijos son ordenadores de uso general que no están equipados para gestionar unidades móviles pero que pueden configurarse para tal efecto. Las estaciones base están equipadas con interfaces inalámbricas y pueden comunicarse con las unidades móviles para servir de soporte al acceso de datos.

Las **Unidades Móviles (UM) (o Computadores Móviles)** y las estaciones base se comunican a través de canales inalámbricos que tienen anchos de banda considerablemente inferiores a la de las de una red cableada. Se emplea un **canal descendente** para enviar datos desde una EB a una UM y un **canal ascendente** para enviar datos desde una UM a su EB. Los productos recientes para portátiles inalámbricos cuentan con un límite máximo de 1 Mbps (megabits por segundo) para la comunicación por infrarrojos, de 2 Mbps para la comunicación por radio, y 9,14 Kbps (kilobits por segundo) para la telefonía celular. En comparación, Ethernet proporciona 10 Mbps para Ethernet rápida, FDDI ofrece 100 Mbps y ATM (modalidad de transferencia asincrónica) proporciona 155 Mbps.

Las unidades móviles son computadores portátiles que funcionan con batería y que se transportan libremente dentro de un **dominio de movilidad geográfica**, un área restringida por la limitación de ancho de banda de los canales de comunicación inalámbrica. Para gestionar la movilidad

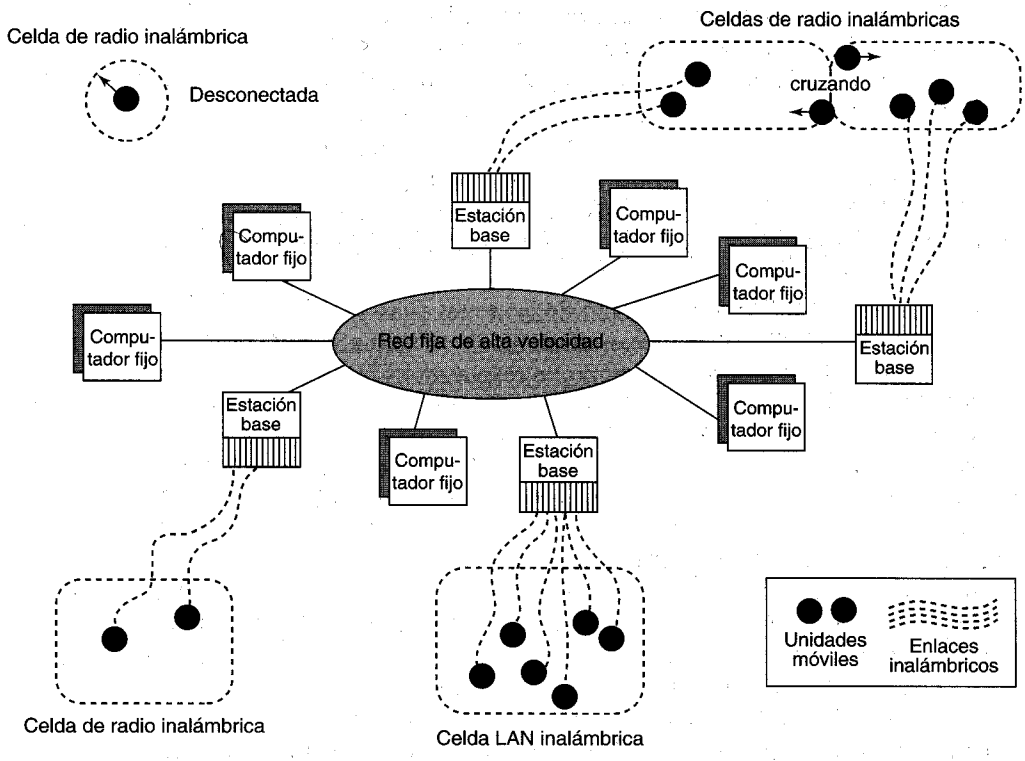

Figura 27.4. Arquitectura general de una plataforma móvil. (Adaptada de Dunham y Helal (1995).)

de las unidades, la totalidad del dominio de movilidad geográfica se divide en dominios menores denominados **celdas**. La disciplina móvil requiere que el movimiento de las unidades móviles no esté restringido dentro del dominio de movilidad geográfica (movimiento entre celdas), en tanto que disponer de **contigüidad de acceso** a la información durante el movimiento garantiza que el movimiento de una unidad móvil a través de los límites de las celdas no afectará al proceso de recuperación de datos.

La plataforma de informática móvil puede describirse eficazmente bajo el paradigma cliente-servidor, lo que significa que podemos referirnos a una unidad móvil unas veces como cliente y otras como usuario, y a las estaciones base como servidores. Cada celda es gestionada por una estación base, que contiene transmisores y receptores para dar respuesta a las necesidades de procesamiento de información de los clientes ubicados en la celda. Suponemos que el tiempo medio de respuesta a consultas es mucho menor que el tiempo que precisa el cliente para atravesar físicamente una celda.

Los clientes y servidores se comunican por medio de canales inalámbricos. El enlace de comunicación entre un cliente y un servidor puede modelarse a modo de canales de datos múltiples o de un único canal.

Características de los entornos móviles. En los entornos de bases de datos móviles, por lo general los datos varían con gran rapidez. Con frecuencia, los usuarios consultan a los servidores para mantenerse actualizados. Más concretamente, éstos desean a menudo hacer un seguimiento de cada transmisión de elementos de datos de su interés. Entre los ejemplos de este tipo de datos se encuentran la información sobre el mercado bursátil, datos meteorológicos e información sobre

compañías aéreas. La base de datos se actualiza de forma asíncrona mediante un proceso externo independiente.

Los usuarios son móviles y entran y salen de las celdas arbitrariamente. La duración media de la estancia de un usuario en la celda se conoce con el nombre de **latencia de residencia (LR)**, un parámetro que se calcula (y se ajusta continuamente) mediante la observación de los tiempos de permanencia (residencia) del usuario en las celdas. De este modo, cada celda tiene un valor de LR. El comportamiento de referencia del usuario tiende a ser localizado, es decir, los usuarios tienden a acceder a ciertas partes de la base de datos con mucha frecuencia. Los servidores no conservan ni los patrones de llegada y salida del cliente ni la información de solicitud de datos específica del cliente.

Las redes inalámbricas difieren de las redes fijas en muchos aspectos. Los usuarios de bases de datos de una red fija permanecen conectados no sólo a la red sino también a una fuente de energía continua. Por este motivo, el tiempo de respuesta es la medida clave del rendimiento. Sin embargo, en una red inalámbrica son importantes tanto el tiempo de respuesta como la autonomía de la fuente de energía del usuario (batería). Mientras una unidad móvil está recibiendo o trasmitiendo *on-line*, se considera que está en modalidad activa. Para los ordenadores portátiles actuales con unidades de CD-ROM, la duración aproximada de la batería en modo activo es inferior a 3 horas. Con el fin de ahorrar energía y ampliar la duración de la batería, los clientes pueden cambiar a la **modalidad suspendida (*doze mode*)**, en la que éstos no están conectados al canal de forma activa y pueden consumir una cantidad de energía considerablemente inferior que si se encuentran en modalidad activa. Los clientes pueden recibir aviso para salir de la modalidad suspendida cuando el servidor necesite comunicarse con el cliente.

27.3.2. Tipos de datos en aplicaciones móviles

Las aplicaciones que operan en computadores móviles tienen diferentes requisitos de datos. Los usuarios se dedican a realizar comunicaciones personales o actividades administrativas, o simplemente reciben actualizaciones sobre información que varía constantemente. Las aplicaciones móviles pueden clasificarse de dos maneras: (1) aplicaciones verticales y (2) aplicaciones horizontales.[3] En las **aplicaciones verticales** los usuarios acceden a los datos dentro de una celda específica y se niega el acceso a aquellos usuarios que se encuentran fuera de dicha celda. Por ejemplo, los usuarios pueden obtener información sobre la localización de médicos o centros de urgencias dentro de una celda o datos sobre la disponibilidad de plazas de aparcamiento en una celda del aeropuerto. En las **aplicaciones horizontales**, los usuarios cooperan a la hora de llevar a cabo la tarea, y pueden manejar datos distribuidos por todo el sistema. El mercado de aplicaciones horizontales es enorme; los dos tipos de aplicaciones más populares son las aplicaciones que dan acceso al correo electrónico y los servicios de información a los usuarios móviles.

Los datos pueden clasificarse en tres categorías:

1. *Datos privados*: el propietario de estos datos es un solo usuario que es quien los maneja. Ningún otro usuario puede acceder a ellos.
2. *Datos públicos*: estos datos los puede utilizar todo aquél que pueda leerlos. Los actualiza una sola fuente. Ejemplos de estos datos son los partes meteorológicos o los precios de las acciones.
3. *Datos compartidos*: a estos datos acceden grupos de usuarios tanto en las modalidades de escritura como de lectura. Ejemplos de éstos son los datos de inventario de los productos de una empresa.

[3] Esta clasificación de aplicaciones y tipos de datos la proponen Imielinski y Badrinath (1994).

Los datos públicos se gestionan principalmente mediante aplicaciones verticales, mientras que las aplicaciones horizontales emplean los datos compartidos, posiblemente con alguna replicación. Las copias de los datos compartidos pueden almacenarse tanto en estaciones base como móviles. Esto presenta una serie de dificultades para la consistencia de la gestión de las transacciones así como para la integridad y escalabilidad de la arquitectura.

27.3.3. Cuestiones de la gestión de datos

Desde el punto de vista de la gestión de datos, la informática móvil puede considerarse como una variante de la informática distribuida. Las bases de datos móviles pueden distribuirse en dos ámbitos posibles:

1. La base de datos completa se distribuye principalmente entre los componentes conectados, posiblemente con replicación total o parcial. Una estación base gestiona su propia base de datos con una funcionalidad similar a la del SGBD, con una funcionalidad adicional para localizar unidades móviles y características adicionales de gestión de consultas y transacciones destinadas a satisfacer los requisitos de los entornos móviles.
2. La base de datos se distribuye entre componentes conectados e inalámbricos. La responsabilidad de la gestión de datos es compartida entre las estaciones base y las unidades móviles.

Por lo tanto, las cuestiones sobre la gestión bases de datos distribuidas que abordamos en el Capítulo 24 también pueden aplicarse a las bases de datos móviles con las siguientes consideraciones y variantes adicionales:

1. *Distribución y replicación de datos*: los datos se distribuyen desigualmente entre las estaciones base y las unidades móviles. Las restricciones de consistencia agravan el problema de la gestión de la caché. Las cachés intentan proporcionar a las unidades móviles los datos que son accedidos y actualizados más frecuentemente. Las unidades móviles procesan sus propias transacciones y pueden estar desconectadas durante largos períodos.
2. *Modelos de transacción*: las cuestiones de tolerancia a fallos y corrección de las transacciones se ven agravadas en el entorno móvil. Una transacción móvil se ejecuta secuencialmente a través de diversas estaciones base y posiblemente en múltiples conjuntos de datos que dependen del movimiento de la unidad móvil. No existe una coordinación central de la ejecución de la transacción, especialmente en el ámbito (2) descrito anteriormente. Por lo tanto, puede que sea necesario modificar las tradicionales propiedades ACID de las transacciones (*véase* el Capítulo 19) y definir nuevos modelos de transacción.
3. *Procesamiento de consultas*: es importante saber dónde se encuentran los datos. Esto repercute en el análisis del coste/beneficio del procesamiento de las consultas. La respuesta a la consulta ha de ser devuelta a las unidades móviles que pueden encontrarse en tránsito o que pueden cruzar los límites de las celdas. A pesar de ello deben recibir los resultados de la consulta completos y correctos.
4. *Recuperación y tolerancia a los fallos*: el entorno de bases de datos móviles debe hacer frente a fallos de sitios, de medios, de transacciones y de comunicación. El fallo de sitio en una UM se debe frecuentemente a que la batería tiene un límite de energía disponible. Si una UM realiza una interrupción voluntaria, ésta *no* debería considerarse como un fallo. Los fallos en las transacciones son más frecuentes durante el proceso conocido como *hand-off* que tiene lugar cuando una UM atraviesa las celdas. El fallo en la UM provoca una partición de la red y afecta a los algoritmos de encaminamiento (*routing*).

5. *Diseño de bases de datos móviles*: el problema de la resolución de nombres globales para tratar las consultas se ve agravado debido a la movilidad y a las frecuentes interrupciones. El diseño de bases de datos móviles debe considerar numerosas cuestiones de gestión de metadatos, por ejemplo, la constante actualización de la información sobre la localización.

27.3.4. Bases de datos móviles sincronizadas intermitentemente

Un escenario diferente promete convertirse en algo cada vez más común, a medida que la gente se lleva el trabajo fuera de sus oficinas y de sus hogares y realizan una amplia gama de actividades y funciones: todo tipo de ventas, particularmente en el sector farmacéutico, bienes de consumo y componentes industriales; la aplicación de la ley; el asesoramiento y la planificación de seguros y financiera; actividades de gestión de bienes inmuebles o propiedades, etc. En estas aplicaciones, un servidor o grupo de servidores controla la base de datos central y los clientes transportan los computadores portátiles (*laptop*) y computadores de bolsillo (*palmtop*) con un software de SGBD interno para realizar una actividad de transacciones «locales» la mayor parte del tiempo. Los clientes se conectan con el servidor a través de una red o de una conexión telefónica (o posiblemente incluso a través de Internet), por lo general durante un breve espacio de tiempo, digamos que de 30 a 60 minutos. Estos envían sus actualizaciones al servidor, y el servidor debe a su vez introducirlos en su base de datos central, que debe mantener los datos actualizados y preparar copias apropiadas para todos los clientes que se encuentren en el sistema. De este modo, cada vez que los clientes se conecten (a través de un proceso que se conoce en la industria como la sincronización de un cliente con un servidor), estos reciben una serie de actualizaciones que han de instalarse en su base de datos local. La característica principal de este escenario es que los clientes se encuentran en su mayor parte desconectados; el servidor no tiene por qué ser capaz de llegar al cliente. Este entorno tiene dificultades similares a los de las bases de datos distribuidas y a las de cliente-servidor, y algunos procedentes de las bases de datos móviles, pero presenta varios problemas de investigación adicionales para su estudio. Nos referiremos a este entorno como **Entorno de Bases de Datos Sincronizadas Intermitentemente (ISDBE)**,[4] y a las bases de datos correspondientes como Bases de Datos Sincronizadas Intermitentemente (ISDB).[5]

Las siguientes características de las ISDB las *diferencian* de las bases de datos móviles que hemos estudiado hasta ahora:

1. Un cliente se conecta al servidor cuando desea recibir actualizaciones de un servidor o enviar sus actualizaciones a un servidor o procesar transacciones que precisan datos no locales. Esta comunicación puede ser *unidestinatario* (una comunicación uno a uno entre el servidor y el cliente) o *multidestinatario* (un emisor o servidor puede comunicarse periódicamente con un conjunto de ordenadores o actualizar un grupo de clientes).
2. Un servidor no puede conectarse con el cliente cuando lo desea.
3. Las cuestiones sobre si las conexiones de los clientes son inalámbricas o bien son por cable, y el ahorro de la energía son irrelevantes.
4. Un cliente es libre de gestionar sus propios datos y transacciones mientras esté desconectado. También puede realizar su propia recuperación de datos hasta cierto punto.
5. Un cliente tiene múltiples formas de conectarse a un servidor y, en el caso de muchos servidores, éste puede elegir un servidor concreto al que conectarse de acuerdo a su proximidad, nodos de comunicación disponibles, etc.

[4] ISDBE significa *Intermittently Synchhronized Database Environment.*
[5] ISDB significa *Intermittently Synchronized Databases.*

Debido a estas diferencias, hay una necesidad de abordar una serie de problemas relacionados con las ISDB que difieren de los que atañen generalmente a los sistemas de bases de datos móviles. Entre estos se incluye el diseño de la base de datos del servidor para las bases de datos del servidor, la gestión de la coherencia entre el procesamiento de las transacciones y actualizaciones de las bases de datos del cliente y del servidor, el uso eficiente del ancho de banda del servidor, y el logro de la escalabilidad en los entornos de las ISDB.

27.3.5. Bibliografía seleccionada para las bases de datos móviles

Se ha producido un repentino aumento del interés por la informática móvil, y la investigación en torno a las bases de datos móviles ha experimentado un crecimiento significativo durante los últimos cinco o seis años. Entre los libros publicados en torno a este tema, Dhawan (1997) constituye una fuente excelente en lo que respecta a la informática móvil. Holtzman y Goodman (1993) examinan las redes inalámbricas y su futuro. Imielinski y Badrinath (1994) proporcionan un interesante análisis sobre cuestiones relacionadas con las bases de datos. Dunham y Helal (1995) abordan los problemas del procesamiento de consultas, la distribución de datos y la gestión de las transacciones en las bases de datos móviles. Foreman y Zahorjan (1994) describen las posibilidades y problemas de la informática móvil y ofrecen argumentos en su favor como una solución viable para numerosas aplicaciones de sistemas de información en el futuro. Pitoura y Samaras (1998) describen todos los aspectos de los problemas y soluciones de las bases de datos móviles. Chintalapati *et al.* (1997) proporcionan un algoritmo de gestión de localización adaptativa mientras que Bertino *et al.* (1998) examinan los métodos para la tolerancia a fallos y la recuperación en las bases de datos móviles. El número de junio 1995 de la revista *Byte* trata numerosos aspectos de la informática móvil. Para un examen inicial de las cuestiones sobre la escalabilidad en las ISBD y un enfoque según la agregación de datos y agrupación de clientes, véase Mahajan *et al.* (1998).

27.4. Sistemas de información geográfica

Los **sistemas de información geográfica (GIS)**[6] se emplean para recoger, modelar, almacenar y analizar información que describe las propiedades físicas del mundo geográfico. En líneas generales, el ámbito de los GIS abarca dos tipos de datos: (1) datos espaciales, procedentes de mapas, imágenes digitales, fronteras administrativas y políticas, carreteras, redes de transporte; datos físicos tales como ríos, características del suelo, regiones climáticas, elevaciones del terreno, y (2) datos no espaciales como cómputos del censo, datos económicos e información sobre ventas o marketing. Los GIS constituyen un dominio de rápido desarrollo que ofrecen métodos sumamente innovadores para hacer frente a algunas demandas técnicas que constituyen un reto.

27.4.1. Las aplicaciones GIS

Es posible dividir los GIS en tres categorías: (1) aplicaciones cartográficas, (2) aplicaciones para el modelado digital de terrenos, y (3) aplicaciones de objetos geográficos. La Figura 27.5 resume estas categorías.

En las aplicaciones cartográficas y de modelado de terrenos, se capturan variedad de atributos espaciales, por ejemplo, las características del suelo, densidad de cultivos y calidad del aire. En las

[6] GIS significa *Geographic Information System.*

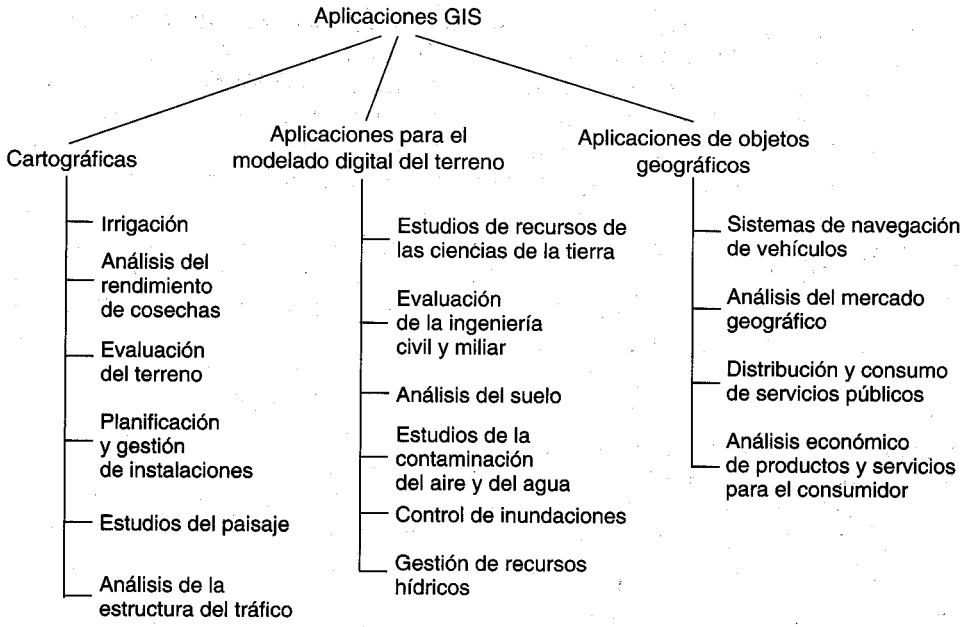

Figura 27.5. Una clasificación posible de las aplicaciones GIS. (Adaptada de Adam y Gangopadhyay (1997).)

aplicaciones de objetos geográficos, se identifican los objetos de interés de un dominio físico, por ejemplo, centrales eléctricas, distritos electorales, parcelas de terreno, áreas de distribución de productos y edificios o lugares emblemáticos de una ciudad. Estos objetos están relacionados con datos de aplicaciones pertinentes, que pueden ser, para este ejemplo concreto, el consumo eléctrico, las pautas de voto, los volúmenes de venta de propiedades, el volumen de venta de productos y la densidad del tráfico.

Las dos primeras categorías de aplicaciones GIS requieren una representación basada en campos, mientras que la tercera categoría precisa de una basada en objetos. Las aplicaciones de tipo cartográfico conllevan unas funciones especiales que pueden incluir la superposición de varias capas de mapas para combinar datos de atributos que permitirán, por ejemplo, la medición de distancias en un espacio tridimensional y la reclasificación de datos en el mapa. El modelado digital de terrenos requiere una representación digital de partes de la superficie terrestre utilizando elevaciones del terreno en puntos de muestreo que se interconectan para dar lugar a un modelo de superficie como, por ejemplo, una red tridimensional (líneas conectadas en 3D) que muestre la superficie del terreno. Esto precisa de unas funciones de interpolación entre los puntos observados así como unas funciones de visualización. En las aplicaciones geográficas basadas en objetos, se necesitan funciones espaciales adicionales para manejar los datos referentes a carreteras, conductos físicos, cables de comunicación, cables de alta tensión, etc. Por ejemplo, para una región dada, se pueden emplear mapas comparables para contrastarlos en diferentes puntos del tiempo y mostrar los cambios producidos en determinados datos como son la ubicación de carreteras, cables, edificios y ríos.

27.4.2. Requisitos de los GIS para la gestión de datos

Los requisitos funcionales de las aplicaciones GIS descritas anteriormente se traducen en los siguientes requisitos para las bases de datos.

Modelado y representación de datos. En líneas generales, los datos GIS pueden representarse en dos formatos: (1) vector y (2) matriz de puntos (*raster*). Los datos del vector representan objetos geométricos como son puntos, líneas y polígonos. De este modo, un lago puede representarse como un polígono, un río como una serie de segmentos de línea. Los datos de la matriz de puntos se caracterizan por una serie de puntos, en la que cada punto representa el valor de un atributo para una localización del mundo real. De manera informal, las imágenes de las matrices de puntos son series n-dimensionales en las que cada entrada constituye una unidad de la imagen y representa un atributo. Las unidades bidimensionales se denominan *pixels*, mientras que las unidades tridimensionales reciben el nombre de *voxels*. Los datos de elevación tridimensional se almacenan en un formato de **modelo de elevación digital (DEM)**[7] basado en matrices de puntos. Otro formato de matriz de puntos denominado **red irregular triangular (TIN)**[8] es un método topológico basado en vectores que modela las superficies conectando puntos de muestra a modo de vértices de triángulos y tiene una densidad de puntos que puede variar con la aspereza del terreno. Las cuadrículas rectangulares (o matrices de elevación) son estructuras de series bidimensionales. En el **modelado digital del terreno (DTM)**,[9] también puede emplearse el modelo sustituyendo la elevación por algún atributo de interés como puede ser la densidad de población o la temperatura ambiental. Los datos GIS incluyen a menudo una estructura temporal además de la estructura espacial. Por ejemplo, la densidad del tráfico puede medirse cada 60 segundos en un conjunto de puntos.

Análisis de datos. Los datos GIS experimentan diversos tipos de análisis. Por ejemplo, en aplicaciones como los estudios de la erosión del suelo, estudios sobre el impacto medioambiental, o en simulaciones de afluencia hidrológica, los datos DTM pueden experimentar varios tipos de **análisis geomorfométricos**, mediciones como son los valores de laderas, *gradiente* (la tasa de cambio en altitud), *aspecto* (la dirección magnética del gradiente), *convexidad del perfil* (la tasa de de cambio en el gradiente), *convexidad de plano* (la convexidad de contornos y otros parámetros). Cuando se emplean los datos GIS para aplicaciones de soporte a la toma de decisiones, estos pueden sufrir operaciones de agregación y expansión mediante el almacenamiento de datos, como ya vimos en la Sección 26.1.5. Además, se realizan operaciones geométricas (para calcular distancias, áreas y volúmenes), operaciones topológicas (para calcular superposiciones, intersecciones, caminos más cortos), y operaciones temporales (para calcular consultas internas o basadas en eventos). El análisis conlleva una serie de operaciones temporales y espaciales que ya tratamos en las Secciones 23.2 y 23.3.

Integración de datos. Los GIS deben integrar datos tanto de vectores como de matrices de puntos procedentes de diversas fuentes. Algunas veces se infieren los bordes y regiones de una imagen de matriz de puntos para formar un modelo de vector, o a la inversa, se emplean imágenes de matriz de puntos como fotografías aéreas para actualizar modelos de vectores. Se emplean diversos sistemas coordinados como el *Universal Transverse Mercator* (UTM), sistemas de latitud/longitud y sistemas catastrales locales para identificar localizaciones. Los datos procedentes de diferentes sistemas coordinados requieren unas transformaciones adecuadas. Muchas herramientas para el trazado de mapas basadas en la Web (por ejemplo, http://maps.yahoo.com) hacen uso de las principales fuentes públicas de datos geográficos, entre los que se incluyen los ficheros TIGER de los que se ocupa el Ministerio de Comercio de los EE.UU., para el trazado de mapas de carreteras. Con frecuencia, hay mapas de gran precisión y pocos atributos que tienen que unirse con mapas de poca precisión y que contienen numerosos atributos. Esto se realiza mediante un proceso denominado

[7] *Digital Elevation Model.*
[8] *Triangular Irregular Network.*
[9] *Digital Terrain Modeling.*

«rubber banding» donde, el usuario define un conjunto de puntos de control en ambos mapas y se realiza la transformación del mapa de menor precisión para alinear los puntos de control. Una cuestión de integración fundamental es la de crear y mantener la información de atributos (como son la calidad del aire o la densidad del tráfico) que con el tiempo pueden relacionarse e integrarse con la información geográfica correspondiente a medida que ambos evolucionan.

Captura de datos. El primer paso a la hora de crear una base de datos espacial para el modelado cartográfico es el de obtener información geográfica bidimensional o tridimensional en forma digital, un proceso que a veces se ve dificultado por las características del mapa original tales como la resolución, el tipo de proyección, las escalas de los mapas, la autorización cartográfica, la diversidad de técnicas de medición así como las diferencias del sistema de coordenadas. Asimismo, también pueden obtenerse datos espaciales procedentes de sensores remotos en satélites como Landsat, NORA, y *Advanced Very High Resolution Radiometer* (AVHRR: Radiómetro de Resolución Muy Alta Avanzado) así como SPOT HRV (*High Resolution Visible Range Instrument* o Instrumento de Campo Visible de Alta Resolución), que está libre de sesgo interpretativo y es muy preciso. En lo que se refiere al modelado digital del terreno, los métodos de captación de datos varían desde los manuales hasta los plenamente automatizados. Los estudios del terreno suelen ser el método tradicional y el más preciso, pero llevan mucho tiempo. Otras técnicas incluyen el muestreo fotogramétrico y la digitalización de documentos cartográficos.

27.4.3. Operaciones específicas de datos GIS

Las aplicaciones GIS se llevan a cabo mediante el empleo de los siguientes operadores especiales:

- *Interpolación*: este proceso obtiene datos de elevación para puntos en los que no se han obtenido muestras. Incluye el cálculo para un solo punto, el cálculo de una cuadrícula rectangular o de un contorno, etc. La mayoría de los métodos de interpolación se basan en la triangulación que emplea el método TIN para interpolar elevaciones dentro del triángulo tomando como base las de sus vértices.
- *Interpretación*: el modelado digital del terreno conlleva la interpretación de operaciones sobre datos del terreno como son la definición, el allanado, la reducción de detalles y su mejora. Las operaciones adicionales implican el arreglo o cierre de los bordes de los triángulos (en los datos TIN), y la fusión, lo que supone la combinación de modelos que se superponen y la resolución de conflictos entre los datos de los atributos. Las conversiones entre los modelos de cuadrículas, los modelos de contornos y los datos TIN son aspectos integrantes de la interpretación del terreno.
- *Análisis de proximidad*: varias clases de análisis de proximidad incluyen el cálculo de «zonas de interés» en torno a objetos, como son el establecimiento de un búfer en torno a un vehículo en una autopista. Los algoritmos de camino más corto mediante el empleo de información en 2D o 3D constituyen un tipo importante de análisis de proximidad.
- *Procesamiento de imágenes en una matriz de puntos*: este proceso puede dividirse en dos categorías (1) el álgebra de mapas, que se emplea para integrar características geográficas en diferentes capas de mapas para dar lugar a nuevos mapas algebraicamente; y (2) el análisis digital de imágenes, que se encarga del análisis de una imagen digital para características como son la detección de bordes y la detección de objetos. La detección de carreteras en una imagen de una ciudad obtenida por satélite es un ejemplo de esta última.
- *Análisis de redes*: las redes aparecen en los GIS en numerosos contextos que deben analizarse y que pueden estar sujetos a segmentaciones, superposiciones, etc. La superposición de

redes hace referencia a un tipo de unión espacial en la que una red dada, por ejemplo una red de autopistas, se combina con una base de datos concreta (por ejemplo, localizaciones de accidentes) para dar lugar, en este caso, a un perfil de carreteras con una alta siniestralidad.

Otra funcionalidad de las bases de datos. La funcionalidad de una base de datos GIS está también sujeta a otras consideraciones.

- *Extensibilidad*: es preciso que los GIS sean extensibles para dar cabida a una serie de aplicaciones en constante evolución y a los tipos de datos correspondientes. Si se emplea un SGBD estándar, éste debe permitir un conjunto básico de tipos de datos con recursos para definir nuevos tipos y métodos para dichos tipos.
- *Control de calidad de los datos*: como en muchas otras aplicaciones, la calidad de los datos originales es de primordial importancia a la hora de proporcionar unos resultados precisos a las consultas. Este problema resulta especialmente significativo en el contexto de los GIS debido a la variedad de datos, fuentes y técnicas de medición empleadas y a la absoluta precisión que esperan los usuarios de las aplicaciones.
- *Visualización*: una función crucial de los GIS está relacionada con la visualización (la exposición gráfica de la información del terreno y la representación correspondiente de los atributos de la aplicación). Las principales técnicas de visualización incluyen (1) el *contorneado* mediante el empleo de *isolíneas*, unidades espaciales de líneas o arcos de valores de atributos iguales; (2) *sombreado de montañas*, un método de iluminación empleado para la representación cualitativa de relieves empleando intensidades distintas de luz para las facetas individuales del modelo de terreno; y (3) *visualizaciones de perspectivas*, imágenes tridimensionales de las facetas del modelo de terreno mediante la utilización de métodos de proyección de perspectivas de los gráficos creados por ordenador. Estas técnicas aplican datos cartográficos y otros objetos tridimensionales a los datos del terreno proporcionando versiones animadas de las escenas como ocurre en las simulaciones de vuelo y en películas de dibujos animados.

Este tipo de requisitos muestran claramente que los SGBDR (SGBD relacionales) o SGBDO (SGBD de objetos) no satisfacen las necesidades concretas de los GIS. Por lo tanto, es preciso diseñar sistemas que sirvan de soporte para las representaciones en forma de vector y de matriz de puntos y para la funcionalidad espacial así como las características necesarias del SGBD. En la subsección siguiente, se examina brevemente un conocido GIS denominado ARC-INFO, que *no* es un SGBD pero que integra la funcionalidad de un SGBDR en la parte INFO del sistema. Es probable que en el futuro se diseñen más sistemas que operen con bases de datos relacionales u orientadas a objetos y que contengan algo de información espacial y la mayor parte de información no espacial.

27.4.4. Ejemplo de un GIS: ARC-INFO

ARC/INFO, un conocido GIS lanzado en 1981 por el *Environmental System Research Institute* (ESRI), emplea el modelo de nodo de arco para almacenar datos espaciales. Una disposición geográfica, denominada *cobertura* en ARC/INFO, consta de tres componentes básicos: (1) nodos (puntos), (2) arcos (similares a líneas), y (3) polígonos. El arco es el más importante de los tres y almacena gran cantidad de información topológica. Un arco consta de una nodo inicial y un nodo final (y, por lo tanto, tiene también dirección). Además, los polígonos a la derecha e izquierda del arco también se almacenan junto con cada arco. Dado que no existe restricción en lo que a la forma del

arco se refiere, los puntos de la forma que no contienen información topológica también se almacenan junto con cada arco. De este modo, la base de datos gestionada por el SGBDR INFO está formada necesariamente por tres tablas: (1) tabla de atributos de nodos (NAT),[10] (2) tabla de atributos de arcos (AAT),[11] y (3) tabla de atributos de polígonos (PAT).[12] La información adicional puede almacenarse en tablas distintas y combinarse con cualquiera de estas tres tablas.

La NAT contiene un identificador (ID) interno para el nodo, un ID especificado por el usuario, las coordenadas del nodo, y cualquier otra información relacionada con dicho nodo (por ejemplo, los nombres de las carreteras con intersección en ese nodo). La AAT contiene un ID interno para el arco, un ID especificado por el usuario, el ID interno de los nodos inicial y final, el ID interno de los polígonos a la derecha e izquierda, una serie de coordenadas de los puntos de la forma (si los hay), la longitud del arco, y cualquier otro dato relacionado con el arco (por ejemplo, el nombre de la carretera que representa el arco). La PAT está formada por un ID interno para el polígono, un ID especificado por el usuario, el área del polígono, el perímetro del polígono, y cualquier otro dato relacionado (por ejemplo, el nombre del municipio que representa el polígono).

Las consultas espaciales habituales están relacionadas con la adyacencia, la contención y la conectividad. El modelo de nodo de arco contiene suficiente información como para satisfacer los tres tipos de consultas, pero el SGBDR no resulta adecuado para este tipo de consultas. Un simple ejemplo pondrá de relieve el número de veces que debe consultarse una base de datos relacional para obtener información sobre adyacencia. Supongamos que intentamos determinar si dos polígonos, A y B, son adyacentes entre sí. Tendríamos que examinar toda la AAT de manera exhaustiva para establecer si hay un borde que tiene a A en un lado y a B en el otro. La búsqueda no puede limitarse a los bordes de cada polígono puesto que no almacenamos explícitamente todos los arcos que constituyen un polígono en la PAT. El almacenaje de todos los arcos en la PAT resultaría redundante dado que toda la información ya está contenida en la AAT.

ESRI ha creado Arc/Storm (*Arc Store Manager*), el cual permite que numerosos usuarios empleen el mismo GIS, maneja bases de datos distribuidas, y se integra con otros SGBDR comerciales como ORACLE, INFORMIX y SYBASE. Aunque ofrece muchas ventajas funcionales y de rendimiento respecto a ARC/INFO, es esencialmente un SGBDR incorporado a un GIS.

27.4.5. Problemas y cuestiones futuras en los GIS

Los GIS constituyen un área de aplicación de bases de datos en expansión, que refleja una explosión en el número de usuarios finales que emplean mapas digitalizados, datos sobre terrenos, imágenes espaciales, datos meteorológicos, así como datos de soporte de información del tráfico. Como consecuencia de ello, ha surgido una serie de problemas cada vez mayor que afectan a las aplicaciones GIS y que han de resolverse:

- *Nuevas arquitecturas*: las aplicaciones GIS precisarán una nueva arquitectura cliente-servidor que se beneficiará de los avances ya existentes de la tecnología de los SGBDR y SGBDO. Una solución posible es la de separar los datos espaciales de los no espaciales para que estos últimos puedan ser totalmente controlados por un SGBD. Un proceso de este tipo requiere un modelado y una integración adecuadas, a medida que evolucionan ambos tipos de datos. Los distribuidores comerciales consideran que resulta más viable contar con un nú-

[10] *Node Attribute Table.*
[11] *Arc Attribute Table.*
[12] *Polygon Attribute.*

mero reducido de bases de datos independientes que se envíen entre sí las actualizaciones de modo automático. Para ello, se necesitarán unas herramientas adecuadas para la transferencia de datos, la gestión de cambios y la gestión del flujo de trabajo.

- *Versionado y método del ciclo de vida del objeto*: debido a que las características geográficas están en constante evolución, los GIS deben mantener unos minuciosos datos cartográficos y de terreno, un problema de gestión que podría reducirse mediante una actualización incremental combinada con unos esquemas de autorización de actualizaciones para diferentes niveles de usuarios. Con el método del ciclo de vida del objeto, que abarca las actividades de creación, destrucción y modificación de objetos así como la promoción de versiones en objetos permanentes, se puede predefinir un conjunto completo de métodos para controlar estas actividades para los objetos GIS.
- *Estándares de datos*: debido a la diversidad de los esquemas y modelos de representación, la formalización de los estándares de transferencia de datos resulta crucial para el éxito de los GIS. La organización para la estandarización internacional (ISO TC211) y la organización de estándares europeos (CEN TC278) se encuentran en el proceso de debate de cuestiones relevantes, entre ellas la conversión de datos entre vector y matriz de puntos para un rápido rendimiento de las consultas.
- *Aplicaciones equiparables y estructuras de datos*: si observamos de nuevo la Figura 27.5, veremos que una clasificación de las aplicaciones GIS se basa en la naturaleza y organización de los datos. En el futuro, los sistemas que comprendan un amplio abanico de funciones (desde análisis de mercados y servicios públicos hasta navegación de vehículos) precisarán de datos en torno a fronteras así como de funcionalidad. Por otra parte, las aplicaciones en ciencia medioambiental, hidrología y agricultura requerirán datos más orientados al área y al modelo de terreno. No resulta evidente que un único GIS de ámbito general pueda servir de soporte a toda esta funcionalidad. Las necesidades especializadas de los GIS harán necesario que deban mejorarse los SGBD de uso general con tipos de datos y funcionalidad adicionales antes de que puedan servir de soporte a aplicaciones GIS completamente desarrolladas.
- *Ausencia de semántica en las estructuras de datos*: esto resulta especialmente patente en los mapas. La información relativa a cruces en autopistas y carreteras puede resultar difícil de precisar tomando como base los datos almacenados. Las vías de sentido único también son difíciles de representar en los GIS actuales. Los sistemas de transporte CAD han incorporado este tipo de semántica en los GIS.

27.4.6. Bibliografía seleccionada para los GIS

Hay una serie de libros en torno a los GIS. Adam y Gangopadhyay (1997) y Laurini y Thompson (1992) analizan los problemas de la gestión de las bases de datos GIS y la gestión de la información. Kemp (1993) da una visión general de las cuestiones y fuentes de datos de los GIS. Maguire *et al.* (1991) tienen un conjunto de artículos sobre los GIS. Sarasua y O'Neill (1999) se centran en los GIS para los sistemas de transporte. El Ministerio de Comercio de los EE.UU. (1993) está a cargo de los ficheros TIGER sobre datos viarios. El sitio Web de Laser-Scan (http://www.lsl.co.uk/papers) constituye una buena fuente de información.

El Environmental System Research Institute (ESRI, Instituto de Investigación de Sistemas Medioambientales) cuenta con una biblioteca excelente de publicaciones en torno a los GIS para todos los niveles en la dirección http://www.esri.com. La terminología GIS se define en la siguiente dirección: http://www.esri.com/library/glossary/glossary.html.

27.5. Gestión de datos del genoma

27.5.1. Ciencias biológicas y genética

Las ciencias biológicas abarcan una enorme variedad de información. La ciencia medioambiental nos ofrece una visión del modo en el que viven e interactúan las especies, en un mundo lleno de fenómenos naturales. La biología y la ecología estudian especies concretas. La anatomía se centra en la estructura general de un organismo, documentando los aspectos físicos de los organismos individuales. La medicina y la fisiología tradicionales dividen el organismo en sistemas y tejidos e intentan obtener información sobre el funcionamiento de estos sistemas y sobre el organismo en general. La histología y la biología celular ahondan en los niveles de los tejidos y de las células y proporcionan conocimientos sobre la estructura y función internas de la célula. Esta riqueza de información que se ha generado, clasificado y almacenado durante siglos se ha convertido tan sólo muy recientemente en una aplicación fundamental de la tecnología de bases de datos.

La **genética** ha surgido como un campo ideal para la aplicación de la tecnología de la información. En un sentido amplio, puede considerarse como la construcción de modelos basada en la información sobre genes (que pueden definirse como las unidades básicas de herencia) y poblaciones, y la búsqueda de relaciones en esa información. El estudio de la genética puede dividirse en tres ramas: (1) genética mendeliana, (2) genética molecular, y (3) genética de poblaciones. La genética mendeliana constituye el estudio de la transmisión de características entre generaciones. La genética molecular es el estudio de la estructura química y función de los genes a escala molecular. La genética de poblaciones es el estudio del modo en el que varía la información genética a través de colonias de organismos.

La genética molecular proporciona un examen más detallado de la información genética al permitir que los investigadores examinen la composición, estructura y función de los genes. Los orígenes de la genética molecular pueden remontarse a dos descubrimientos importantes. El primero tuvo lugar en 1869 cuando Friedrich Miescher descubrió la nucleína y su principal componente, el ácido desoxirribonucleico (ADN). En investigaciones posteriores, se descubrió que el ADN y su compuesto asociado, el ácido ribonucleico (ARN), estaban compuestos de nucleótidos (un azúcar, un fosfato y una base, que se combinaban para constituir el ácido nucleico) enlazados en polímeros de gran longitud por medio del azúcar y del fosfato. El segundo descubrimiento fue la demostración que realizó Oswald Avery en 1944 de que el ADN era realmente la sustancia molecular que portaba la información genética. De ese modo, se demostró que los genes estaban formados por cadenas de ácidos nucleicos dispuestas linealmente en los cromosomas y que cumplían tres funciones principales: (1) reproducir la información genética entre generaciones, (2) proporcionar moldes para la creación de polipéptidos, y (3) acumular cambios, permitiendo así que se produzca la evolución.

27.5.2. Características de los datos biológicos

Los datos biológicos ponen de manifiesto numerosas características especiales que hacen de la gestión de información biológica un problema que supone un reto especial. Por ello, comenzaremos haciendo un resumen de las características relacionadas con la información biológica, y centrándonos en un campo multidisciplinar denominado **bioinformática,** que en la actualidad cuenta con programas de licenciaturas ya implantados en varias universidades. La bioinformática aborda la gestión de la información genética haciendo especial hincapié en el análisis de la secuencia del ADN. Ésta ha de ampliarse a un ámbito más general para utilizar todos los tipos de información biológica: su modelado, almacenamiento, recuperación y gestión.

Característica 1: *Los datos biológicos resultan sumamente complejos cuando se comparan con la mayoría de los demás dominios o aplicaciones.* Por ello, las definiciones de datos deben ser capaces de representar una subestructura compleja de datos así como relaciones y de garantizar que no se pierde ninguna información durante el modelado de datos biológicos. La estructura de datos biológicos proporciona con frecuencia un contexto adicional para la interpretación de la información. Los sistemas de información biológica deben ser capaces de representar cualquier nivel de complejidad de cualquier esquema de datos, relación o subestructura de esquema(no sólo datos jerárquicos, binarios o de tabla). A modo de ejemplo, MITOMAP es una base de datos que documenta el genoma mitocondrial humano.[13] Este genoma único es un pequeño fragmento circular de ADN que comprende información sobre 16.569 bases de nucleótidos; 52 loci de gen que codifican el ARN mensajero, el ARN ribosomal, y el ARN de transferencia; 1.000 variantes de poblaciones conocidas; más de 60 asociaciones con enfermedades conocidas; y un conjunto limitado de conocimientos sobre las interacciones moleculares complejas de la energía bioquímica que producen el camino de la fosforilación oxidativa. Como cabría esperarse, su gestión se ha encontrado con un gran número de problemas; no hemos sido capaces de emplear los métodos tradicionales de los SGBDR o de los SGBDO para obtener todos los aspectos de los datos.

Característica 2: *El grado y abanico de variabilidad de los datos es alto.* De ahí, que los sistemas biológicos deban ser flexibles a la hora de manejar los tipos de datos y sus valores. Con un abanico tan amplio de valores de datos posibles, las restricciones a los tipos de datos deben ser limitadas puesto que esto puede excluir valores inesperados (por ejemplo, valores más externos) que son particularmente comunes en el campo biológico. La exclusión de este tipo de valores da como resultado una pérdida de información. Además, las frecuentes excepciones de las estructuras de datos biológicos puede hacer necesario que se disponga de una selección del tipo de datos para un dato concreto.

Característica 3: *Los esquemas de las bases de datos biológicas varían a gran velocidad.* Por ello, para mejorar el flujo de información entre generaciones o versiones de bases de datos se debe soportar la evolución de esquemas y la migración de objetos de datos. La capacidad de ampliar el esquema, algo frecuente en el campo biológico, no cuenta con un soporte en la mayoría de los sistemas de bases de datos relacionales y de objetos. En la actualidad, sistemas como Genbank vuelven a lanzar una o dos veces al año toda la base de datos incorporando nuevos esquemas en lugar de modificar el sistema gradualmente a medida que van siendo necesarios los cambios. Una base de datos evolutiva de este tipo proporcionaría un mecanismo oportuno y ordenado para seguir los cambios de las entidades de datos individuales en las bases de datos biológicas a lo largo del tiempo. Este tipo de seguimiento es importante para que los investigadores biológicos puedan acceder y reproducir los resultados previos.

Característica 4: *Es probable que las representaciones de los mismos datos por parte de biólogos diferentes sean distintas (empleando incluso el mismo sistema).* Por este motivo, deberían soportarse los mecanismos destinados a «alinear» los diferentes esquemas biológicos o versiones diferentes. Dada la complejidad de los datos biológicos, existen múltiples formas de modelar una entidad dada, en las que con frecuencia los resultados reflejan el enfoque concreto del científico. Aunque dos investigadores puedan obtener modelos de datos dispares si se les pide que interpreten la misma entidad, es probable que estos modelos tengan numerosos puntos en común. En este tipo de situaciones, sería útil que los investigadores biológicos pudieran realizar las consultas basándose en estos puntos comunes. Esto se lograría relacionando elementos de datos en una red de esquemas.

Característica 5: *La mayor parte de los usuarios de datos biológicos no precisan tener acceso de escritura a la base de datos; es adecuado el acceso de sólo lectura.* El acceso de escritura está

[13] Los detalles sobre MITOMAP así como sobre la complejidad de su información pueden consultarse en Kogelnik *et al.* (1997, 1998) y en http://www.gen.emory.edu/mitomap.html.

limitado a usuarios privilegiados llamados *administradores* (*curators*). Por ejemplo, la base de datos creada como parte del proyecto MITOMAP tiene un promedio de más de 15.000 usuarios al mes en Internet. Hay menos de veinte propuestas al mes procedentes de usuarios que no son administradores. Es decir, el número de usuarios que requieren acceso de escritura es reducido. Los usuarios generan una amplia variedad de patrones de acceso de lectura en la base de datos, pero estos patrones no son los mismos que los que hemos visto en las bases de datos relacionales tradicionales. Las búsquedas ad hoc solicitadas por el usuario exigen la indexación de combinaciones de clases de instancias de datos que son con frecuencia inesperadas.

Característica 6: *Es probable que la mayoría de los biólogos no tenga ningún conocimiento de la estructura interna de la base de datos o del diseño del esquema.* Las interfaces de bases de datos biológicas deberían mostrar la información a los usuarios de forma que ésta sea aplicable al problema del que se ocupan y que refleje la estructura subyacente de los datos. Por lo general, los usuarios biológicos conocen los datos que necesitan, pero no tienen los conocimientos técnicos de la estructura de los datos o de la forma en la que un SGBD representa los datos. Estos confían en que los usuarios técnicos les proporcionen vistas de la base de datos. Los esquemas relacionales no consiguen ofrecer al usuario indicaciones o información intuitiva respecto al significado de su esquema. En concreto, las interfaces Web ofrecen a menudo interfaces de búsqueda preestablecidos, lo que puede limitar el acceso a la base de datos. Sin embargo, si estas interfaces se generan directamente a partir de las estructuras de la base de datos, es probable que den lugar a un campo de acceso más amplio, aunque puede que no garanticen la facilidad de uso.

Característica 7: *El contexto de los datos aporta un significado añadido para su uso en las aplicaciones biológicas.* Por lo tanto, el contexto debe mantenerse y transmitirse al usuario cuando sea preciso. Además, debería ser posible integrar tantos contextos como sea posible para maximizar la interpretación de un valor de datos biológico. Los valores aislados son de menor utilidad en los sistemas biológicos. Por ejemplo, la secuencia de un eslabón de ADN no es especialmente útil sin la información adicional que describe su organización, función, etc. Por ejemplo, un único nucleótido en un eslabón de ADN, visto en contexto con eslabones de ADN que no causan enfermedades, podría considerarse un elemento causante de la anemia drepanocítica.

Característica 8: *La definición y la representación de consultas complejas son sumamente importantes para un biólogo.* Por lo tanto, los sistemas biológicos deben soportar consultas complejas. Sin tener ningún conocimiento de la estructura de datos (*véase* la Característica 6), un usuario medio no puede construir por sí mismo una consulta compleja donde intervengan varios conjuntos de datos. Por ello, con el fin de resultar realmente útiles, los sistemas deben ofrecer determinadas herramientas para generar estas consultas. Como hemos dicho anteriormente, muchos sistemas proporcionan plantillas de consultas predefinidas.

Característica 9: *Los usuarios de información biológica a menudo necesitan los valores «antiguos» de los datos, especialmente cuando se verifican los resultados obtenidos anteriormente.* Por ello, los cambios realizados a los valores de los datos de la base de datos se tienen que conservar mediante un sistema de ficheros. En el campo biológico son importantes tanto el acceso a la versión más reciente del valor de un dato como a su versión anterior. Los investigadores desean consultar sistemáticamente los datos más recientes, pero también deben poder reconstruir el trabajo anterior y volver a evaluar la información previa y la actual. En consecuencia, los valores que están a punto de actualizarse en una base de datos biológica no pueden simplemente ser simplemente eliminados.

Todas estas características sirven claramente de evidencia de que los SGBD actuales no satisfacen plenamente los requisitos de los datos biológicos complejos. Es necesaria una nueva dirección en los sistemas de gestión de bases de datos.[14]

[14] En Kogelnik (1998) se da como prototipo para hacer frente a estas cuestiones un entorno de base de datos denominado GENOME (*Georgia Tech Emory Network Object Management Environment*); *véase* también Kogelnik *et al.* (1997, 1998).

27.5.3. El proyecto del genoma humano y las bases de datos biológicas actuales

El término *genoma* se define como la información genética total que puede obtenerse sobre una entidad. El **genoma humano**, por ejemplo, hace referencia en general al conjunto completo de genes necesarios para crear un ser humano (que se calcula entre 100.000 y 300.000 genes que se extienden en 23 pares de cromosomas, con un número aproximado de 3 a 4 billones de nucleótidos). El objetivo del Proyecto del Genoma Humano (PGH) es el de obtener la secuencia completa (el orden de las bases) de esos nucleótidos. En la actualidad, sólo se han identificado 8.000 genes y se ha conseguido la secuencia de menos del 10 por ciento del genoma humano. Sin embargo, se espera que para el año 2002 se completará toda la secuencia. De manera aislada, la secuencia del ADN no resulta especialmente útil. Sin embargo, la secuencia puede combinarse con otros datos y emplearse como una herramienta poderosa para ayudar a resolver cuestiones en los campos de la genética, bioquímica, medicina, antropología y la agricultura. En las bases de datos del genoma existentes, el interés se ha centrado en «conservar» (o recabar con un determinado examen inicial y control de calidad) y clasificar la información sobre los datos de la secuencia del genoma. Además del genoma humano, se han investigado numerosos organismos como *E.coli*, *Drosophila*, y *C. Elegans*. Examinaremos brevemente algunos de los sistemas de bases de datos actuales que están sirviendo de soporte o que han surgido del Proyecto del Genoma Humano.

Genbank. La base de datos de la secuencia del ADN preeminente en el mundo hoy en día es Genbank, de la que se ocupa el Centro Nacional de Información sobre Biotecnología (*National Center for Biotechnology Information*, NCBI) de la Biblioteca Nacional de Medicina (National Library of Medicine, NLM). Fue establecida en 1978 como un almacén central de los datos de la secuencia del ADN. Desde entonces, su ámbito se ha ido expandiendo para incluir datos de identificación de la secuencia, datos de la secuencia de proteínas, la estructura de proteínas tridimensionales, la taxonomía, y enlaces con la bibliografía biomédica (MEDLINE). Su última versión contiene más de 602.000.000 de bases de nucleótidos de más de 920.000 secuencias sobre más de 16.000 especies con un número aproximado de diez organismos nuevos que se añaden cada día. El tamaño de la base de datos se ha duplicado aproximadamente cada dieciocho meses durante cinco años.

Aunque se trata de una base de datos compleja y global, el ámbito de su cobertura se centra en las secuencias humanas y en enlaces con la bibliografía. Otras fuentes de datos limitadas (por ejemplo, la estructura tridimensional y OMIM, que examinaremos a continuación) se han añadido recientemente dando un nuevo formato a las bases de datos ya existentes, OMIM y PDB, y volviendo a diseñar la estructura del sistema Genbank para dar cabida a estos conjuntos de datos nuevos.

El sistema se mantiene como una combinación de ficheros planos, bases de datos relacionales, y ficheros que contienen una ***Abstract Syntax Notation One*** **(ASN.1 o Notación de Sintaxis Abstracta Uno)**, que es una sintaxis destinada a definir estructuras de datos desarrollada para la industria de telecomunicaciones. El NCBI asigna a cada entrada de Genbank un identificador único. A las actualizaciones se les asigna un nuevo identificador, mientras que el identificador de la identidad original sigue sin modificarse a efectos de ser archivado. De este modo, las referencias más antiguas de una entidad no indican inadvertidamente un valor nuevo y posiblemente inadecuado. Los conceptos más actuales también reciben un segundo conjunto de identificadores únicos (UID), los cuales establecen la forma más actualizada de un concepto al tiempo que permiten que se acceda a versiones más antiguas por medio de su identificador original.

El usuario medio de la base de datos no puede acceder a la estructura de los datos directamente para realizar consultas u otras funciones, aunque se pueden obtener instantáneas completas de la

base de datos para su exportación en una serie de formatos, incluyendo la ASN.1. El mecanismo de consulta que se proporciona es mediante la aplicación *Entrez* (o su versión en la World Wide Web), que permite búsquedas por una palabra clave, una secuencia y un UID de Genbank por medio de una interfaz estática.

La Base de Datos del Genoma (GDB).[15] Creada en 1989, la Base de Datos del Genoma (GDB) es un catálogo de datos sobre genes humanos. Almacena correspondencias entre una información y una localización concreta en el genoma humano. El grado de precisión de esta localización la correspondencia depende de la fuente de los datos, pero no se da generalmente al nivel de las bases de nucleótidos individuales. Los datos de la GDB incluyen datos que describen principalmente información sobre dicha correspondencia (distancia y límites del intervalo de confianza), y datos de investigación (condiciones experimentales, bases de PCR, y reactivos empleados) de la *Polymerase Chain Reaction* (PCR). Más recientemente, se han realizado esfuerzos para incorporar datos sobre mutaciones asociadas a los loci genéticos, líneas de células empleadas en experimentos, bibliotecas de investigación del ADN, y algunos datos limitados sobre el polimorfismo y poblaciones.

El sistema GDB se creó en torno a SYBASE, un SGBD relacional comercial, y sus datos se configuran empleando las técnicas estándar del modelo Entidad-Relación (*véanse* los Capítulos 3 y 4). Los creadores de la GDB han apreciado la existencia de dificultades a la hora de emplear este modelo para capturar algo más que simples correspondencias y datos de investigación. Con el fin de mejorar la integridad de los datos y de simplificar la programación de aplicaciones, la GDB distribuye un Juego de Herramientas de Acceso a la Base de Datos (*Database Access Toolkit*). Sin embargo, la mayoría de los usuarios emplean una interfaz Web para buscar los diez gestores de datos interrelacionados. Cada gestor hace un seguimiento de los vínculos (relaciones) de una de las diez tablas dentro del sistema GDB. Como sucedía con Genbank, los usuarios reciben únicamente una visión de los datos de nivel muy alto en el momento de la búsqueda por lo que no pueden emplear fácilmente ningún conocimiento extraído de la estructura de las tablas de la GDB. Los métodos de búsqueda resultan mucho más prácticos cuando los usuarios buscan simplemente un índice de la correspondencia o de los datos de investigación. Las interfaces actuales no fomentan la búsqueda preliminar ad hoc de la base de datos. La integración de las estructuras de base de datos de GDB y OMIM (*véase* más abajo) no se estableció nunca plenamente.

Herencia mendeliana en el hombre *on-line*. La herencia mendeliana en el hombre *on-line* (OMIM: *On-line Mendelian Inheritance in Man*) es un compendio electrónico de información sobre la base genética de las enfermedades humanas. Fue iniciada en forma impresa por Victor McCusick en 1966 con 1500 entradas, y convertida en formato electrónico con texto completo entre 1987 y 1989 por la GDB. En 1991 su administración fue transferida de la Johns Hopkins University al NCBI, y toda la base de datos fue convertida al formato Genbank del NCBI. Hoy en día, contiene más de 7.000 entradas.

La OMIM abarca material en cinco áreas de enfermedades basadas en términos generales en órganos y sistemas. Toda propiedad morfológica, bioquímica, de conducta o de otro tipo que se encuentre bajo investigación recibe el nombre de **fenotipo** de un individuo (o de una célula). Mendel observó que los genes pueden existir de numerosas formas distintas conocidas como **alelos**. Un **genotipo** hace referencia a la composición aléica real de un individuo.

La estructura de las entradas de fenotipos y genotipos contiene datos textuales estructurados en términos generales como descripciones generales, nomenclaturas, modos de herencia, variaciones, estructura de genes, correspondencias y numerosas categorías menores. Las entradas con texto

[15] GBD significa *Genome Database*.

completo se transformaron a un formato estructurado ASN.1 cuando la OMIM fue transferida al NCBI. Esto mejoró considerablemente la capacidad de asociar datos OMIM con otras bases de datos y también proporcionó una estructura rigurosa de los datos. Sin embargo, la forma básica de la base de datos continuó siendo difícil de modificar.

EcoCyc. La Enciclopedia de Genes y Metabolismo *Escherichia Coli* (EcoCyc) es un experimento reciente que combina información sobre el genoma y el metabolismo de *E.coli K*-12. La base de datos fue creada en 1966 como una colaboración entre el Instituto de Investigación de Stanford y el Laboratorio Biológico Marino. Ésta clasifica y describe los genes conocidos de *E.coli*, las enzimas codificadas por dichos genes, y las reacciones bioquímicas catalizadas por cada enzima y su organización en caminos metabólicos. De esta manera, EcoCyc extiende los ámbitos de las secuencias y funciones de la información genómica. Contiene 1.283 compuestos con 965 estructuras así como listas de enlaces y átomos, pesos moleculares, y fórmulas empíricas. Contiene 3.038 reacciones bioquímicas descritas mediante 269 clases de datos.

Primeramente, se empleó un modelo de datos orientado a objetos para implementar el sistema, con datos almacenados en Ocelot, un sistema de representación del conocimiento mediante *frames*. Los datos de EcoCyc se dispusieron en una jerarquía de clases de objetos basadas en las observaciones de que (1) las propiedades de una reacción son independientes de la enzima que lo cataliza, y (2) una enzima tiene una serie de propiedades que son «lógicamente distintas» de sus reacciones.

EcoCyc proporciona dos métodos de consulta: (1) directo (por medio de consultas predefinidas) y (2) indirecto (por medio de navegación por hipertexto). Las consultas directas se realizan empleando menús y diálogos que pueden lanzar un conjunto de consultas finito pero muy numeroso. No existe soporte para la navegación a través de las estructuras de datos reales. Además, no se documenta ningún mecanismo para la evolución del esquema.

La Tabla 27.1 resume las características de las principales bases de datos relacionadas con el genoma, así como las bases de datos HGMDB y ACEDB. Existen algunas bases de datos de proteínas adicionales; éstas contienen información sobre las estructuras de las proteínas. Las bases de datos fundamentales de proteínas incluyen SWISS-PROT de la University of Geneva, *Protein Data Bank* (PDB) del Brookhaven National Laboratory, y *Protein Identification Resource* (PIR) de la National Biomedical Research Foundation.

Durante los últimos diez años, se ha producido un creciente interés por las aplicaciones de las bases de datos en la biología y la medicina. Genbank, GDB, y OMIM se han creado como almacenes centrales de determinados tipos de datos biológicos pero, aunque son sumamente útiles, siguen sin abarcar todo el espectro de datos sobre el Proyecto del Genoma Humano. Sin embargo, se están realizando esfuerzos en todo el mundo encaminados al diseño de nuevas herramientas y técnicas que paliarán el problema de gestión de datos para los científicos biológicos e investigadores médicos.

27.5.4. Bibliografía seleccionada para la bases de datos del genoma

La bioinformática se ha convertido en los últimos años en una área de investigación conocida y se están organizando numerosos seminarios y conferencias en torno a este tema. Robbins (1993) ofrece una buena perspectiva general mientras que Frenkel (1991) estudia el proyecto del genoma humano considerando el papel especial que desempeña en la bioinformática en general. Cutticchia *et al.* (1993), Benson *et al.* (1996), y Pearson *et al.* (1994) constituyen referencias sobre GDB, Genbank, y OMIM. Wallace (1995) ha sido pionero en la investigación del genoma mitocondrial,

Tabla 27.1. Resumen de las principales bases de datos relacionadas con el genoma.

Nombre de base de datos	Contenido principal	Tecnología inicial	Tecnología actual	Áreas de problemas en BD	Tipos de datos principales
Genbank	Secuencia ADN/ARN, proteína	Ficheros de texto	Fichero plano/ASN.1	Navegación de esquemas, evolución de esquemas, vinculación con otras bd	De texto, numéricos, algunos tipos complejos
OMIM	Fenotipos y genotipos de enfermedades, etcétera	Tarjetas de índices/ficheros de texto	Fichero plano/ASN.1	Entradas de texto libre no estructurado vinculadas a otras bd	De texto
GDB	Datos de vinculación con el mapa genético	Fichero plano	Relacional	Expansión/evolución de esquemas, objetos complejos, vinculación con otras bd	De texto, numéricos
ACEDB	Datos de vinculación con el mapa genético, datos de secuencias (no humanas)	OO	OO	Expansión/evolución de esquemas, vinculación con otras bd	De texto, numéricos
HGMDB	Secuencias y variantes de secuencia	Fichero plano, específico de la aplicación	Fichero plano, específico de la aplicación	Expansión/evolución de esquemas, vinculación con otras bd	De texto
EcoCyc	Reacciones y caminos bioquímicos	OO	OO	Cerradas en jerarquías de clases, evolución de esquemas	Tipos complejos, de texto, numéricos

que se ocupa de una parte específica del genoma humano; los detalles sobre la secuencia y organización de esta área aparecen en Anderson *et al.* (1981). Los trabajos recientes de Kogelnik *et al.* (1997, 1998) y Kogelnik (1998) abordan el desarrollo de una solución genérica para el problema de la gestión de datos en las ciencias biológicas mediante la creación de una solución prototipo. Se puede acceder a la base de datos MITOMAP elaborada por Kogelnik (1998) en la dirección http://www.gen.emory.edu/mitomap.html. También puede accederse a la mayor base de datos de proteínas SWISS-PROT en http://expasy.hcuge.ch/sprot/. La información sobre la base de datos ACEDB está disponible en http://probe.nalusda.gov:8080/acedocs/.

27.6. Bibliotecas digitales

Las bibliotecas digitales constituyen un área de investigación importante y activa. Desde el punto de vista conceptual, una biblioteca digital es análoga a una biblioteca tradicional (un gran conjunto de fuentes de información en medios diferentes) junto con las ventajas de las tecnologías digitales. Sin embargo, las bibliotecas digitales difieren de las tradicionales de forma significativa: el almacenamiento es digital, el acceso a distancia es rápido y fácil, y los materiales se copian de una versión original. Además, resulta sencillo tener copias adicionales a mano y no se ve obstaculizado por restricciones de presupuesto o de almacenamiento, que constituyen los principales proble-

Tabla 27.2. Bases de datos y bibliotecas digitales: similitudes y diferencias.

Similitudes	
• Ambas recogen, organizan, almacenan, buscan, recuperan, procesan y proporcionan datos. • Ambas contienen y manejan soportes múltiples. • Ambas contienen y manejan tipos de datos heterogéneos y objetos complejos.	
Diferencias	
Bibliotecas digitales	*Bases de datos*
No hay un gestor centralizado (gestión mediante interfaces convenidas)	Gestor centralizado (ABD)
Calidad de datos no controlada/menos controlada	Calidad de datos controlable

mas en las bibliotecas tradicionales. Por lo tanto, las tecnologías digitales reducen muchas de las limitaciones físicas y económicas de las bibliotecas tradicionales.

La introducción del número especial de abril 1995 de *Communications of the ACM* en torno a las bibliotecas digitales las describe grandiosamente como la «oportunidad... de cumplir el viejo sueño de todo ser humano: tener fácil acceso al almacén de información de la humanidad». En el Capítulo 1, definimos una base de datos en términos bastante generales como un «conjunto de datos relacionados». A diferencia de los datos relacionados en una base de datos, una biblioteca digital comprende multitud de fuentes, muchas de las cuales no están relacionadas. Lógicamente, las bases de datos pueden ser componentes de las bibliotecas digitales (*véase* la Tabla 27.2).

La *Digital Library Initiative* (DLI) fundada conjuntamente por NSF, DARPA, y NASA ha sido una impulsora fundamental del desarrollo de las bibliotecas digitales. Esta iniciativa proporcionó una importante financiación a seis de los principales proyectos de seis universidades en su primera fase, abarcando un amplio espectro de tecnologías instrumentales (como veremos más adelante). Las páginas Web de la Iniciativa (*véase* dli.grainger.uiuc.edu/national.htm) definen su objetivo de «potenciar de manera espectacular los medios para obtener, almacenar y organizar la información en formas digitales, y hace que esté disponible para la investigación, recuperación y procesamiento por medio de redes de comunicación, todo de manera que resulte fácil de utilizar para el usuario».

La magnitud de estas colecciones de datos así como su diversidad y multitud de formatos constituyen retos en una nueva escala. Es probable que la progresión futura del desarrollo de las bibliotecas digitales avance de la tecnología actual de recuperación vía Internet, por medio de búsquedas en la Red de información indexada en almacenes, a una etapa de correlación y análisis de la información por parte de redes inteligentes. Las técnicas para la recogida de información, almacenarla y organizarla para servir de soporte a las necesidades informativas aprendidas en las décadas de diseño e implementación de bases de datos, ofrecerán la línea de fondo para el desarrollo de métodos adecuados para las bibliotecas digitales. La búsqueda, la recuperación y el procesamiento de numerosos formatos de información digital harán uso de lo aprendido con las operaciones de bases de datos en esos formatos de información.

27.6.1. La iniciativa de las bibliotecas digitales

La Universidad de Illinois en Urbana-Champaign está coordinando la sincronización nacional de los seis proyectos DLI en curso en seis universidades participantes. Los proyectos muestran los principales aspectos de las necesidades del desarrollo de bibliotecas digitales. Los participantes y sus objetivos son los siguientes:

- *University of California en Berkeley*: *Sistemas de Planificación del Medioambiente e Información Geográfica* (*véase* la Sección 27.4 anterior para más detalles sobre los GIS). Este proyecto pondrá en práctica una biblioteca digital como un conjunto de servicios de información distribuida, haciendo hincapié en la indexación automatizada, la recuperación inteligente, soporte por parte de bases de datos distribuidas, un mejor protocolo de recuperación cliente-servidor, una mejor tecnología de obtención de datos y un nuevo paradigma de interacción.
- *University of California en Santa Bárbara*: *El Proyecto Alejandría*. Información de mapas con referencias espaciales. Este proyecto desarrollará servicios de biblioteca distribuidos para recogida de información gráfica e indexada espacialmente (*véase* la Sección 23.3 en torno al modelado espacial y la 27.4 en torno a los GIS).
- *Carnegie Mellon University*: *Biblioteca de Video Digital Informedia*. Mediante el empleo de bibliotecas de vídeo digitales, este proyecto se centrará en la búsqueda y recuperación de todos los contenidos.
- *University of Illinois en Urbana-Champaign*: *Federación de Almacenes de Literatura Científica* (*Federating Repositories of Scientific Literature*). Este proyecto ampliará una biblioteca digital para miles de usuarios y documentos.
- *University of Michigan*: *Agentes Inteligentes para la Localización de la Información* (*Intelligent Agents for Information Location*). Este proyecto se centrará en la tecnología de agentes, incluyendo el empleo de agentes de interfaz de usuario, agentes de mediación para coordinar búsquedas, y agentes de recogida.
- *Stanford University*: *Mecanismos de Interoperación entre Servicios Heterogéneos* (*Interoperation Mechanisms Among Heterogeneus Services*). Este proyecto desarrollará tecnologías instrumentales para una biblioteca digital integrada proporcionando un acceso uniforme a servicios novedosos mediante colecciones de información interconectada.

27.6.2. Bibliografía seleccionada sobre bibliotecas digitales

El número de abril de 1995 de *Communications of the ACM* está dedicado a las bibliotecas digitales. En él, Wiederhold (1995) examina la importancia de las bibliotecas digitales en la creciente productividad humana a la hora de descubrir nuevos conocimientos. El número de abril 1998 de *Communications of the ACM* está dedicado a ofrecer un ámbito global y acceso ilimitado a las bibliotecas digitales. Schatz (1995, 1997) proporciona un excelente tratamiento de la recuperación, búsqueda y análisis de la información relacionada con la información en Internet. Puede accederse a las diversas universidades integrantes de la Iniciativa de Bibliotecas Digitales en las siguientes direcciones:

Digital Libraries Initiative: dli.grainger.uiuc.edu/national.htm
University of Berkeley: elib.cs.berkeley.edu
University of Santa Barbara: alexandria.sdc.ucsb.edu
Carnegie-Mellon University: www.informedia.cs.cmu.edu
University of Illinois Urbana-Champaign: dli.grainger.uiuc.edu/default.htm
University of Michigan: www.si.umich.edu/UMDL
Stanford University: walrus.stanford.edu/diglib