

Granularidad automatizada para integrar información digital: el estudio de caso de la "Antarctic Treaty Searchable Database"

Paul Arthur Berkman^{1,2*}, George James Morgan III^{2,3}, Reagan Moore⁴ y Babak Hamidzadeh⁵

*1 Bren School of Environmental Science & Management, University of California, Santa Barbara, CA 93106, United States, Email: berkman@bren.ucsb.edu

*2 EvREsearch LTD, 1611 Tennyson Court, Columbus, OH 43235, United States, Email: paul@evresearch.com

3 Native Voices International, 4639 Cleveland Road, Wooster, OH 44691, United States, Email: sysop@nvi.net

4 San Diego Supercomputer Center, University of California, San Diego, CA 92093, United States, Email: moore@sdsc.edu

5 Office of Strategic Initiatives, Library of Congress, Washington, DC 20540, United States, Email: babak@loc.gov

Fuente: Traducido y publicado con autorización de CODATA, Data Science Journal. Publicado originalmente en: CODATA, Data Science Journal, Vol. 5 (2006), pp. 84-99. ISSN 1683-1470. URL: http://www.jstage.jst.go.jp/article/dsj/5/0/84/_pdf

Traducción: Alejandro Delgado Gómez

Resumen

El acceso a la información es necesario, pero no suficiente en nuestra era digital. El reto es integrar de manera objetiva los recursos digitales basándose en objetivos definidos por el usuario al efecto de descubrir relaciones de información que faciliten las interpretaciones y la toma de decisiones. La *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>), que se encuentra en su sexta edición, proporciona un ejemplo de integración digital basada en la generación automatizada de gránulos de información que pueden combinarse de manera dinámica para revelar relaciones objetivas dentro y entre recursos de información digital. Este estudio de caso demuestra además que la granularidad automatizada y la integración dinámica pueden lograrse simplemente utilizando la estructura inherente de los recursos de información digital. Tal integración de la información es relevante para los programas bibliotecarios y archivísticos que requieren la conservación a largo plazo de recursos digitales auténticos.

Palabras clave: integración, dinámico, documentos de archivo, digital, archivo, biblioteca

1 Introducción

1.1 La era de la información digital

Hemos llegado a un umbral en nuestra "sociedad de la información mundial" en que acceder a más información no equivale a generar más conocimiento. El conocimiento, que emerge de la comprensión de las relaciones dentro y entre recursos de información, se deriva del proceso de integración. Las distinciones entre acceso a la información e integración subyacen a las soluciones tecnológicas del futuro, en el que "*el conocimiento es la riqueza común de la humanidad*", como expresó Su Excelencia Adama Samassekou en el congreso de CODATA en el 2004 en Berlín. El propósito de este ensayo es ponderar los retos, estrategias y eficiencias para integrar recursos de información digital.

Ponderar los retos del soporte digital es instructivo para adoptar una visión amplia de las comunicaciones escritas en nuestra civilización. Desde la piedra y la arcilla hasta el papel sobre soportes digitales, cada era ha incrementado nuestra capacidad para transportar, producir e integrar información (Fig. 1). Por ejemplo, Internet ha estado evolucionando desde finales de los años sesenta (Berners-Lee et al. 2001, Pastor-Satorras y Vespignani 2004) con un número creciente de hosts de Internet, de 213 en 1981 a más de 350.000.000 en 2005 (Internet Systems Consortium 2005). Desde 1972, la velocidad de los microprocesadores ha incrementado 5 veces su magnitud (Intel Corporation 2005), mientras que los sistemas de satélite han hecho posible recoger y transmitir información a escala global (Evans 2000). Más aún, el volumen de información digital se dobló en los tres años posteriores a 1999 con más de 5 hexabytes (10¹⁸ bytes) de información almacenada sólo en soporte impreso, óptico y magnético (Lyman et al. 2003). También tenemos poderosos motores de búsqueda para recuperar información digital de vastos depósitos. Todas estas características apuntan hacia la observación de que el acceso a la información digital se ha vuelto efectivamente infinito e instantáneo.

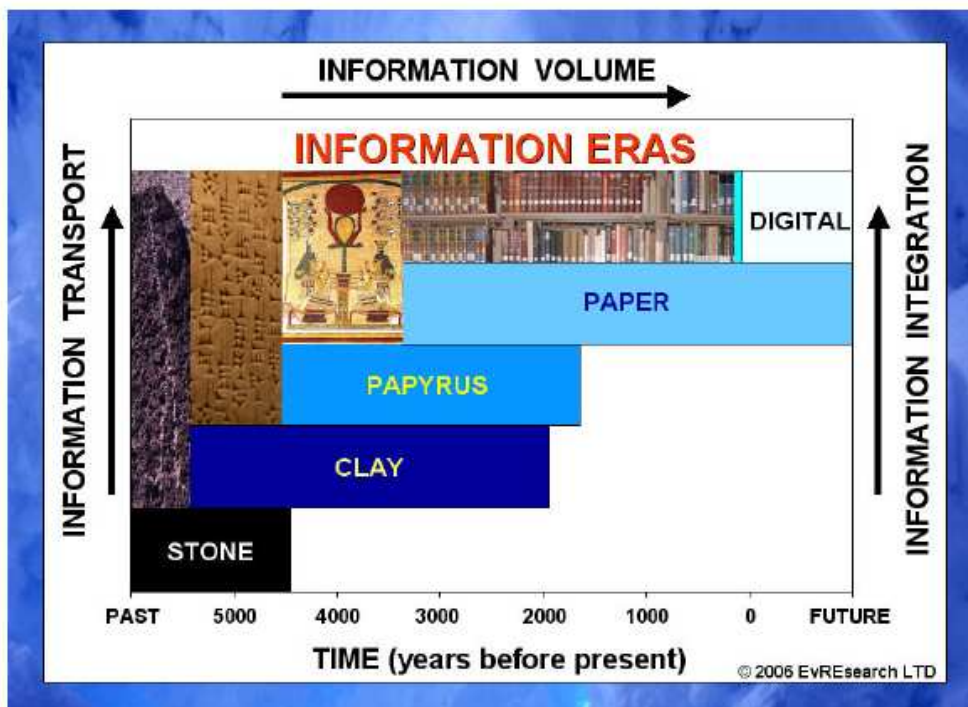


Figura 1: Umbrales en la conservación y diseminación de información escrita en nuestra civilización. Cada uno de los soportes anteriores al digital han sido utilizados durante milenios (Senner 1989). Desde la piedra a los soportes digitales: (a) el transporte de información en el tiempo y el espacio se ha incrementado; (b) el volumen y tasa de información producida se ha incrementado; y (c) la capacidad para integrar información en nuevo conocimiento se ha incrementado.

Aunque es fácil comprender que las capacidades para transportar y producir información se han incrementado con cada era (Fig. 1), es menos obvio que la capacidad para integrar información también se haya incrementado. Hoy en día, se considera que más del 80% de la información digital es "desestructurada", lo que significa que no puede descomponerse automáticamente en un schema relacional. En consecuencia, la integración de información se limita de manera efectiva al restante 20% de recursos digitales que están estructurados en bases de datos, metadatos y marcado.

Uno de los principales retos del soporte digital es ser capaz de integrar información, con independencia de que sea "estructurada" o "desestructurada" (Blumberg y Atre 2003).

1.2 Granularidad automatizada

La información tiene tres elementos indivisibles –contenido, contexto y estructura- que unidos proporcionan significado (Fig. 2). Por ejemplo, cuando se encripta un mensaje (esto es, se altera la estructura), todavía tiene contenido y contexto, pero no significado. De manera similar, si se

quitan los nombres o fechas y lugares de un recurso de información, todavía tiene contexto y estructura, pero un significado limitado, sin sus hechos sobresalientes. Quitar las características del contexto que pueden utilizarse para autenticar un recurso de información también comprometerá su significado.

El desplazamiento de paradigma creado por las tecnologías digitales es la oportunidad para utilizar la estructura de la información, así como su contenido y contexto. Un libro impreso puede gestionarse basándose en su contenido (como en las bibliotecas) o en su contexto (como en los archivos), pero no es posible dividir un libro en unidades más pequeñas que puedan gestionarse automáticamente. Es esta capacidad para manipular automáticamente la granularidad de los recursos de información lo que distingue los soportes digitales de cualquiera de sus predecesores en copia dura que han sido aplicados a todo lo largo de la civilización humana.

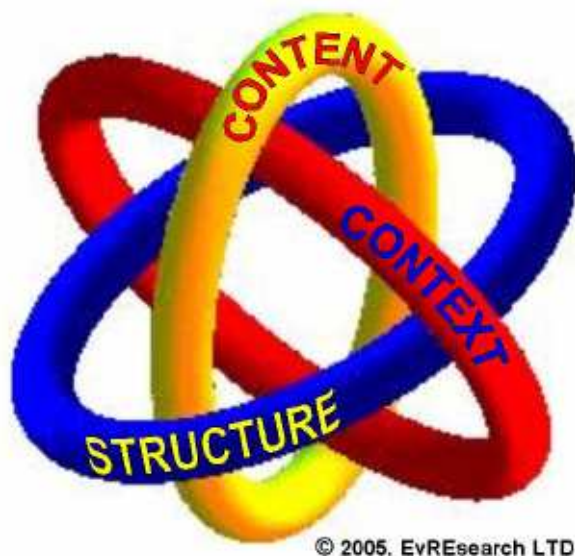


Figura 2: Los anillos de Borromeo que ilustran los tres elementos indivisibles de la información que unidos proporcionan significado.

Este concepto de granularidad se refiere a las unidades conceptuales inherentes (esto es, gránulos de información) que componen un recurso de información. Con el texto, los gránulos podían ser tan pequeños como letras o caracteres individuales, cada uno de los cuales podía identificarse dentro de una ontología de bytes de imprenta, con relación a su recurso padre (Berkman y Morgan 2003). De manera más razonable, los gránulos serán lo suficientemente grandes como para permanecer independientes, con suficiente contenido y contexto, como párrafos o capítulos. Una característica crítica de cada gránulo es que internamente retiene información acerca de su posición jerárquica única dentro de su recurso padre.

La granularidad automatizada ha sido considerada anteriormente para implantar sistemas en red de ordenadores anidados con la "visión de un mundo lleno de grandes números de elementos de computación, muchos de los cuales están ocultos dentro de otros objetos y unidos en red" (National Research Council 2001).

La granularidad automatizada se extiende de manera similar a los documentos digitales de archivo, que tienen anidado su contenido (Berkman y Morgan 2003). El valor de la granularidad automatizada para bibliotecas, archivos y depósitos digitales es que proporciona una estrategia dinámica para buscar, recuperar, organizar e integrar recursos de información tanto "estructurados" como "desestructurados". Este ensayo utiliza la *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>, anteriormente <http://webhost.nvi.net/aspire>) para ponderar las aplicaciones e implicaciones de la granularidad automatizada.

2 Implantación de la *Antarctic Treaty Searchable Database*

2.1 Tecnología de implantación

La tecnología subyacente para implantar la *Antarctic Treaty Searchable Database* es el *Digital Integration System* (DigIn®). El funcionamiento de DigIn®, que se basa en tecnologías patentadas asignadas a EvREsearch LTD (Maynard 2001, 2002, 2003, 2004, 2006), implica cuatro módulos principales que pueden utilizarse juntos o separadamente:

- **MÓDULO DE GRANULARIDAD:** crea gránulos de información utilizando la estructura y los patrones inherentes que vinculan a unidades relevantes de contenido. A cada gránulo se le asigna una etiqueta única de categoría, basándose en un análisis de su procedencia, localización padre-hijo y contenido. La etiqueta de categoría contiene información para generar jerarquías expansibles-colapsables.
- **MÓDULO DE ÍNDICE:** genera una base de datos con la dirección (a la que se hace referencia dentro de cada etiqueta de categoría), cadenas de contenido (palabras, números u otros símbolos) y sus frecuencias dentro de cada gránulo de información.
- **MÓDULO DE INTEGRACIÓN:** busca mediante el índice para recuperar los gránulos de información con términos o cadenas de contenido que se ajusten a las peticiones de búsqueda definidas por el usuario, en forma textual, numérica, o en otras forma simbólicas.
- **MÓDULO DE AGREGACIÓN:** combina gránulos de información relevantes basándose en sus relaciones jerárquicas y en criterios definidos por el usuario.

Cada uno de los módulos actúa a partir de una serie de reglas expertas que definen su funcionamiento automatizado. Estas reglas, que pueden escribirse de manera conveniente con expresiones regulares (Friedl 2002), se optimizan de manera iterativa para integrar y desplegar los gránulos de información relevantes dentro de jerarquías expansibles-colapsables. Además, puesto que DigIn® es modular, puede interactuar con interfaces estadísticos, gráficos, web semántica, lenguaje natural u otros tipos de soluciones de software, que podrían tratarse como módulos adicionales.

DigIn® proporciona un método general que funciona de manera independiente de cualquier hardware o software específico. DigIn® funciona con ASCII o UNICODE, así como con schemas propietarios. DigIn® también funciona con metadatos, marcado y bases de datos que tengan patrones estructurados para organizar la información de manera estructurada (p.ej., Sowa 1984). Más aún, DigIn® está actualmente escrito en PERL, que proporciona un lenguaje de programación estable entre plataformas que puede leer y escribir ficheros binarios, así como procesar ficheros muy grandes. DigIn® también podría escribirse en otras lenguas, dependiendo de las circunstancias. En consecuencia, DigIn® es un método interoperable que puede utilizarse en el futuro de manera persistente.

2.2 Diseño de la implantación

Las actividades generales para crear el documento digital de la *Antarctic Treaty Searchable Database*, así como bases de datos similares de documentos de política, se ilustran en la Figura 3. El primer paso es definir los parámetros de la colección, que incluye los componentes de las colecciones, así como la granularidad resultante y la organización de las visualizaciones jerárquicas que se generarán de manera dinámica en respuesta a las peticiones de integración. Después de compilar los elementos de la colección, el siguiente paso es implantar la granularidad adecuada con una etiqueta de cabecera en cada gránulo que describa su posición jerárquica única en relación con su recurso padre.

Estas etiquetas, que conservan la procedencia de cada gránulo, se utilizarán para generar dinámicamente jerarquías expansibles-colapsables que desplieguen comprensiva y objetivamente las relaciones de los gránulos dentro y entre recursos de información. Después de buscar e integrar los gránulos, las visualizaciones del gránulo se ponderan para determinar si la colección debiera revisarse o si el documento digital completado debiera fijarse a efectos archivísticos.

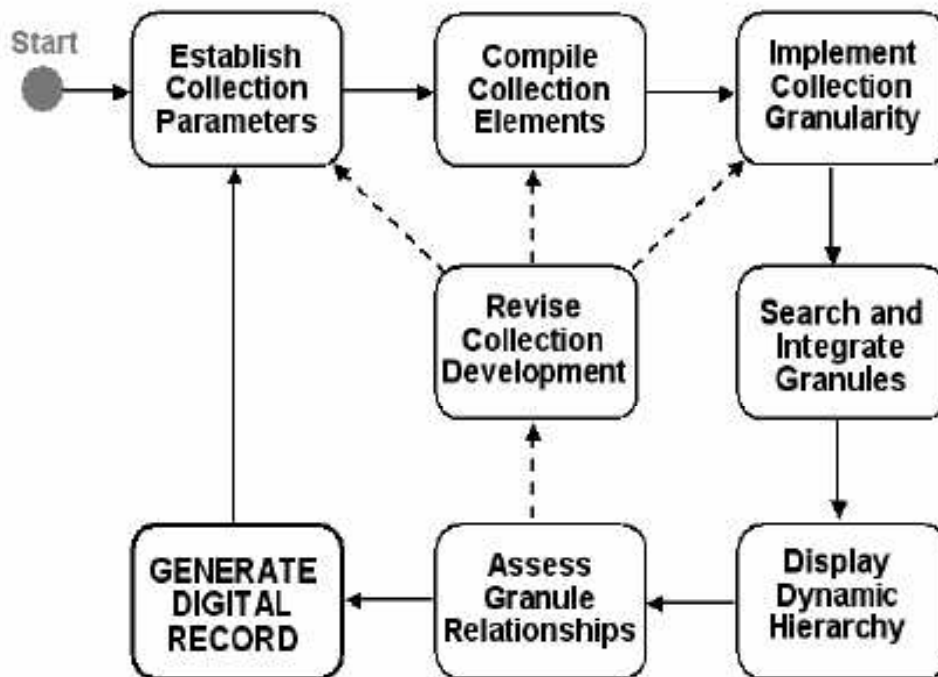


Figura 3: Un diagrama generalizado de actividad-flujo (Bobak 1997) de los procesos para crear la *Antarctic Treaty Searchable Database* u otros documentos digitales con el *Digital Integration System™* (DigIn®), de EvREsearch LTD. Adaptado de Berkman et al. (2005).

De manera más específica, la edición inicial de la *Antarctic Treaty Searchable Database* se implantó en colaboración con la National Science Foundation y el Departamento de Estado de los Estados Unidos. Basándose en las características del *Antarctic Treaty Handbook. 8th Edition* (United States Department of State 1994), se utilizaron las siguientes reglas para compilar los contenidos de la *Antarctic Treaty Searchable Database* inicial:

Regla 1: Incluir sólo las “medidas” que fueron adoptadas por las Partes Consultivas del Tratado Antártico “en apoyo de los principios y objetivos del Tratado”.

Regla 2: El contenido de cada “medida” adoptada incluiría su texto junto con cualquier tabla o figura.

Regla 3: Excluir cualquier “extracto”, “nota introductoria” u otras adiciones del Departamento de Estado de los Estados Unidos, que es el gobierno depositario, porque no habían sido formalmente adoptadas por las Partes Consultivas del Tratado Antártico.

La siguiente decisión fue identificar la granularidad adecuada de los documentos de política que serían buscables. Cada Reunión Consultiva del Tratado Antártico (ATCM) producía un informe

con "recomendaciones", "decisiones", "medidas" o "resoluciones" adoptadas, que a veces incluían "apéndices", "anexos" o "adjuntos". Periódicamente, las Partes Consultivas del Tratado Antártico también adoptaban Convenciones o documentos más amplios de política que incluían "artículos" específicos junto con "anexos". Basadas en estos tipos de medidas adoptadas, las siguientes reglas definen la granularidad de los documentos de política de la *Antarctic Treaty Searchable Database*:

Regla 4: Cada "recomendación", "decisión", "medida" o "resolución" se trataría como un gránulo completo de información (dentro del contexto de la ATCM y el año de adopción con los dos niveles jerárquicos cobertores).

Regla 5: Cada "apéndice", "anexo" o "adjunto" se trataría como un gránulo completo de información (dentro del contexto de la "recomendación", "decisión", "medida" o "resolución", así como dentro de la ATCM y el año de adopción, como los tres niveles jerárquicos cobertores).

Regla 6: Cada "artículo" y "anexo" se trataría como un gránulo completo de información (dentro de una Convención o Protocolo y el año de adopción como los dos niveles jerárquicos cobertores).

La edición inicial de la *Antarctic Treaty Searchable Database*, que se construyó de manera automatizada basándose en las reglas de más arriba, ha sido continuamente actualizada a medida que:

- (1) las Partes Consultivas del Tratado Antártico han adoptado nuevas medidas;
- (2) se han identificado medidas perdidas; y
- (3) se han identificado componentes perdidos de las medidas (p. ej., tablas o figuras).

Estas actualizaciones implican la inserción, etiquetado y edición de gránulos individuales. Cada actualización o edición de la *Antarctic Treaty Searchable Database* ha sido fijada conservando todos los ficheros y su funcionalidad en un webCDserver™ (Berkman 2002). Desde el principio hasta el fin, los contenidos de la *Antarctic Treaty Searchable Database* se han incorporado directamente a partir de fuentes auténticas (esto es, el Departamento de Estado de los Estados Unidos, la Marine Mammal Commission, el Committee for Environmental Protection, y las naciones anfitrionas de la ATCM). La implantación general de la *Antarctic Treaty Searchable Database* se ilustra en la Figura 4.

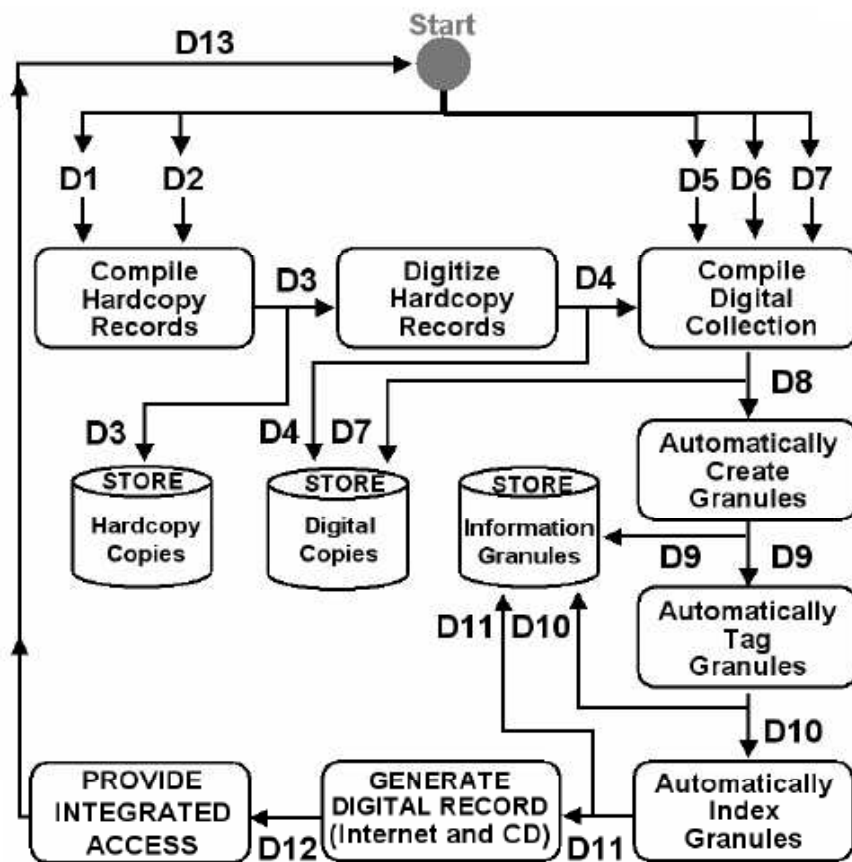


Figura 4: Un diagrama de flujo de datos (Bobak 1997) para ilustrar las actividades junto con elementos de datos específicos y almacenes de datos para implantar la *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>). Los elementos de datos son: **D1** Marine Mammal Commission Compendia (Marine Mammal Commission 1994); **D2** informes de la Antarctic Treaty Consultative Meeting (ATCM); **D3** Documentos relevantes en copia dura; **D4** Documentos digitalizados; **D5** Ficheros digitales del Departamento de Estado de los Estados Unidos (1994); **D6** Sitios web de la ATCM alojada por las diferentes Partes Consultivas del Tratado Antártico; **D7** Ficheros digitales del Committee on Environmental Protection (<http://www.cep.aq>); **D8** Documentos archivísticos digitales de documentos completos; **D9** Gránulos de información; **D10** Gránulos etiquetados de información; **D11** Gránulos indizados-etiquetados de información; **D12** Base de datos completa de gránulos; y **D13** Nuevo sitio web y documento (anual) de webCDserver™. Adaptado de Berkman et al. (2005).

2.3 Historia de la implantación

La historia de la *Antarctic Treaty Searchable Database* se remonta a 1998, cuando se contactó con el Departamento de Estado de los Estados Unidos respecto al acceso a las versiones digitales de los documentos de política que estaba gestionando como gobierno depositario para el *Tratado Antártico de 1959*. Esta petición vino inducida porque la gestión de la información se estaba desplazando rápidamente hacia los soportes digitales y el *Antarctic Treaty Handbook* (United States Department of State 1994) se había vuelto difícil para las actividades de estudios de caso

en el curso de licenciatura sobre ciencia y política antárticas que se estaba impartiendo desde 1982 (Berkman 2002). En 1999, pasado un mes de la implantación inicial de la *Antarctic Treaty Searchable Database*, el Departamento de Estado la presentó en 23ª ATCM en Lima, Perú.

Aunque originalmente se pretendía que fuera un complemento del curso universitario sobre ciencia y política antárticas (Berkman 2002), la *Antarctic Treaty Searchable Database* pronto evolucionó hacia un archivo digital que ha sido mantenido y actualizado posteriormente para beneficio de la diversa comunidad de interesados en la Antártida (Tabla 1). El propósito redefinido de la *Antarctic Treaty Searchable Database* ha sido facilitar el descubrimiento de conocimiento acerca de las políticas y estrategias que promueven la "cooperación internacional" y el "uso de la Antártida sólo para usos pacíficos", como se establecía en el *Preámbulo del Tratado Antártico de 1959*.

Tabla 1. Enlaces a sitios web de representantes de la *Antarctic Treaty Searchable Database*
(<http://aspire.nvi.net>, anteriormente <http://webhost.nvi.net/aspire>)

| <u>Sede del sitio web</u> | <u>URL del sitio web</u> |
|--|---|
| Instituciones gubernamentales internacionales | |
| Antarctic Treaty Secretariat | http://www.ats.aq/ |
| Agencias gubernamentales nacionales | |
| Australian Antarctic Division | http://www.aad.gov.au/default.asp?casid=3638 |
| Canadian Department of Foreign Affairs | http://www.dfait-maeci.gc.ca/circumpolar/sec05_antarctic-en.asp |
| Library of Congress | http://www.loc.gov/rr/international/frd/antarctica/government_law.htm |
| Organizaciones no gubernamentales | |
| Antarctic Southern Ocean Coalition | http://www.asoc.org/links.htm |
| Arctic Council | http://www.arctic-council.org/en/main/infopage/81/ |
| Joint Committee on Antarctic Data Management | http://www.jcadm.scar.org/links1.html#AT |
| Scientific Committee on Antarctic Research | http://www.scar.org/information/links/ |
| The National Academies | http://dels.nas.edu/prb/links.shtml |
| Empresa | |
| American Society of International Law | http://users.erols.com/jackbobo/ |
| Expedition Medicine | http://www.expeditionmedicine.co.uk/ |
| French National Sea Experience Centre | http://www.nausicaa.fr/links/ |
| International Assoc. Antarctic Tour Operators | http://www.iaato.org/resources.html |

| | |
|---|---|
| M/S 'Nordnorge' | http://www.granfoss.net/arne/k100e/restipsr/hurtigru/k1nnca/k1nnkali.htm |
| Programas educativos | |
| George Washington University Law School | http://www.law.gwu.edu/burns/research/intl/env.htm |
| Katholieke Universiteit Leuven | http://www.kuleuven.ac.be/iir/linkse.htm |
| Link Up Alaska | http://www.linkupalaska.com/science/polar/ |
| McMurdo Long-Term Ecological Research | http://huey.colorado.edu/LTER/links.html |
| Oxford University | http://www.oup.uk/pdf/bt/cassese/cases/part1/ch03/614.pdf |
| Students on Ice | http://www.studentsonice.com/antarctica2004/html/antarctica.html |
| Texas A&M University | http://antarctica.tamu.edu/links/index_html |
| University of California, Santa Barbara | http://fiesta.bren.ucsb.edu/~gsd/links/links.php?nav=nonprofit |

Dos años después de presentar la primera edición de la *Antarctic Treaty Searchable Database* (Tabla 2), las Partes Consultivas del Tratado Antártico cambiaron de manera fundamental el “*intercambio de información*” en el Sistema del Tratado Antártico, adoptando la Decisión XXIV-1 en la 24ª ATCM para establecer el Secretariado del Tratado Antártico en Buenos Aires. Mientras estas negociaciones relativas al Secretariado del Tratado Antártico estaban en curso, la *Antarctic Treaty Searchable Database* fue enlazada a los sitios web de las 24ª y 25ª ATCM, en San Petersburgo y Varsovia, respectivamente. Además de ser la primera colección digital de documentos del Tratado Antártico producida, la *Antarctic Treaty Searchable Database* sigue siendo la fuente global más comprehensiva para integrar documentos de política del Sistema del Tratado Antártico.

Tabla 2: Granularidad, cobertura y dimensiones de la *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>, anteriormente <http://webhost.nvi.net/aspire>) a lo largo del tiempo¹

| Año de producción | Edición | Cobertura | Gránulos | | Imágenes anidadas | |
|-------------------|----------------|-----------|----------|-----------------------|-------------------|------------------------|
| | | | Número | Volumen de texto (MB) | Número | Volumen de imagen (MB) |
| 1999 | 1 ^a | 1959-1999 | 608 | 2,65 | 113 | 2,19 |
| 2001 | 2 ^a | 1959-1999 | 608 | 2,65 | 164 | 4,46 |
| 2002 | 3 ^a | 1959-2002 | 661 | 3,11 | 166 | 5,07 |
| 2003 | 4 ^a | 1959-2003 | 720 | 3,67 | 200 | 6,67 |
| 2004 | 5 ^a | 1959-2004 | 740 | 5,60 | 224 | 9,57 |
| 2005 | 6 ^a | 1959-2005 | 822 | 7,63 | 352 | 21,40 |

¹ Copias de cada edición de la *Antarctic Treaty Searchable Database* han sido archivadas en webCDservers™ completamente funcionales e independientes. Adaptado de Berkman et al. (2005).

3 Gestión y descubrimiento del conocimiento

3.1 Limitaciones de la granularidad convencional

El potencial de descubrir relaciones significativas dentro y entre recursos de información es directamente proporcional a su granularidad. Por ejemplo, para una petición de búsqueda dada, dos libros podrían generar 4 posibles resultados (esto es, un libro o el otro, ambos libros, ninguno de los libros). Si cada libro estuviera dividido en dos gránulos, habría 16 posibles combinaciones con 0, 1, 2, 3 ó 4 gránulos. Si cada libro estuviera dividido en cuatro gránulos (esto es, 8 en total), habría 256 posibles combinaciones desde 0 a 8 gránulos. En consecuencia, entre N gránulos hay 2^N combinaciones posibles. Ser capaz de expresar y después descomponer las relaciones ternarias, cuaternarias y de un orden mayor, puede revelar dependencias funcionales entre los gránulos o las entidades digitales (Jones y Song, 1996).

En la práctica, el número de posibles relaciones entre 100 objetos digitales (esto es, 2100) es demasiado grande para gestionarlo de manera comprensiva al comienzo. No obstante, las estrategias convencionales implican descripciones de relaciones al comienzo con lenguajes de marcado (Gill y Ratnakar 2001, Fensel et al. 2003) que añaden estructura a los recursos de información con etiquetas para delimitar, contener o definir los límites de cierto contenido. Por ejemplo, esta limitación desde el comienzo se aplica a las ontologías (McGuinness y Harmelen 2004, Lagoze et al. 2005) que describen relaciones entre componentes, propiedades, funciones y procesos de recursos digitales, así como taxonomías (Szykman et al. 1999, Daconta 2005). Además de estas limitaciones, existe también la característica práctica de que añadir etiquetas de marcado en el recurso de información digital es una forma de contaminación que puede comprometer su contenido auténtico en el futuro. De manera muy importante, definir las relaciones al comienzo al efecto de acceder a la información da como resultado que el final

queda efectivamente restringido, lo que reduce grandemente la oportunidad de ser sorprendido. Dadas estas limitaciones y la sugerencia de que las relaciones no pueden gestionarse de manera comprehensiva al comienzo, ¿qué estrategias pueden revelar las relaciones 2^N entre gránulos al final?

Además de por las aplicaciones de marcado, que se consideran metadatos estructurales, el descubrimiento de conocimiento también viene facilitado por los metadatos descriptivos y administrativos (Hodge 2001). Con respecto a los metadatos descriptivos, existe un universo en expansión de schemas para diferentes disciplinas, instituciones y actividades (p. ej., <http://www.mapageweb.umontreal.ca/turner/meta/english/>), cada uno de los cuales contiene diferentes series de atributos (p. ej., nombre, tamaño, tipo de datos, restricciones de uso, etc.), que deben definirse o documentarse con nomenclaturas especializadas para cada objeto digital (Duval et al. 2001). De manera más importante, los metadatos no escalan, lo cual es una de las principales razones que subyacen a la noción ampliamente mantenida de que existe información “estructurada” y “desestructurada”. Sin embargo, al reconocer que toda la información tiene estructura (Fig. 2), en realidad la información digital es “gestionada” o “no gestionada” con tecnologías convencionales.

Puede construirse un simple experimento para ilustrar las limitaciones de escalabilidad de los metadatos (Fig. 5). Considérese un libro que tiene un volumen de 20 (en unidades arbitrarias) y que cada schema de metadatos completado tiene un volumen de 1 (en unidades arbitrarias). Si el libro está dividido en dos gránulos, cada uno de los cuales debe tener su propio schema de metadatos, entonces el volumen total del libro sigue siendo constante, aunque el volumen total del schema de metadatos se dobla. Si la granularidad se dobla continuamente, el volumen de metadatos pronto sobrepasará el volumen de los datos reales que se están gestionando. Los metadatos adicionales también requieren un esfuerzo incrementado de generación, almacenamiento y proceso –lo que se traduce en costes e ineficacias. Más aún, si los metadatos se almacenan en depósitos separados, entonces la pérdida de los metadatos podría comprometer la conservación de los datos reales.

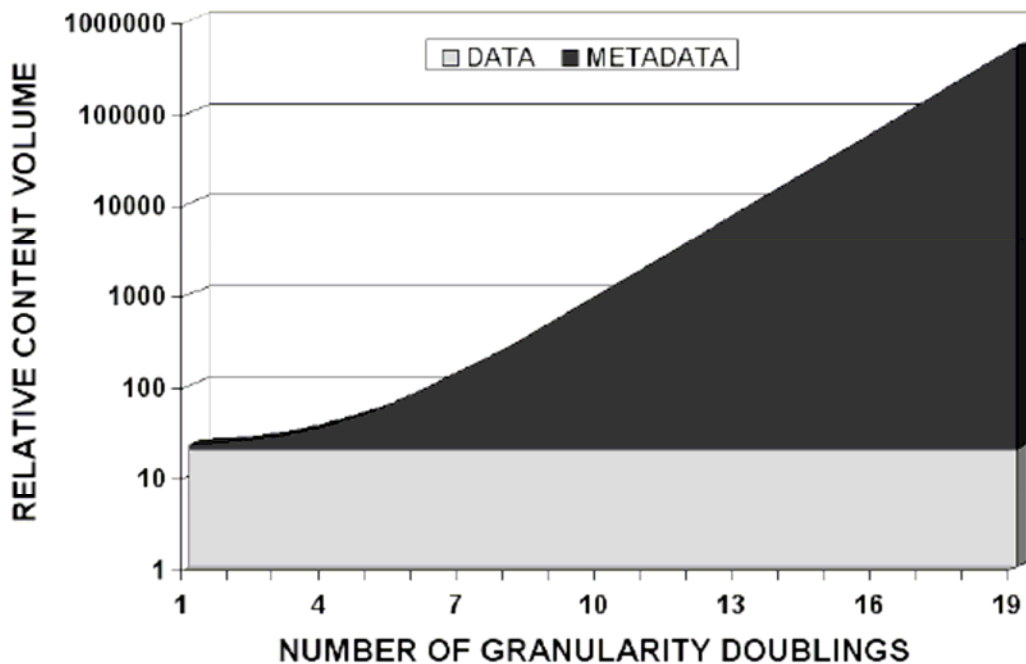


Figura 5: Un modelo simple para ilustrar el volumen exponencialmente creciente de metadatos, a medida que se dobla la granularidad, en relación con los datos reales dentro de un solo recurso digital. En este modelo, el volumen total de todos los gránulos de información es constante (de manera arbitraria, 20 unidades) y el volumen total del schema de metadatos para cada gránulo es fijo (de manera arbitraria, 1 unidad), con independencia del tamaño del gránulo. Adaptado de Berkman y Morgan (2003).

La *Antarctic Treaty Searchable Database* es un ejemplo de recursos de información que están siendo gestionados con una granularidad creciente, pero sin metadatos ni marcado convencionales. Más aún, la *Antarctic Treaty Searchable Database* integra gránulos para lograr 2^N relaciones posibles de información sin las convencionales manipulaciones de tablas en “bases de datos”.

La capacidad para descubrir relaciones con la *Antarctic Treaty Searchable Database* queda adicionalmente reflejada por sus 822 gránulos de información, en contraste con el sitio web del Departamento de Estado de los Estados Unidos (<http://www.state.gov/g/oes/rls/rpts/ant/>) que incluye el [Handbook of the Antarctic Treaty System](#) en 18 ficheros PDF “bloqueados”, junto con ficheros HTML de los cinco principales documentos (p. ej., el *1991 Protocol on Environmental Protection to the Antarctic Treaty*). Con estos sitios web, se requiere que cada usuario lleve a cabo búsquedas a texto completo, un recurso digital a la vez, antes de que el usuario sea capaz de cortar-y-pegar, y luego organizar las piezas relevantes de información –pasos que se han automatizado con la *Antarctic Treaty Searchable Database* y otras aplicaciones DigIn® (p. ej., *Marine Mammal Commission Digital Library of International Environmental and Ecosystem Policy Documents* – <http://nsdl.tierit.com>).

3.2 Una aplicación de integración dinámica

La capacidad para integrar y generar un schema relacional objetivo basado en la estructura inherente de los recursos de información puede ilustrarse con la *Antarctic Treaty Searchable Database*. De los 822 gránulos de la 6ª edición de la *Antarctic Treaty Searchable Database* (Tabla 2), por ejemplo, 23 gránulos contienen el término "pacífico" (Fig. 6).

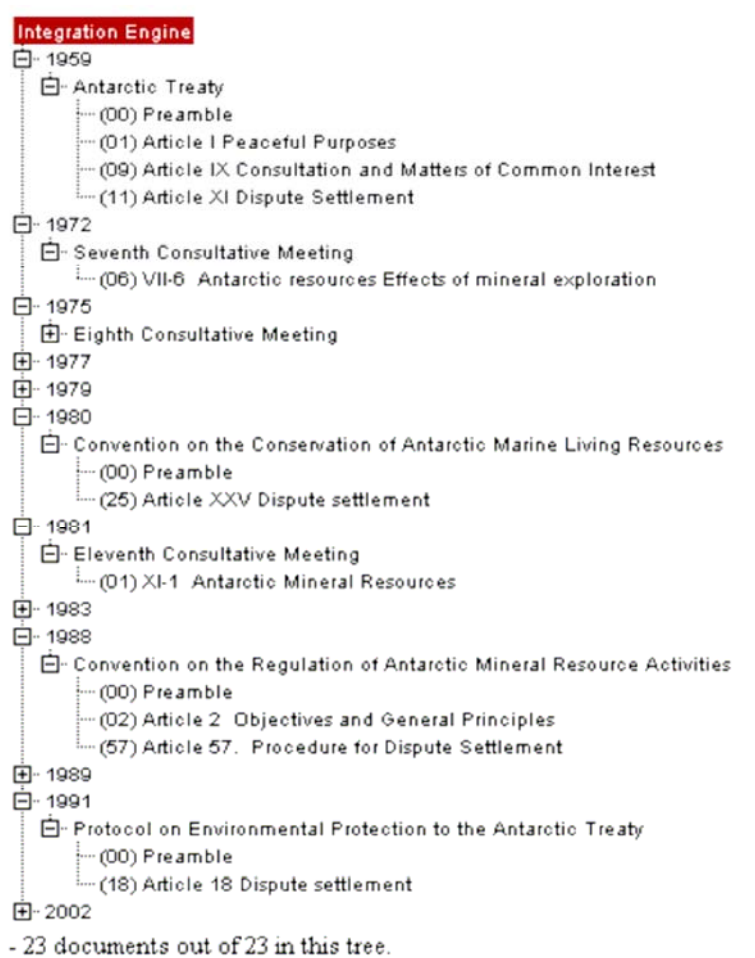


Figura 6: Jerarquía expansible-collapsable que se generó dinámicamente a partir de la 6ª edición de la *Antarctic Treaty Searchable Database* (Tabla 2) con "pacífico" como petición de integración. Las relaciones de política de objetivos dentro y entre años se deriva de los gránulos de información que se generaron sobre la base de las Reglas 1-6 (véase texto).

Los gránulos, que se visualizan en la jerarquía expansible-collapsable, identifican relaciones de política dentro y entre las reuniones del Tratado Antártico que se celebraron de 1959 a 2005. Como puede verse, "pacífico" es una característica común de los "acuerdos en disputa" en las instituciones legales que emergieron del Tratado Antártico en 1980, 1988 y 1991. Más aún, bajo una inspección más estricta de los gránulos individuales, la misma frase se reprodujo cada año

(esto es, "...disputa resuelta por negociación, pesquisa, mediación, conciliación, arbitraje, acuerdo judicial u otros medios **pacíficos**..." Estos resultados son objetivos porque todos los gránulos relevantes (esto es, aquellos con "pacífico") están identificados, y cada gránulo único sólo aparece una vez en la jerarquía.

Las relaciones que pueden visualizarse de manera objetiva también pueden cuantificarse con exactitud para comprobar hipótesis, tal y como los conceptos clave de política se han integrado de manera creciente en las "medidas" adoptadas a lo largo del tiempo. Como ilustración, considérese la protección ambiental antártica, que implica los impactos humanos que se valoran como siendo "*menores o transitorios*" en relación con los diversos valores del Sistema del Tratado Antártico. Basándose en los datos extraídos de las visualizaciones jerárquicas (p ej., Fig. 6), pueden identificarse las tendencias en la incorporación de conceptos ambientales clave a las medidas del Tratado Antártico (Fig. 7).

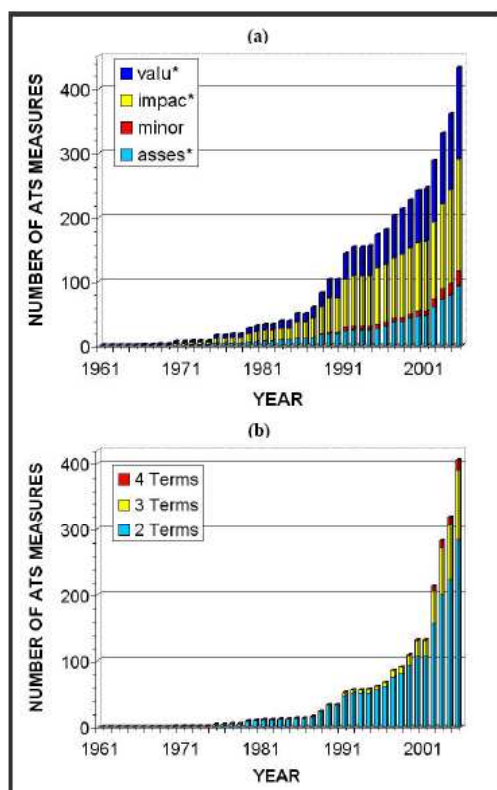


Figura 7: Perfiles de frecuencia acumulativa de medidas de política adoptadas a lo largo del tiempo en el Sistema del Tratado Antártico, que se derivaron de la 6ª edición de la *Antarctic Treaty Searchable Database* (Tabla 2). Estos perfiles relacionales se basaban en: **(a)** búsqueda de términos "minor", "asses*", "impac*" y "valu*", donde * es el carácter comodín; y **(b)** combinaciones de 2, 3 ó 4 de los términos de búsqueda anteriores. El número de medidas de política se iguala con el número de gránulos que se visualizan en las jerarquías expansibles-colapsables (p. ej., Fig. 5). Adaptado de Berkman et al. (2005).

La Figura 7a muestran que los términos clave se han incorporado de manera creciente a las nuevas políticas del Tratado Antártico, con el cambio más grande entre los conceptos de "impacto". Además, puede identificarse la fecha de su primer uso, como en el concepto "valor", que comenzó a aparecer en 1961. De manera similar, la Figura 7b muestra que las medidas de política incorporaron progresivamente 2, luego 3, y finalmente los 4 términos ambientales clave. Los análisis cuantitativos (Figs. 7a, b) no sólo apoyan la hipótesis anterior, sino que revelan que pueden extraerse objetivamente tendencias también de la información cualitativa, en relación con sistemas de coordenadas fijas, como tiempo o espacio.

3.3 Visualizaciones dinámicas persistentes

La *Antarctic Treaty Searchable Database* se ha expandido de 608 a 822 gránulos entre 1999 y 2005 (Tabla 2). Cada una de las ediciones anuales de la *Antarctic Treaty Searchable Database* se conserva en un webCDserver™ (Berkman 2002) que contiene una copia completamente ejecutable e independiente del sitio web con todos sus ficheros asociados. Este tipo de actividad de conservación puede utilizarse para archivar documentos digitales fijos (Gilliland-Swetland y Eppard 2000), mientras que facilita acceso persistente a entornos dinámicos, a pesar de la obsolescencia del hardware y el software originales. Tales soluciones son necesarias para resolver la paradoja de la conservación digital (Chen 2001): "*para mantener la información digital intacta*" mientras se proporciona "*acceso a esta información en un contexto de uso dinámico*"- que es una característica central del *International Research on Permanent Authentic Records in Electronic Systems Project* que implica a los archivos nacionales de 13 países (Duranti 2005a; <http://www.interpares.org>).

Una característica necesaria de los documentos en un archivo es que se fijen en el momento de la conservación para asegurar que no han sido alterados de manera no documentada. El resultado práctico de la fijeza es el acceso coherente y reproducible a los documentos. Con la *Antarctic Treaty Searchable Database*, la integración dinámica de los gránulos da como resultado jerarquías fiables, reproducibles y exactas para peticiones y periodos de tiempo dados (p. ej., Fig. 6). Bajo tales circunstancias, se fijarían los resultados de un proceso dinámico.

Los documentos que se crean en el curso de un asunto, "*constituyen una fuente primaria y privilegiada de evidencia acerca de las actividades y los actores implicados en ellos*" (Thibodeau 2001). Los documentos, que se guardan para su archivo, también tienen características necesarias (Duranti 2005b):

- forma fija que puede representarse;
- contenido no modificable;
- enlaces explícitos con otros documentos;

- contexto administrativo identificable;
- autor, destinatario y escritor; y
- acción en la que el documento participa o a la que apoya.

Sin embargo, en entornos digitales puede que estas seis características no sean necesarias para proporcionar la evidencia necesaria acerca de la exactitud de un documento digital que se generó de manera dinámica mediante un sistema informático en respuesta a una interacción o petición. Por ejemplo, si alguien discutiera la Figura 6 sobre la base de que hubo un error en su generación, ¿qué sería necesario para validar su exactitud?

La solución persistente implica ser capaz de reconstruir el documento con el software original o una emulación, y luego comprobar las anomalías en su contenido o relaciones (Thibodeau 2002). Para lograr esta reconstrucción, sería necesario tener documentación detallada acerca del contenido del sistema, los parámetros y funcionalidades en el momento en que se generó el documento, así como un log de la interacción o petición. Con la *Antarctic Treaty Searchable Database*, la documentación está representada por el contenido de los webCDservers™ (Tabla 2), los diagramas de flujo (Figs. 3 y 4) y las descripciones detalladas del sistema subyacente de integración digital (Berkman y Morgan 2003, Berkman et al. 2005).

El reto con los documentos digitales es proporcionar acceso persistente más allá de un pantallazo estático o un fichero de imagen bloqueado, que son de hecho documentos en copia dura. También resulta relevante considerar la eficiencia y eficacia de costes (Thibodeau 2001) de almacenar grandes volúmenes de documentos estáticos que son generados mediante procesos dinámicos basados en interacciones del usuario, como una petición a un sistema de información geográfica o una base de datos relacional para alguna decisión administrativa. Lo esencial es que los documentos estáticos son insuficientes a todos los efectos evidenciales, como se ilustra más arriba. En consecuencia, es necesario establecer estrategias y métodos para implantar documentos dinámicos que utilicen la estructura de información inherente, que es la única distinción entre recursos de información digitales y en copia dura (Figs. 1 y 2). La *Antarctic Treaty Searchable Database* y sus métodos subyacentes ofrecen un estudio de caso para implantar documentos dinámicos persistentes en los que pueda confiarse.

4 Conclusión

El desplazamiento de paradigma creado por las tecnologías digitales es la oportunidad para gestionar dinámica y objetivamente la estructura de información así como su contenido y contexto. A diferencia de las decisiones subjetivas que pueden variar de persona a persona para describir el contexto y el contenido de un documento, la estructura es un elemento inherente de un documento que puede describirse objetivamente. Es esta capacidad para utilizar

automáticamente la estructura inherente de información lo que distingue la gestión de la información en soportes digitales de los soportes en copia dura que había sido aplicada anteriormente en nuestra civilización (Figs. 1 y 2).

La *Antarctic Treaty Searchable Database* demuestra un método de integración bien definido que utiliza la estructura inherente de los recursos de información digital para generar automáticamente gránulos de información. Basándose en peticiones de integración definidas por el usuario, los gránulos de información pueden ser combinados de manera dinámica en un schema relacional exacto, fiable y reproducible. El poder de la granularidad automatizada reside en descubrir de manera eficiente relaciones objetivas entre recursos de información sin marcado convencional, metadatos o bases de datos (p. ej., Figs. 5-7). Tal integración de la información es relevante para programas bibliotecarios y archivísticos que requieran la conservación a largo plazo de recursos digitales auténticos, como los investigados por el *International Research on Permanent Authentic Records in Electronic Systems Project* (<http://www.interpares.org>). La granularidad automatizada también tiene implicaciones para realizar la visión de la World Summit on the Information Society cuando descubre que “*el conocimiento es la riqueza común de la humanidad.*”

5 Agradecimientos

Este ensayo se basa en una presentación en el congreso de CODATA de 2004 en Berlín, relativo al *International Research on Permanent Authentic Records in Electronic Systems Project*. Me gustaría agradecer a Luciana Duranti, Anne-Gilliland Swetland y Philip Eppard el haberme implicado en el Proyecto InterPARES. También me gustaría agradecer a Robert Chaddock de los National Archives and Records Administration las anteriores oportunidades para aprender acerca de los retos de conservar documentos digitales auténticos persistentes.

Este proyecto de la *Antarctic Treaty Searchable Database* originó discusiones en todo el Departamento de Estado de los Estados Unidos, y me gustaría agradecer a Raymond Arnaudo y Fabio Saturni el haber compartido continuamente información acerca del Sistema del Tratado Antártico. Información adicional acerca del Sistema del Tratado Antártico y otros acuerdos internacionales fue proporcionada por la Marine Mammal Commission, y me gustaría agradecer a Suzanne Montgomery estas oportunidades. El apoyo a la *Antarctic Treaty Searchable Database* ha sido generosamente proporcionado por la National Science Foundation (NSF/DUE-OPP 9652883, NSF/DUE 0329044 y NSF/ACI-9619020) y el Proyecto InterPARES. El acceso al *Digital Integration System* (DigIn®) y el mantenimiento continuo de la infraestructura de la *Antarctic Treaty Searchable Database* ha sido proporcionado por EvREsearch LTD en colaboración con Native Voices International.

6 Referencias

Berkman, P.A. 2002. Science into Policy: Global Lessons from Antarctica. Academic Press, San Diego.

Berkman, P.A. y Morgan, G.J. 2003. Automated granularity of authentic digital records in a persistent archive. Report for the National Archives and Records Administration. *EvREsearch Technical Report 2003-1*, Columbus. (http://www.sdsc.edu/NARA/Publications/EV_Report_2003G2_31aug03.doc).

Berkman, P.A., Morgan, G.J., Moore, R., Marciano, R., Suderman, J., Hamidzadeh, B., y Hofman, H. 2005. Antarctic Treaty Searchable Database Case Study. Final Report for the InterPARES 2 Project. International Research on Permanent Authentic Records in Electronic Archives (<http://www.interpares.org>).

Berners-Lee, T., Hendler, J. y Lassila, O. 2001. The semantic web. *Scientific American* 284(5):34-43.

Blumberg, R. y Atre, S. 2003. The problem with unstructured data. *DM Review* (February 2003).
Bobak, A. 1997. *Data Modeling and Design for Today's Architectures*. Artech House Publishers, Norwood.

Chen, S. 2001. The paradox of digital preservation. *Computer* (March 2001).

Daconta, M. 2005. Formal Taxonomies for the U.S. Government. XML.com (<http://www.xml.com/pub/a/2005/01/26/formtax.html>).

Duranti, L. 2005a. The long-term preservation of accurate and authentic digital data. The InterPARES Project. *CODATA Data Science Journal* 4:106-118.

Duranti, L. 2005b. The concept of record in experimental, interactive and dynamic environments." En: Guimaraes, J.A.C. (ed.). Diplomatic and Technological Approaches to the Analysis of the Record. Universidade Estadual Paulista, Malilia, Brazil. En prensa.

Duval, E., Hodgins, W., Sutton, S. E. y Weibel, S.L. 2002. Metadata Principles and Practicalities. *D-Lib Magazine* 8(4). <http://www.dlib.org/dlib/april02/weibel/04weibel.html>

Evans, B.G. 2000. Satellite Communications Systems. 3rd Edition. Institution of Electrical Engineers, London.

Fensel, D., Hendler, J., Lieberman, H. y Wahlster, W. 2003. *Spinning the Semantic Web*. MIT Press, Cambridge.

Friedl, J.E.F. 2002. *Mastering Regular Expressions*. 2nd Edition. O'Reilly, Sebastopol.

Gill, Y. y Ratnakar, V. 2001. A comparison of (semantic) markup languages. En: *Proceedings of the 15th International FLAIRS Conference*. Penscalo. Pp. 6.

Gilliland-Swetland, A.J. y Eppard, P.B. 2000. Preserving the authenticity of contingent digital objects: The InterPARES Project. *D-Lib Magazine* 6(7/8): <http://www.dlib.org/ar/dlib/july00/eppard/07eppard.html>

Greenberg, J. 2001. A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science and Technology* 52(11): 917 – 924

Hodge, G. 2001. *Metadata Made Simpler*. National Information Standards Organization, Bethesda.

Intel Corporation. 2005. *Microprocessor Quick Reference Guide*. <http://www.intel.com/pressroom/kits/quickref.htm>

Internet Systems Consortium. 2005. *Number of Internet hosts survey*. <http://www.isc.org/index.pl?ops/ds/host-count-history.php>

Jones, T.H. y Song, I-Y. 1996. Analysis of binary/ternary cardinality combinations in entity-relationship modelling. *Data and Knowledge Engineering* 19:39-64.

Lagoze, C., Payette, S., Shin, E. y Wilper, C. 2005. Fedora: An Architecture for Complex Objects and their Relationships. *Journal of Digital Libraries, Special Issue on Complex Objects* (en prensa).

Lyman, P., Varian, H.L., Swearingen, K., Charles, P., Good, N., Jordan, L.L. y Pal, J. 2003. *How much information 2003*. <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/>
Marine Mammal Commission. 1994. *Marine Mammal Commission Compendium of Selected Treaties, International Agreements, and Other Relevant Documents on Marine Resources, Wildlife, and the Environment*. Volumes 1-3. Marine Mammal Commission, Washington, D.C.

Maynard (a.k.a. Morgan), G.J. 2001. Information Management, Retrieval and Display Systems and Associated Methods. Patent No. 6,175,830. United States Patent and Trademark Office, Washington, D.C.

Maynard (a.k.a. Morgan), G.J. 2002. Information Management, Retrieval and Display Systems and Associated Methods. Patent No. 6,484,166. United States Patent and Trademark Office, Washington, D.C.

Maynard (a.k.a. Morgan) , G.J. 2003 Information Management, Retrieval and Display Systems and Associated Methods. Patent No. 515,007. Intellectual Property Office of New Zealand, Wellington.

Maynard (a.k.a. Morgan), G.J. 2004. Information Management, Retrieval and Display Systems and Associated Methods. Patent No. 770,087. IP Australia, Canberra.

Maynard (a.k.a. Morgan), G.J. 2006. Sistema De Administracion, Recuperacion y Despliegue Visual de Informacion y Metodo Asociado. Patent No. 233474. Instituto Mexicano de la Propiedad Industrial, Mexico City.

McGuinness, D.L. y van Harmelen, F. (eds.). 2004. OWL Web Ontology Language Reference. W3C (<http://www.w3.org/TR/owl-features/>).

National Research Council. 2001. Embedded, Everywhere: A Research Agenda for Networked Systems of Embedded Computers. National Academy Press, Washington, D.C.

Pastor-Satorras, R. y Vespignani, A. 2004. Evolution and Structure of the Internet: A Statistical Physical Approach. Cambridge University Press, Cambridge.

Senner, W.M. (ed.). 1989. The Origins of Writing. University of Nebraska Press, Lincoln.

Szykman, S., Senfaute, J. y Sriram, R.D. 1999. The use of XML for describing functions and taxonomies in computer-based design. *Proceedings of the 1999 ASME Design Engineering Technical Conferences*, Las Vegas.

Sowa, J.F. 1984. Conceptual Structures: Information Processing in Mind and Machine. Addison Wesley, Menlo Park.

Thibodeau, K. 2001. Building the archives of the future: advances in preserving electronic records at the National Archives and Records Administration. *D-Lib Magazine* 7(2). <http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html>

Thibodeau, K. 2002. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years. En: *The State of Digital Preservation: An International Perspective Conference Proceedings*. Institute for Information Science, Washington, D.C. Pp. 1-31.

United States Department of State. 1994. *Handbook of the Antarctic Treaty System*. 8th Edition. United States Department of State, Washington, D.C.