

AUTOMATED GRANULARITY TO INTEGRATE DIGITAL INFORMATION: THE “ANTARCTIC TREATY SEARCHABLE DATABASE” CASE STUDY

***Paul Arthur Berkman*^{1,2*}, *George James Morgan III*^{2,3}, *Reagan Moore*⁴ and *Babak Hamidzadeh*⁵**

^{*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

ABSTRACT

Access to information is necessary, but not sufficient in our digital era. The challenge is to objectively integrate digital resources based on user-defined objectives for the purpose of discovering information relationships that facilitate interpretations and decision making. The *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>), which is in its sixth edition, provides an example of digital integration based on the automated generation of information granules that can be dynamically combined to reveal objective relationships within and between digital information resources. This case study further demonstrates that automated granularity and dynamic integration can be accomplished simply by utilizing the inherent structure of the digital information resources. Such information integration is relevant to library and archival programs that require long-term preservation of authentic digital resources.

Keywords: Integration, Dynamic, Records, Digital, Archive, Library

1 INTRODUCTION

1.1 Era of Digital Information

We have reached the threshold in our ‘world information society’ when accessing more information does not equate with generating more knowledge. Knowledge, which emerges from understanding relationships within and between information resources, derives from the process of integration. Distinctions between information access and integration underlie technological solutions for the future when “*knowledge is the common wealth of humanity*”, as expressed by His Excellency Adama Samassekou at the 2004 CODATA meeting in Berlin. The purpose of this paper is to assess the challenges, strategies and efficiencies to integrate digital information resources.

To assess the challenges with the digital medium, it is instructive to take a broad view of written communications in our civilization. From stone and clay to paper onto digital media, each era has increased our capacity to transport, produce and integrate information (Fig. 1). For example, the Internet has been evolving since the late 1960’s (Berners-Lee et al. 2001, Pastor-Satorras and Vespignani 2004) with the number of Internet hosts increasing from 213 in 1981 to over 350,000,000 in 2005 (Internet Systems Consortium 2005). Since 1972, microprocessor speeds have increased 5 orders of magnitude (Intel

Corporation 2005) while satellite systems have made it possible to collect and transmit information on a global scale (Evans 2000). Moreover, the volume of digital information doubled in the three years after 1999 with more than 5 exabytes (10^{18} bytes) of information stored on print, optical and magnetic in 2002 alone (Lyman et al. 2003). We also have powerful search engines to retrieve digital information from vast warehouses. These features all point to the observation that access to digital information has become effectively infinite and instantaneous.

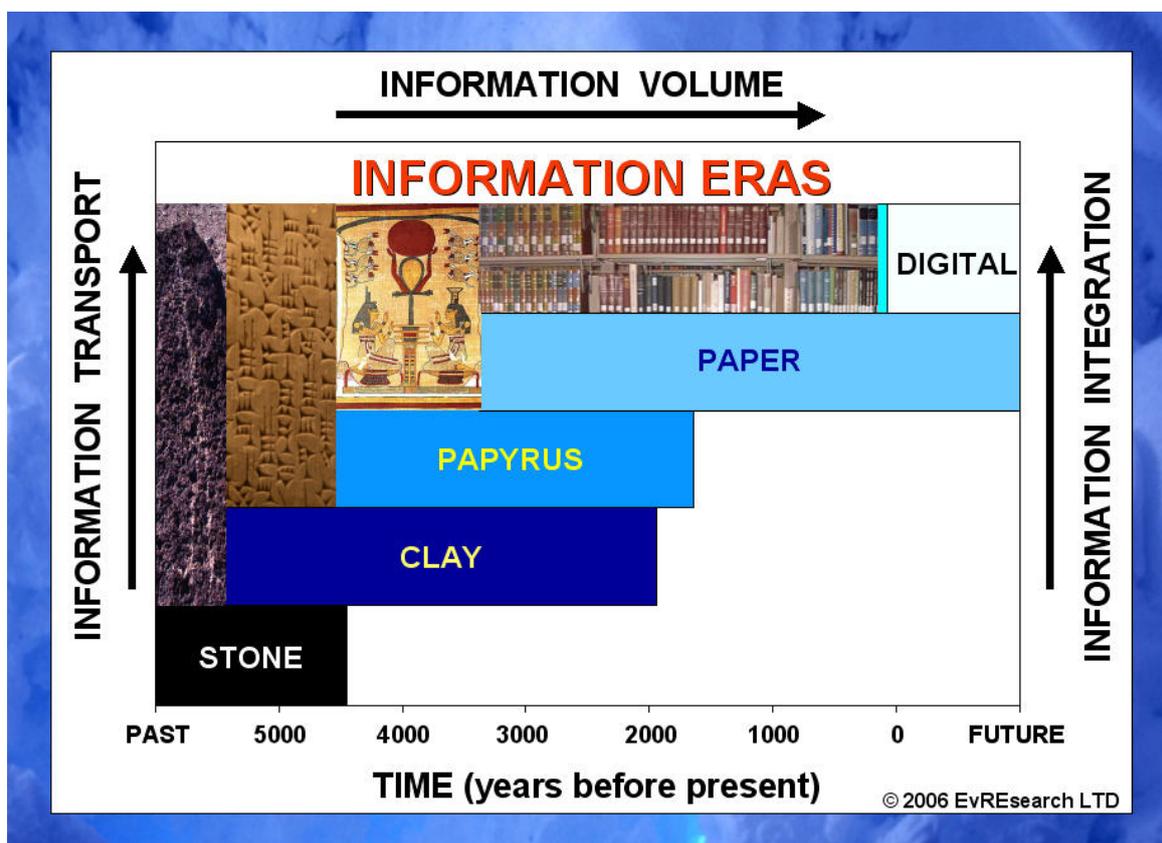


FIGURE 1: Thresholds in the preservation and dissemination of written information in our civilization. Each of the media prior to digital had been used for millennia (Senner 1989). From stone to digital media: (a) the transport of information across time and space has increased; (b) the volume and rate of information produced has increased; and (c) the capacity to integrate information into new knowledge has increased.

While it is easy to understand that the capacities to transport and produce information have increased with each era (Fig. 1), it is less obvious that the capacity to integrate information also has increased. Today, more than 80% of the digital information is considered to be “unstructured”, which means that it cannot be automatically decomposed into relational schema. Consequently, information integration is effectively limited to the remaining 20% of the digital resources that are structured with databases, metadata and markup. A principal challenge with the digital medium is being able to integrate information independent of whether it is “structured” or “unstructured” (Blumberg and Atre 2003).

1.2 Automated Granularity

Information has three indivisible elements – content, context and structure – that together provide meaning (Fig. 2). For example, when a message is encrypted (i.e., the structure is altered) it still has content and context, but no meaning. Similarly, if the names or dates and places are removed from an information

resource, it still has context and structure, but limited meaning without the salient facts. Removing context features that can be used to authenticate an information resource also will compromise its meaning.

The paradigm shift created by digital technologies is the opportunity to utilize the structure of information as well as its content and context. A printed book can be managed based on its content (as in libraries) or its context (as in archives), but it is not possible to break a book into smaller units that can be managed automatically. It is this ability to automatically manipulate the granularity of information resources that distinguishes digital media from all of the hardcopy predecessors that have been applied throughout human civilization.

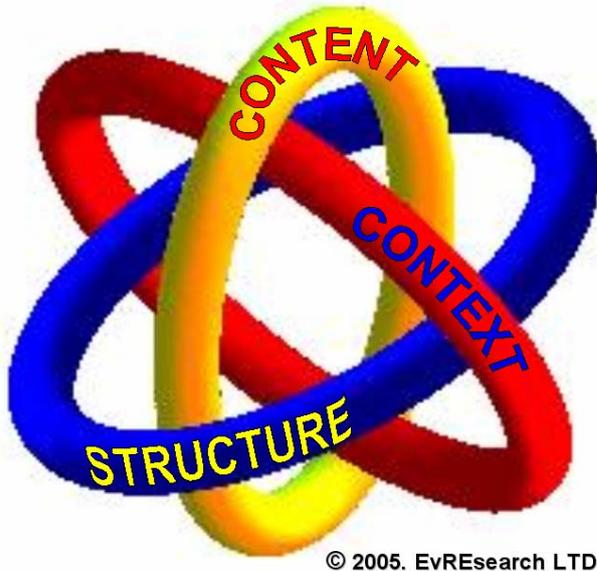


FIGURE 2: Borromean rings illustrating the three indivisible elements of information that together provide meaning.

This concept of granularity refers to the inherent conceptual units (i.e., information granules) that compose an information resource. With text, the granules could be as small as individual letters or characters, each of which could be identified within a byte-offset ontology relative to their parent resource (Berkman and Morgan 2003). More reasonably, the granules will be large enough to stand alone with sufficient content and context, such as paragraphs or chapters. A critical feature of each granule is that it internally retains information about its unique hierarchal position within its parent resource.

Automated granularity has been considered previously to implement networked systems of embedded computers with “*vision of a world filled with large numbers of computing elements, many of which are hidden inside other objects and networked together*” (National Research Council 2001). Automated granularity similarly extends to digital records, which have embedded content (Berkman and Morgan 2003). The value of automated granularity for digital libraries, archives and warehouses is that it provides a dynamic strategy for searching, retrieving, organizing and integrating both “structured” and “unstructured” information resources. This paper uses the *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>, previously <http://webhost.nvi.net/aspire>) to assess the applications and implications of automated granularity.

2 IMPLEMENTATION OF THE ANTARCTIC TREATY SEARCHABLE DATABASE

2.1 Implementation Technology

The underlying technology to implement the *Antarctic Treaty Searchable Database* is the *Digital Integration System* (DigIn®). Operation of DigIn®, which is based on patented technologies assigned to EvREsearch LTD (Maynard 2001, 2002, 2003, 2004, 2006), involves four principal modules that can be used together or separately:

- **GRANULARITY MODULE:** creates information granules by using the inherent structure and patterns that bound relevant units of content. A unique categorical tag is assigned to each granule based upon an analysis of its provenance, parent-child location and contents. The categorical tags contain information to generate expandable-collapsible hierarchies.
- **INDEX MODULE:** generates a database with the address (referenced within each categorical tag), content strings (words, numbers or other symbols) and their frequencies within each information granule.
- **INTEGRATION MODULE:** searches through the index to retrieve the information granules with terms or content strings that match the user-defined search queries in textual, numeric or other symbolic forms.
- **AGGREGATION MODULE:** combines relevant information granules based on their hierarchal relationships and user-defined criteria.

Each of the modules acts upon a set of expert rules that define its automated operation. These rules, which can be conveniently written with regular expressions (Friedl 2002), are optimized iteratively to integrate and display the relevant information granules within expandable-collapsible hierarchies. In addition, because DigIn® is modular, it can interface with statistical, graphical, semantic web, natural language or other types of software solutions that could be treated as additional modules.

DigIn® provides a general method that operates independently from any specific hardware and software. For example, DigIn® operates with ASCII or UNICODE as well as proprietary schema. DigIn® also operates with metadata, mark-up and databases that each have standardized patterns to organize information in a structured manner (e.g., Sowa 1984). Moreover, DigIn® currently is written in PERL, which provides a stable cross-platform programming language that can read and write binary files as well as process very large files. DigIn® also could be written in other languages depending on the circumstances. Consequently, DigIn® is an interoperable method that can be utilized into the future in a persistent manner.

2.2 Implementation Design

The general activities to create the digital record of the *Antarctic Treaty Searchable Database*, as well as similar databases of policy documents, are illustrated in Figure 3. The first step is to define the collection parameters, which includes the components of the collections as well as the resulting granularity and organization of the hierarchal displays that will be dynamically generated in response to the integration queries. After compiling the collection elements, the next step is to implement the appropriate granularity with a header tag in each granule that describe its unique hierarchal position relative to its parent resource. These tags, which preserve the provenance of each granule, will be used to dynamically generate expandable-collapsible hierarchies that comprehensively and objectively display granule relationships within and between the information resources. After searching and integrating the granules, the granule displays are assessed to determine whether the collection should be revised or whether the completed digital record should be fixed for archival purposes.

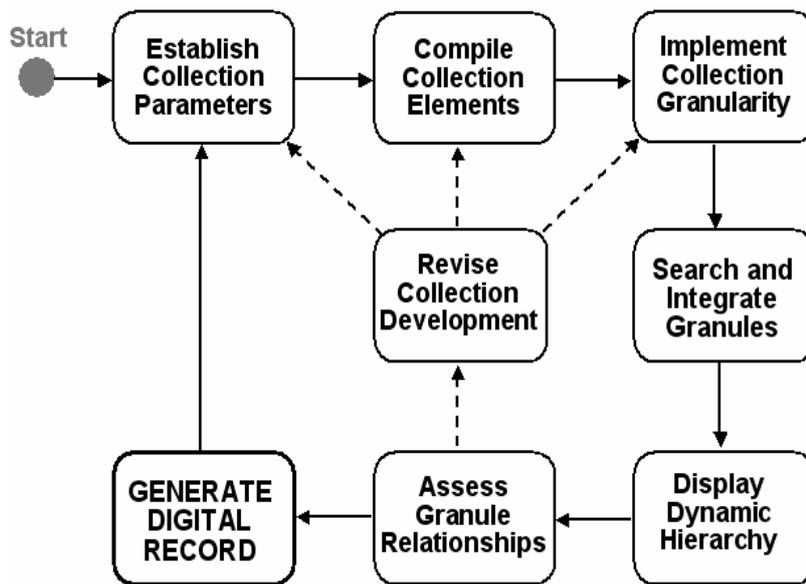


FIGURE 3: A generalized activity-flow diagram (Bobak 1997) of the processes to create the *Antarctic Treaty Searchable Database* or other digital records with the *Digital Integration System™* (DigIn®) from EvREsearch LTD. Adapted from Berkman et al. (2005).

More specifically, the initial edition of the *Antarctic Treaty Searchable Database* was implemented in collaboration with the National Science Foundation and United States Department of State. Based on the characteristics of the *Antarctic Treaty Handbook, 8th Edition* (United States Department of State 1994), the following rules were used for compiling the contents of the initial *Antarctic Treaty Searchable Database*:

Rule 1: Include only the “*measures*” that were adopted by the Antarctic Treaty Consultative Parties “*in furtherance of the principles and objectives of the Treaty.*”

Rule 2: Content of each adopted “*measure*” would include its text along with any tables or figures.

Rule 3: Exclude any “*extracts,*” “*introductory notes*” or other additions from the United States Department of State, which is the depository government, because they were not formally adopted by the Antarctic Treaty Consultative Parties

The next decision was to identify the appropriate granularity of the policy documents that would be searchable. Each Antarctic Treaty Consultative Meeting (ATCM) produced a report with adopted “*recommendations,*” “*decisions,*” “*measures*” or “*resolutions*”, which sometimes included “*appendices,*” “*annexes*” or “*attachments.*” Periodically, the Antarctic Treaty Consultative Parties also adopted Conventions and larger policy documents that included specific “*articles*” along with “*annexes.*” Based on these types of adopted measures, the following rules define the granularity of the policy documents for the *Antarctic Treaty Searchable Database*:

Rule 4: Each “*recommendation,*” “*decision,*” “*measure*” or “*resolution*” would be treated as a complete information granule (within the context of the ATCM and year of adoption as the two overlying hierarchal levels).

Rule 5: Each “*appendix,*” “*annex*” or “*attachment*” would be treated as a complete information granule (within the context of the “*recommendation,*” “*decision,*” “*measure*” or “*resolution*” as well as within the ATCM and year of adoption as the three overlying hierarchal levels).

Rule 6: Each “*article*” and “*annex*” would be treated as a complete information granule (within a Convention or Protocol and year of adoption as the two overlying hierarchal levels).

The initial edition of the *Antarctic Treaty Searchable Database*, which was constructed in an automated manner based on the above rules, has been continuously updated as:

- (1) new measures have been adopted by the Antarctic Treaty Consultative Parties;
- (2) missing measures have been identified; and
- (3) missing components from the measures (e.g., tables or figures) have been identified.

These updates involve the insertion, tagging and editing of individual granules. Each update or edition of the *Antarctic Treaty Searchable Database* has been fixed by preserving all files and functionality on a webCDserver™ (Berkman 2002). Throughout, the contents of the *Antarctic Treaty Searchable Database* have been incorporated directly from authentic sources (i.e., United States Department of State, Marine Mammal Commission, Committee for Environmental Protection, and host nations for the ATCM). Overall implementation of the *Antarctic Treaty Searchable Database* is illustrated in Figure 4.

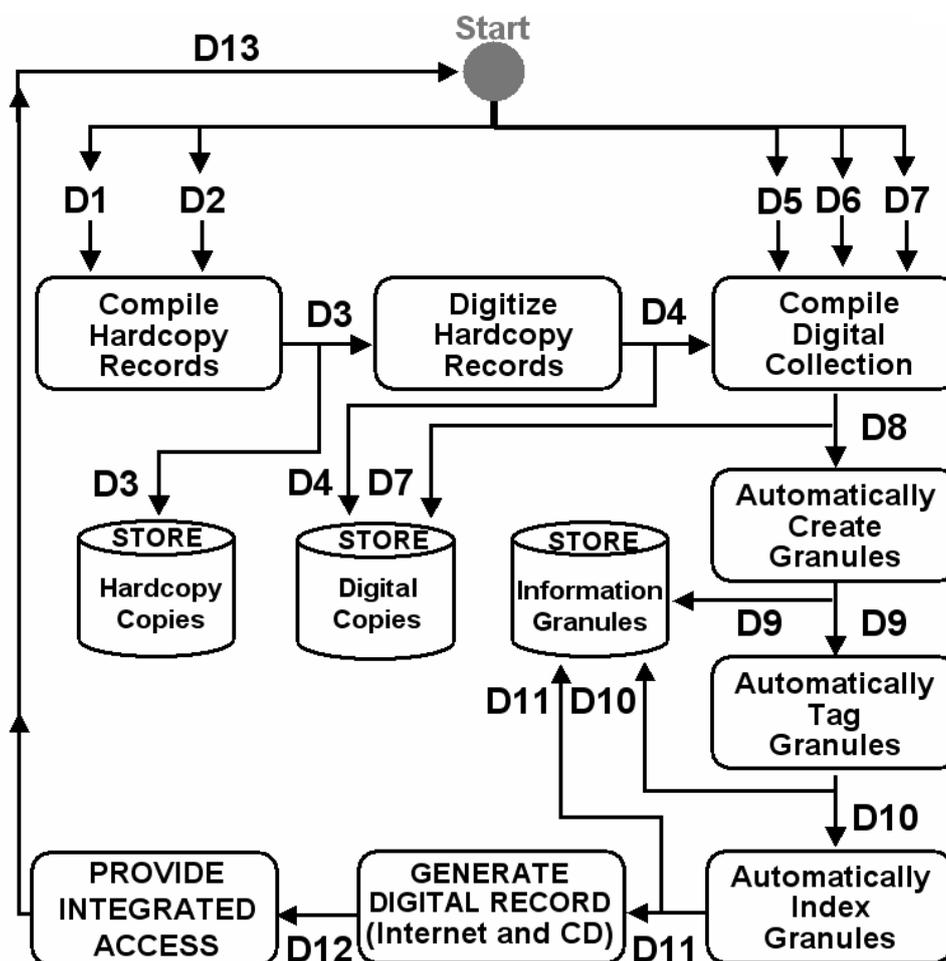


FIGURE 4: A data-flow diagram (Bobak 1997) to illustrate the activities along with specific data elements and data stores to implement the *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>). The data elements are: **D1** Marine Mammal Commission Compendia (Marine Mammal Commission 1994); **D2** Antarctic Treaty Consultative Meeting (ATCM) reports; **D3** Relevant hardcopy records; **D4** Digitized records; **D5** United States Department of State (1994) digital files; **D6** ATCM websites hosted by different Antarctic Treaty Consultative Parties; **D7** Committee for Environmental Protection (<http://www.cep.aq>) digital files; **D8** Digital records of entire documents; **D9** Information granules; **D10** Tagged information granules; **D11** Indexed-tagged information granules; **D12** Complete database granules; and **D13** New Website and webCDserver™ record (annually). Adapted from Berkman et al. (2005).

2.3 Implementation History

The history of the *Antarctic Treaty Searchable Database* goes back to 1998, when the United States Department of State was contacted about access to digital versions of the policy documents that they were managing as depository government for the *1959 Antarctic Treaty*. This query was prompted because information management was rapidly moving toward digital media and the *Antarctic Treaty Handbook* (United States Department of State 1994) had become unwieldy for case-study activities in an undergraduate Antarctic science and policy course that had been taught since 1982 (Berkman 2002). In 1999, within a month of initially implementing the *Antarctic Treaty Searchable Database*, the Department of State introduced it at the 23rd ATCM in Lima, Peru.

Although originally intended as a supplement for the university course on Antarctic science and policy (Berkman 2002), the *Antarctic Treaty Searchable Database* soon evolved into a digital archive that has been maintained and updated subsequently to benefit a diverse community of Antarctic stakeholders (Table 1). The redefined purpose of the *Antarctic Treaty Searchable Database* has been to facilitate knowledge discovery about the policies and strategies that promote “*international cooperation*” and the “*use of Antarctica for peaceful purposes only*” as stated in the *Preamble* of the *1959 Antarctic Treaty*.

TABLE 1. Representative Website links to the *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>, previously <http://webhost.nvi.net/aspire>)

<u>Website Host</u>	<u>Website URL</u>
International Government Institutions	
Antarctic Treaty Secretariat	http://www.ats.aq/
National Government Agencies	
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
Non-Governmental Organizations	
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
Businesses	
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/hurtigr/k1nncali.htm
Education Programs	
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

Two years after introducing the first edition of the *Antarctic Treaty Searchable Database* (Table 2), the Antarctic Treaty Consultative Parties fundamentally changed “*information exchange*” in the Antarctic

Treaty System by adopting Decision XXIV-1 at the 24th ATCM to establish the Antarctic Treaty Secretariat in Buenos Aires. As these international negotiations regarding the Antarctic Treaty Secretariat were underway, the *Antarctic Treaty Searchable Database* was linked to the websites for the 24th and 25th ATCM in St. Petersburg and Warsaw, respectively. In addition to being the first digital collection of Antarctic Treaty documents ever produced, the *Antarctic Treaty Searchable Database* remains as the most comprehensive source globally for integrating policy documents from the Antarctic Treaty System.

TABLE 2: Granularity, Coverage and Dimensions of the *Antarctic Treaty Searchable Database* (<http://aspire.nvi.net>, previously <http://webhost.nvi.net/aspire>) Through Time¹

Year Produced	Edition	Coverage	Granules		Embedded Images	
			Number	Text Volume (MB)	Number	Image Volume (MB)
1999	1 st	1959-1999	608	2.65	113	2.19
2001	2 nd	1959-1999	608	2.65	164	4.46
2002	3 rd	1959-2002	661	3.11	166	5.07
2003	4 th	1959-2003	720	3.67	200	6.67
2004	5 th	1959-2004	740	5.60	224	9.57
2005	6 th	1959-2005	822	7.63	352	21.40

¹ Copies of each *Antarctic Treaty Searchable Database* edition have been archived on fully-functional, stand-alone webCDserversTM. Adapted from Berkman et al. (2005).

3 KNOWLEDGE MANAGEMENT AND DISCOVERY

3.1 Conventional Granularity Limitations

The potential to discover meaningful relationships within and between the information resources is directly proportional to their granularity. For example, for a given search query, two books could generate 4 possible results (i.e., one book or the other, both books together, neither of the books). If each book were divided into two granules, there would be 16 possible combinations with 0, 1, 2, 3 or 4 granules. If each book were divided into four granules (i.e., 8 total), there would be 256 possible combinations with 0 to 8 granules. Consequently, among N granules there are 2^N possible relationships. Being able to express and then decompose the ternary, quaternary and higher-order relationships may reveal functional dependencies among the granules or digital entities (Jones and Song, 1996).

Practically, the number of possible relationships among even 100 digital objects (i.e., 2^{100}) is too large to manage comprehensively on the front-end. Nonetheless, conventional strategies involve descriptions of relationships on the front end with markup languages (Gill and Ratnakar 2001, Fensel et al. 2003) that add structure to information resources with tags to delimit, contain, or define the borders of certain content. For example, this front-end limitation applies to ontologies (McGuinness and Harmelen 2004, Lagoze et al. 2005) that describe relationships among components, properties, functions and processes of digital resources as well as taxonomies (Szykman et al. 1999, Daconta 2005). Aside from these limitations, there also is the practical feature that adding markup tags throughout a digital information resource is a form of contamination that may compromise its authentic content into the future. Importantly, by defining the relationships on the front end for the purpose of accessing information, results on the back end are effectively constrained, which greatly reduces the opportunity to be surprised. Given, these limitations and the suggestion that relationships cannot be managed comprehensively on the front-end, what strategies are available to reveal the 2^N relationships among granules on the back end?

In addition to applications of markup, which is considered to be structural metadata, knowledge discovery also is facilitated by descriptive and administrative metadata (Hodge 2001). With regard to descriptive metadata, there is an expanding universe of schema for different disciplines, institutions and activities (e.g., <http://www.mapageweb.umontreal.ca/turner/meta/english/>) that each contains different sets of attributes (e.g., name, size, data type, use restrictions, etc.) that must be defined or documented with

specified nomenclatures for every digital object (Duval et al. 2001). More importantly, metadata does not scale, which is a principal reason behind the widely-held notion that there is “structured” and “unstructured” information. Recognizing that all information has structure (Fig. 2), however, in reality digital information is either “managed” or “unmanaged” with conventional technologies.

A simple experiment can be constructed to illustrate the scalability limitations of metadata (Fig. 5). Consider a book that has a volume of 20 (in arbitrary units) and that each completed metadata schema has a volume of 1 (in arbitrary units). If the book is divided into two granules, each of which must have its own metadata schema, then the total volume of the book remains constant while the total volume of metadata schema has doubled. If the granularity is continuously doubled, the volume of metadata soon will overwhelm the volume of the actual data that is being managed. The additional metadata also requires increased effort to generate, store and process – which translates into costs and efficiencies. Moreover, if the metadata is stored in separate repositories, then loss of the metadata could compromise the preservation of the actual data.

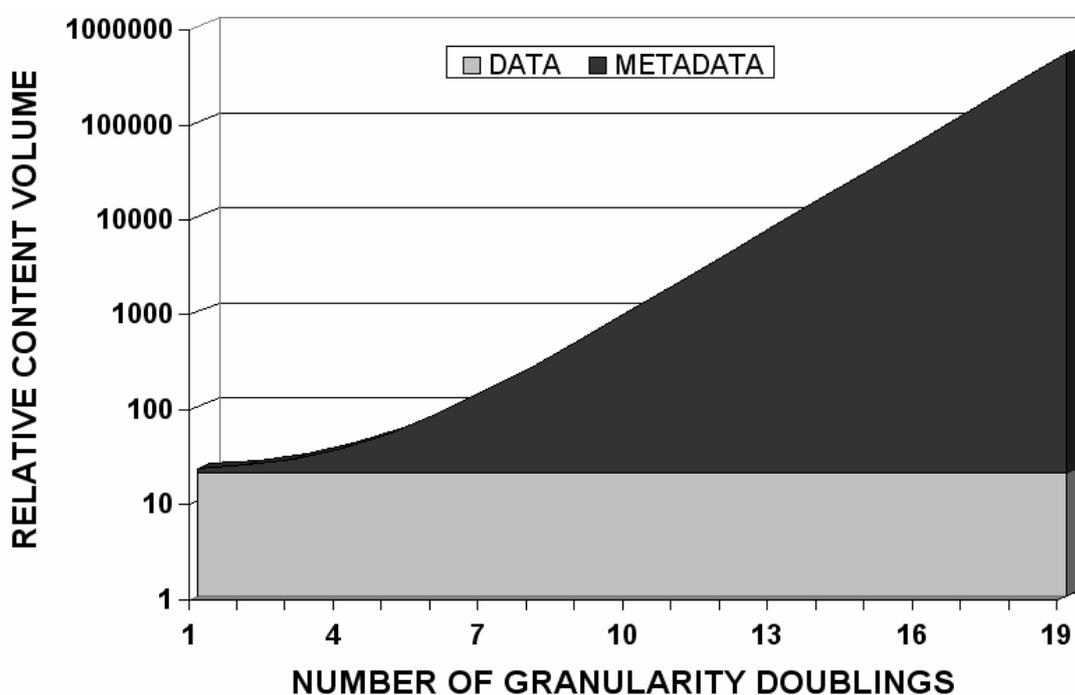


FIGURE 5: A simple model to illustrate the exponentially increasing volume of metadata, as the granularity is doubled, relative to the actual data within a single digital resource. In this model, the total volume of all information granules is constant (arbitrarily 20 units) and total volume of the metadata schema for each granule is fixed (arbitrarily 1 unit) independent of granule size. Adapted from Berkman and Morgan (2003).

The *Antarctic Treaty Searchable Database* is an example of information resources that are being managed with increased granularity, but without conventional metadata or markup. Moreover, the *Antarctic Treaty Searchable Database* integrates granules to achieve 2^N possible information relationships without conventional “database” manipulations of tables.

The capacity to discover relationships with the *Antarctic Treaty Searchable Database* is further reflected by its 822 information granules, in contrast to the Website for the United States Department of State (<http://www.state.gov/g/oes/rls/rpts/ant/>) that includes the *Handbook of the Antarctic Treaty System* in 18 ‘locked’ PDF files along with HTML files of five major documents (e.g., *1991 Protocol on Environmental Protection to the Antarctic Treaty*). With such websites, each user is required to conduct full-text searches, one digital resource at a time before the user is able to cut-and-paste and then organize the relevant pieces of

information – steps that are automated with the *Antarctic Treaty Searchable Database* and other DigIn[®] applications (e.g., *Marine Mammal Commission Digital Library of International Environmental and Ecosystem Policy Documents* – <http://nsdl.tierit.com>).

3.2 A Dynamic Integration Application

The ability to integrate and generate objective relational schema based on the inherent structure of the information resources can be illustrated with the *Antarctic Treaty Searchable Database*. From the 822 granules in the 6th edition of the *Antarctic Treaty Searchable Database* (Table 2), for example, 23 granules contain the term “peaceful” (Fig. 6).

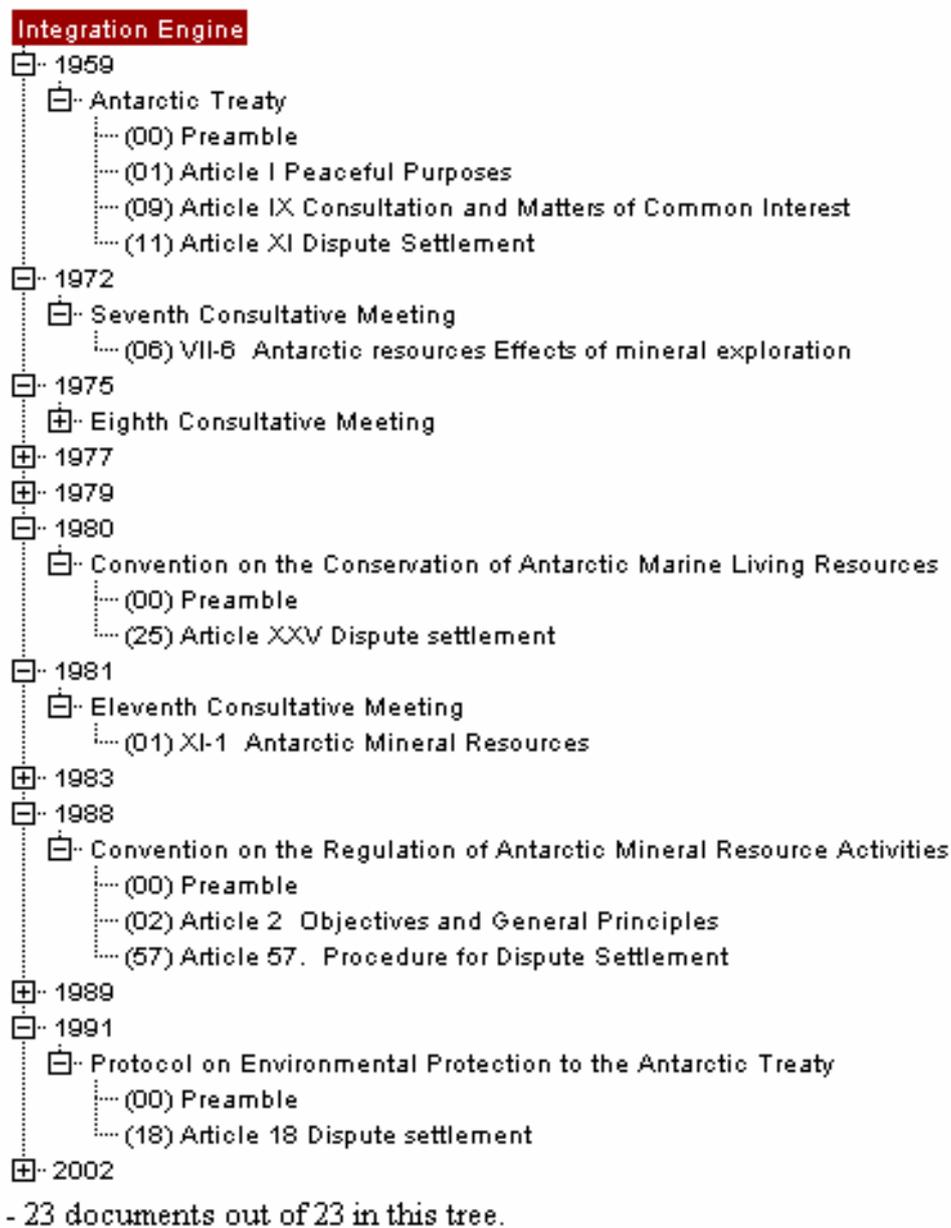


FIGURE 6: Expandable-collapsible hierarchy that was dynamically generated from the 6th Edition of the *Antarctic Treaty Searchable Database* (Table 2) with “peaceful” as the integration query. Objective policy relationships within and between years are derived from the information granules that were generated based on Rules 1-6 (see text).

The granules, which are displayed in the expandable-collapsible hierarchy, identify policy relationships within and between the Antarctic Treaty meetings that were convened from 1959 to 2005. As can be seen, “peaceful” is a common feature of “dispute settlement” in the legal institutions that emerged from the Antarctic Treaty in 1980, 1988 and 1991. Moreover, upon closer inspection of the individual granules, the same phrase was reproduced in each year (i.e., “...dispute resolved by negotiation, inquiry, mediation, conciliation, arbitration, judicial settlement or other *peaceful* means...”). These results are objective because all relevant granules (i.e., those with “peaceful”) are identified and each unique granule only occurs once in the hierarchy.

Relationships that can be displayed objectively also can be quantified accurately to test hypotheses, such as key policy concepts have been increasingly integrated into the adopted “measures” over time. As an illustration, consider Antarctic environmental protection, which involves human impacts that are assessed as being “minor or transitory” in relation to various Antarctic Treaty System values. Based on data extracted from the hierarchal displays (e.g., Fig. 6), trends in the incorporation of key environmental concepts into Antarctic Treaty measures can be identified (Fig. 7).

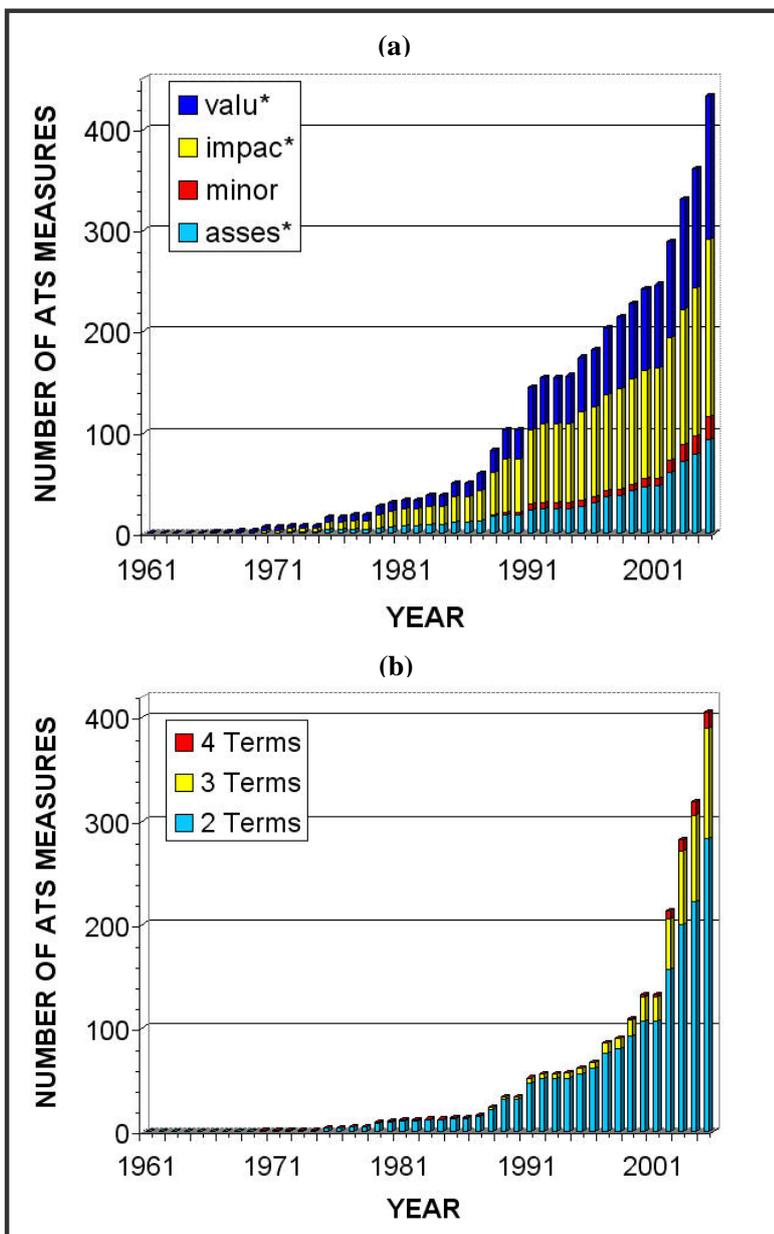


FIGURE 7: Cumulative frequency profiles of policy measures adopted over time in the Antarctic Treaty System that were derived from the 6th edition of *Antarctic Treaty Searchable Database* (Table 2). These relational profiles were based on: (a) search terms of “minor”, “asses*”, “impac*” and “valu*” where * is the wildcard character; and (b) combinations of 2, 3 or 4 of the above search terms. The number of policy measures equates with the number of granules that are displayed in the expandable-collapsible hierarchies (e.g., Fig. 5). Adapted from Berkman et al. (2005).

Figure 7a shows that key terms have been incorporated increasingly into new Antarctic Treaty policies, with the largest change among “impact” concepts. In addition, their date of first use can be identified, as with the “value” concept that began appearing in 1961. Similarly, Figure 7b shows that policy measures progressively incorporated 2 then 3 and finally all 4 of the key environmental terms. Not only do the quantitative analyses (Figs. 7a,b) support the above hypothesis, but they reveal that trends can be objectively extracted from otherwise qualitative information in relation to fixed coordinate systems, such as time or space.

3.3 Persistent Dynamic Displays

The *Antarctic Treaty Searchable Database* has expanded from 608 to 822 granules between 1999 and 2005 (Table 2). Each of the annual editions of the *Antarctic Treaty Searchable Database* is preserved on a webCDserver™ (Berkman 2002) that contains a fully-executable, stand-alone copy of the Website with all of the associated files. This type of preservation activity can be used to archive fixed digital records (Gilliland-Swetland and Eppard 2000) while facilitating persistent access in dynamic environments, despite the obsolescence of the original hardware and software. Such solutions are necessary to resolve the paradox of digital preservation (Chen 2001): “to maintain digital information intact” while providing “access to this information in a dynamic use context” – which is a central feature of the *International Research on Permanent Authentic Records in Electronic Systems Project* that involves the national archives from 13 countries (Duranti 2005a; <http://www.interpares.org>).

A necessary feature of records in an archive is that they are fixed at the time of preservation to ensure that they have not been altered in an undocumented manner. The practical result of fixity is consistent and reproducible access to records. With the *Antarctic Treaty Searchable Database*, the dynamic integration of granules results in reliable, reproducible and accurate hierarchies for given queries and time periods (e.g., Fig. 6). Under such circumstances, the results of a dynamic process would be fixed.

Records, which are created in the course of business, “constitute a primary and privileged source of evidence about the activities and the actors involved in them” (Thibodeau 2001). Records, which are set aside for archiving, also have necessary characteristics (Duranti 2005b):

- fixed form that can be rendered;
- unchangeable content;
- explicit linkages to other records;
- identifiable administrative context;
- author, addressee and writer; and
- action in which the record participates or supports.

In digital environments, however, these six characteristics may not be sufficient to provide the necessary evidence about the accuracy of a digital record that was generated dynamically by a computer system in response to an interaction or query. For example, if someone contested Figure 6 on the grounds that there was an error with its generation, what would be necessary to validate its accuracy?

The persistent solution involves being able to reconstruct the record with the original software or an emulation and then to test for anomalies in its content or relationships (Thibodeau 2002). To accomplish this reconstruction, it would be necessary to have detailed documentation about the system content, parameters and functionalities at the time the record was generated as well as a log of the interaction or query. With the *Antarctic Treaty Searchable Database*, the documentation is represented by the content of the webCDservers™ (Table 2), the flow diagrams (Figs. 3 and 4) and detailed descriptions of the underlying digital integration system (Berkman and Morgan 2003, Berkman et al. 2005).

The challenge with digital records is to provide persistent access beyond a static screenshot or locked image file, which are effectively hardcopy records. It also is relevant to consider the efficiency and cost-effectiveness (Thibodeau 2001) of storing large volumes of static records that are generated by dynamic processes based on user interactions, such as querying a geographic information system or relational database for some administrative decision. The bottom line is that static records are insufficient for all evidentiary

purposes, as illustrated above. Consequently, it is necessary to establish strategies and methods to implement dynamic records that utilize the inherent structure of information, which is the unique distinction between digital and hardcopy information resources (Figs. 1 and 2). The *Antarctic Treaty Searchable Database* and its underlying methods offer a case study to implement persistent dynamic records that can be trusted.

4 CONCLUSION

The paradigm shift created by digital technologies is the opportunity to dynamically and objectively manage the structure of information as well as its content and context. Unlike the subjective decisions that may vary from person to person to describe the context and content of a record, the structure is an inherent element of a record that can be described objectively. It is this ability to automatically utilize the inherent structure of information that distinguishes information management with digital media from the hardcopy media that had been applied previously in our civilization (Figs. 1 and 2).

The *Antarctic Treaty Searchable Database* demonstrates a well-defined integration method that utilizes the inherent structure of digital information resources to automatically generate information granules. Based on user-defined integration queries, the information granules then can be dynamically combined into accurate, reliable and reproducible relational schema. The power of automated granularity is in efficiently discovering objective relationships among information resources without conventional markup, metadata or databases (e.g., Figs. 5-7). Such information integration is relevant to library and archival programs that require long-term preservation of authentic digital resources, as investigated by the *International Research on Permanent Authentic Records in Electronic Systems Project* (<http://www.interpares.org>). Automated granularity also has implications for realizing the vision of the World Summit on the Information Society when discovering “*knowledge is the common wealth of humanity.*”

5 ACKNOWLEDGEMENTS

This paper is based on a panel presentation at the 2004 CODATA meeting in Berlin regarding the *International Research on Permanent Authentic Records in Electronic Systems Project*. I would like to thank Luciana Duranti, Anne-Gilliland Swetland and Philip Eppard for involving me in the InterPARES Project. I also would like to thank Robert Chadduck at the National Archives and Records Administration for earlier opportunities to learn about the challenges of preserving persistent authentic digital records.

This *Antarctic Treaty Searchable Database* project originated with discussions through the United States Department of State and I would like to thank Raymond Arnaudo and Fabio Saturni for continuously sharing information about the Antarctic Treaty System. Additional information about the Antarctic Treaty System and other international agreements was provided by the Marine Mammal Commission and I would like to thank Suzanne Montgomery for these opportunities. Support for the *Antarctic Treaty Searchable Database* has been generously provided by the National Science Foundation (NSF/DUE-OPP 9652883, NSF/DUE 0329044 and NSF/ACI-9619020) and the InterPARES Project. Access to the *Digital Integration System* (DigIn[®]) and continuous maintenance of the infrastructure for the *Antarctic Treaty Searchable Database* have been provided by EvREsearch LTD in collaboration with Native Voices International.

6 REFERENCES

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

Berkman, P.A. and 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., and 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. and Lassila, O. 2001. The semantic web. *Scientific American* 284(5):34-43.
- Blumberg, R. and 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." In: Guimares, J.A.C. (ed.). *Diplomatic and Technological Approaches to the Analysis of the Record*. Universidade Estadual Paulista, Malilia, Brazil. in press.
- Duval, E., Hodgins, W., Sutton, S. E. and 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. and Wahlster. W. 2003. *Spinning the Semantic Web*. MIT Press, Cambridge.
- Friedl, J.E.F. 2002. *Mastering Regular Expressions*. 2nd Edition. O'Reilly, Sebastapol.
- Gill, Y. and Ratnakar, V. 2001. A comparison of (semantic) markup languages. In: Proceedings of the 15th International FLAIRS Conference. Penscalo. Pp. 6.
- Gilliland-Swetland, A.J. and 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. and 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. and Wilper, C. 2005. Fedora: An Architecture for Complex Objects and their Relationships. *Journal of Digital Libraries, Special Issue on Complex Objects* (in press).

Lyman, P., Varian, H.L., Swearingen, K., Charles, P., Good, N., Jordan, L.L. and 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. and 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. and 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. and 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. In: 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.