White Paper

Knowledge Discovery in Our World Information Society: Automated Granularity and the End of Unstructured Information

July 10, 2004

The voyage of discovery is not in seeking new landscapes but in having new eyes.

Marcel Proust (1871 – 1922)

INTRODUCTION

The playing field has changed. We have entered the digital frontier into a black hole of structured and unstructured information.

There is no controversy about the exponential growth of digital information and there is no real consensus about the best way to manage it. As the volume of information vaults from terabytes to petabytes to exabytes, we find that current technologies are inadequate to utilize all the information that is being created. In effect, access to digital information already has become infinite and instantaneous. The challenge now and into the distant future is to integrate digital information to discover knowledge based on our own individually defined objectives.

Today, technologies to integrate digital information are rooted in:

- Databases (e.g., DB2, Access, 9iRAC);
- Mark-up languages (e.g., SGML, HTML or XML); and
- Metadata (e.g., Dublin Core, IEEE).

These technologies impose their own structure on digital information; hence ‘structured information’ was born. The problem is that the vast majority of digital information is not structured with these technologies.

> ***“According to analysts, more than 80% of corporate data comes in the form of unstructured content…”***
>
> **Documentum®**

Consequently, there a huge bottleneck in our ability to integrate digital information at the rate it is being produced (Fig. 1).

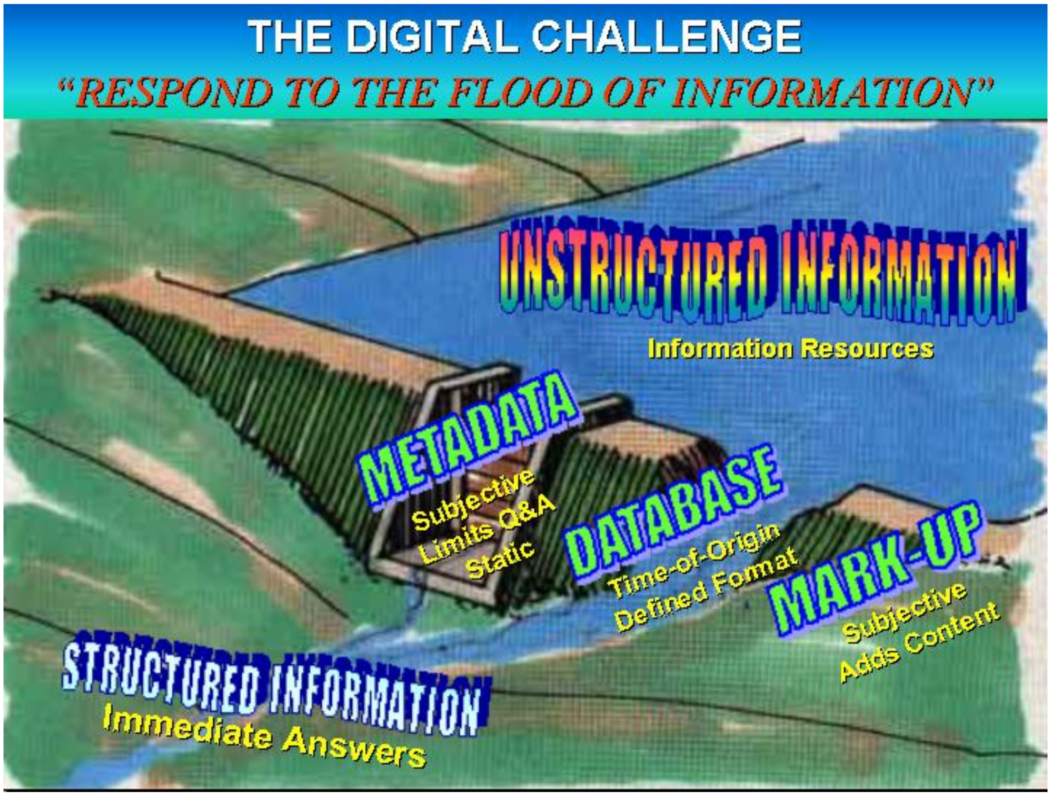


***FIGURE 1**: The digital information bottleneck. Limitations of content-centric technologies (i.e., metadata, databases, and mark-up) prevent knowledge discovery from the full 100% of digital information at the rate it is being produced.*

Most of the digital information that exists (and is being created) is referred to as unstructured information. This is information provided in free form, such as news articles, research papers, e-mails and memos, digital audio and video files as well as most of what you find on the World Wide Web. However, because unstructured information has content that is not organized with relational databases, mark-up languages or metadata - it is not readily accessible for computer analysis.

> ***"In order to allow organizations to extract the full value of their content, it is necessary to make the content more accessible and understandable to their IT systems. This is enabled by augmenting the unstructured text with structured elements called meta-tags."***
>
> **ClearForest®**

Consequently, it is generally believed that:

> ***"Unstructured data is data that cannot be decomposed into a relational schema."***
>
> ***Oracle®***

- Does the existence of "unstructured" information mean that it is impossible to decompose a report into its constituent sections and subsections?
- More practically, is unstructured information merely a contrived concept to identify the arena of digital information that remains unmanaged with current technologies?
- In fact – does information even exist without structure?

Conventional technologies require an *a priori* interpretation of the content to manage the "structured" 20% of the digital information. **However, information has two basic ingredients - *content and structure* - and it is the structure of the information that enables the content to express itself.** These two intertwined characteristics of 'information' are illustrated by an encrypted message that has content, but with a hidden structure that obscures any meaning.

Unstructured or unmanaged, in relation to business and the operation of our information society from home PCs to government agencies, the new playing field involves 100% of the digital information. The paradigm shift - to achieve meaningful management of the so-called unstructured information - will come with tools that utilize the structure as well as the content of digital records. A key tool to escape the paper paradigm – as well as the black hole of digital information - is automated granularity.

In this white paper, we discuss the advantages of utilizing the inherent structure of digital information to facilitate comprehensive access, content integration and knowledge discovery independent of scale or format of the digital records. We then discuss the operation of the patented EvREsearch® *Digital Integration System*™ (DIGIN™) to identify relationships among the constituent elements within and between digital records - comprehensively, objectively and automatically.

DIGIN™ ARCHITECTURE

DIGIN™ - which derives from the *“Information Management, Retrieval and Displays Systems and Methods”* (United States Patent Nos. 6,175,830 and 6,484,166) – has a modular architecture with four principal modules:

- **BREAK MODULE**: objectively generates and automatically creates categorical tags for information granules based on inherent patterns in the parent information resource(s);

- **INDEX MODULE:** automatically generates a database with the address (referenced within each categorical tag) and content strings (words, numbers or other symbols) of each information granule;
- **SEARCH MODULE:** dynamically generates expandable-collapsible hierarchal displays based on user-defined queries; and

- **UN-BREAK MODULE:** automatically combines relevant information granules based on user-defined criteria.

Each of these processing modules acts as an expert engine operating upon a set of expert rules that define its operation. These rule sets are optimized iteratively to objectively organize, identify and display related units of information within collections of digital records.

The BREAK MODULE is a unique feature of this patented technology. Based on rule sets, the break module parses through an information resource to break it into information granules (as determined by user-defined requirements). These rule sets utilize the inherent structural patterns or boundaries of the logical information units within and between digital-record groups, series and entities.

The BREAK MODULE also creates categorical tags for each of these information granules. The categorical tags are assigned to each of the granules, based upon an analysis (defined by the set of expert system rules) of the location and contents of each logical information unit. In addition, the categorical tag can include standard classifications, such as Dewey Decimal-type numbers or metadata. Consequently, these logical information units and their associated tags can be used to generate hierarchies that objectively define relationships within and between collections of information. No other system utilizes this interoperable approach to manage and integrate digital records.

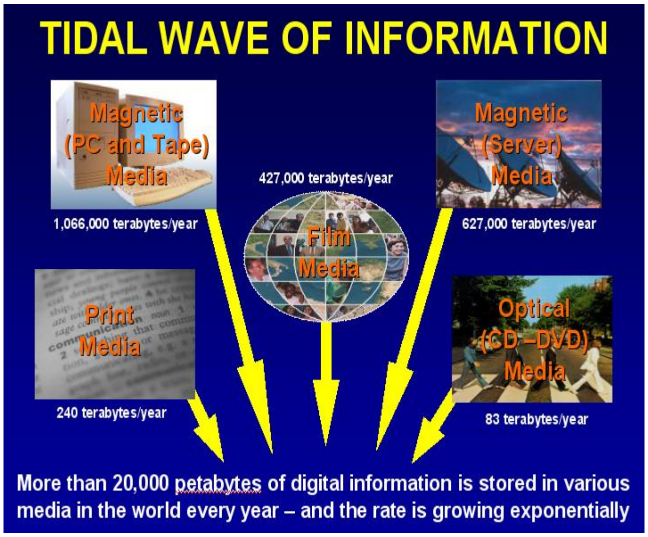
The rule sets for the INDEX MODULE and SEARCH MODULE reflect the user-defined requirements for accessing the logical information units. Each indexed record includes the information strings contained within (e.g., words, phrases, symbols) and their frequencies (i.e., weight). Based on the user's requirements, the rule sets for the INDEX MODULE may be configured for static digital records (such as reports or e-mail messages) as well as dynamic streams of information (such as sensor data or news feeds).

Utilizing user-defined search queries (in textual, numeric or other symbolic forms), the SEARCH MODULE then searches comprehensively through the reverse index for the information granules with matching terms or strings. By applying the categorical tags and/or weight of their matching search strings, the relevant information granules then can be integrated and displayed objectively in expandable-collapsible hierarchies (tree structures) with combinations of tiers and ordering schemes that are virtually limitless.

The UN-BREAK MODULE assembles related information granules objectively after each search to reconstruct a contiguous portion of an original information resource or to create new information resources. Because of its modularity, this patented technology is *expandable* and can include any other knowledge discovery modules (e.g., statistical, pattern recognition, semantic-web, natural language, concept-map, etc.) that further support the knowledge discovery goals of each and every user.

OUR WORLD INFORMATION SOCIETY[1]

Innovations in computing have revolutionized our capacity to access information in all shapes and forms from boardrooms to classrooms. Those same innovations and our ability to communicate instantly worldwide have accelerated the growth of information beyond anything imaginable even a few years ago. The maturation of the Internet, combined with the huge (and growing) memory capacity of today's computers, has created our current condition of information overload with more information at our fingertips than we can utilize effectively.



TIDAL WAVE OF INFORMATION

More than 20,000 petabytes of digital information is stored in various media in the world every year – and the rate is growing exponentially

FIGURE 2: Production rates of digital information from various storage media estimated in 2000. One petabyte equals one billion megabytes. Data from http://www.sims.berkeley.edu/research/projects/how-much-info/).

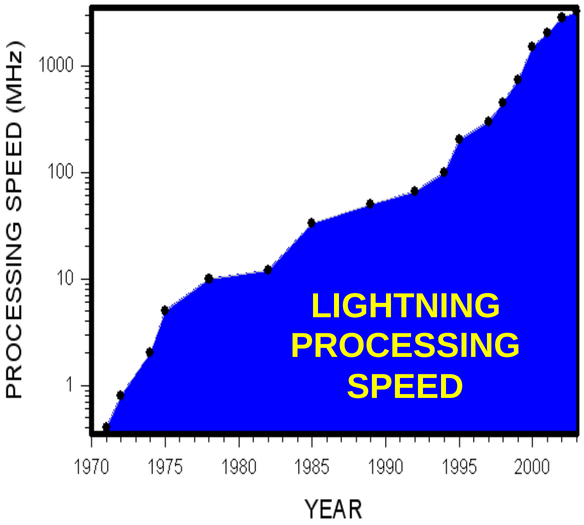


FIGURE 3: Exponentially increasing speed of computer microprocessors since their commercial inception in the early 1970's Data from *http://www.intel.com*.

With the unprecedented access to information via the internet and computers (Figs. 2 and 3), a growing problem in our new information society is that:

more information does not equal more knowledge.

This message is poignantly illustrated by the notion that more than 80% of corporate data comes in the form of unstructured content that "*cannot be decomposed into relational schema"* to discover knowledge. The challenges and opportunities with digital information

[1] Reflected by the goals of the World Summit on the Information Society (http://www.itu.int/wsis/).

are now a matter of scale. With exploding volumes of digital information, there is increasing urgency to describe, store, archive, process and access more and more information to discover knowledge that has meaning for each of us individually.

Integrating information to discover knowledge from digital records is a wide open playing field. Consider that 20% of the digital information market generates $100 billion, while the remaining 80% accounts for maybe $1 billion in annual revenues at this point. **There are no comprehensive, persistent, interoperable solutions to integrate digital records that are commercially available – yet.**

This new playing field is represented by the hundreds of small businesses that are jockeying for position, creating dozens of original-equipment-manufacturer (OEM) relationships, like weeds that rapidly colonize barren landscapes. Eventually the more robust companies emerge like shrubs that outcompete the weeds and consolidate the playing field by merging with other companies. Those companies that have deep roots and the strength to survive over the long term will appear as trees that shade out the underlying shrubs. In this market succession (Fig. 4), the existence of all the OEM relationships alone indicates that the knowledge-discovery market is immature and ripe with opportunity.

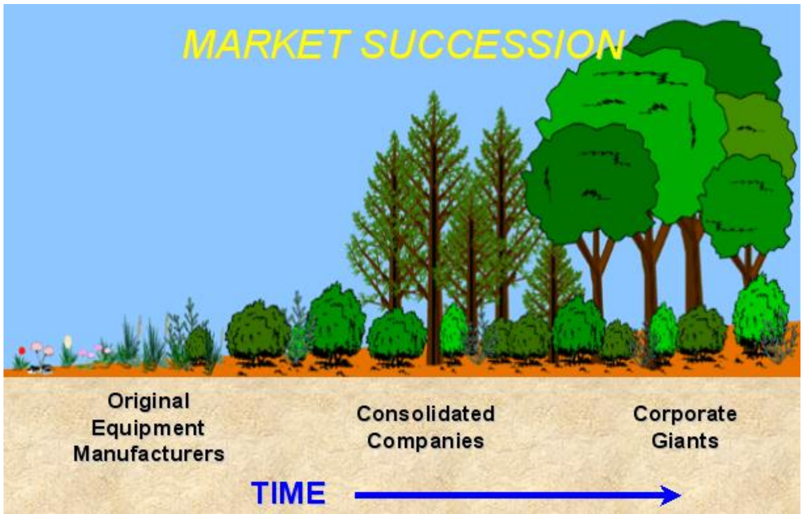


FIGURE 4: *Sequence of corporate stages in a maturing market. The predominance of original equipment manufacturer (OEM) relationships to create knowledge-discovery solutions indicates that this digital information market still is immature.*

MANAGING DIGITAL INFORMATION

The most common solution to manage digital information is by herding it into unstructured storage bins called Binary Large Objects (BLOBS).

> ***"Binary Large Objects or Large Objects (LOBs) in general, are used to store unstructured data."***
>
> **Oracle®**

As an example of the problem and opportunity, consider a simple back-of-the-envelope calculation for e-mail from a large corporation. The corporation has 10,000 employees each of which is receiving a mere 25 messages each day, which means that 250,000 e-mail messages need to be stored daily. For simplicity, the corporation backs up the e-mail messages every four days into a folder (each of which is a BLOB) that is described with a metadata tag. In the course of a year, the corporation would have produced over 120 e-mail BLOBS with more than 120,000,000 e-mail messages.

Suddenly, the corporation has to identify all of the e-mail messages that related to Company X with the results organized by date and employee. What could the corporation do to accomplish this task by the next day?

Search Engines:

To access the relevant information, the corporation could use a search engine that would create long lists of e-mail messages for each BLOB. With brute force, the corporation then could tackle the problem by creating successive subsets of the e-mail list for each employee over time. However, the corporation must comply with regulatory requirements (e.g., the Sarbannes-Oxley Act of 2002) and have the resulting e-mail lists the following day or face large fines. The corporation is running out of time - there isn't enough manpower – what is the corporation to do?

Search engines (which retrieve and provide access to information) are the most common tools used for handling digital information, especially for managing the biggest BLOB of all – the World Wide Web. However, applications of search engines are like building a library without an effective cataloguing scheme. Consequently, for each search, we would stack the pertinent books in a large pile (just like the ranked lists of "hits" that we generate with each digital search query). We then would need to check titles and open pages to identify information of interest. Subsequent analyses to describe relationships and trends within and between the books would involve manual integration by "cutting and pasting" into a new document – and an unbearably inefficient investment of time to turn a collection of data sources into usable insights.

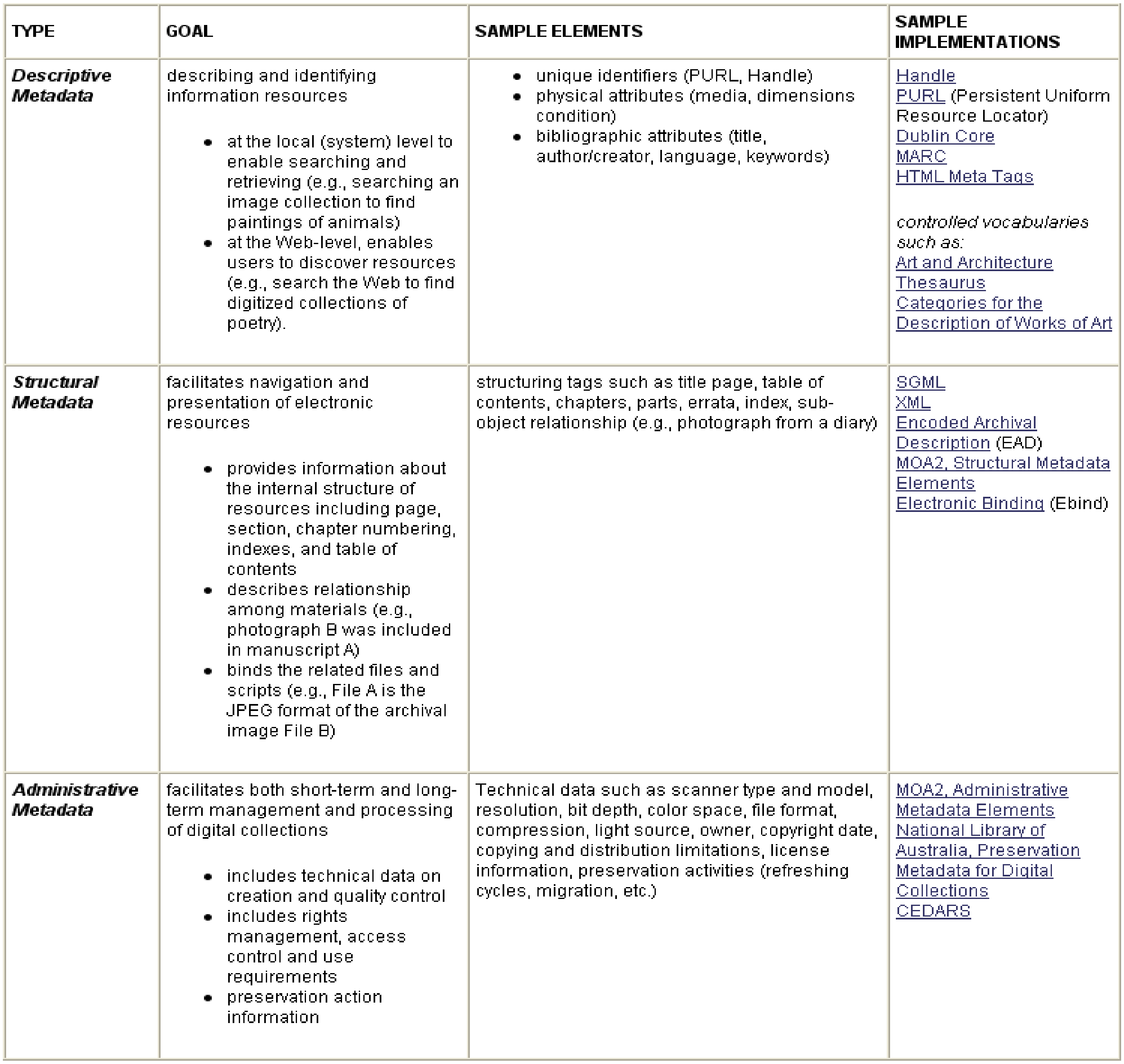
Metadata:

In a world of paper, information management strategies (like the Dewey Decimal System or the Library of Congress system) have been created to tag, relate and locate materials in hierarchies of shelves on floors within libraries. However, unlike books that are tagged once and forever, digital records have the potential to be automatically re-tagged within the context of ever-changing collections based on user-defined objectives.

For these digital collections of information, various "standards" of metadata tags have emerged to describe, structure and administrate digital records (Table 1). However, because of the free-form nature of most digital records, metadata tags contain numerous fields that must be addressed for each digital record to create a standardized collection that can managed, organized and searched. Dublin-Core metadata (http://dublincore.org), for example, contains 15 common fields that must be addressed for each digital record, including:

- Subject (with keywords);
- Subject (with controlled vocabularies);
- Subject (with Dewey Decimal classification);
- Description (abstract);
- Title:
- Date;
- Creator;
- Publisher; and
- Type (category of the resource).

TABLE 1: Metadata types (from http://library.cornell.edu/preservation/tutorial/metadata/table5-1.html).



TYPE	GOAL	SAMPLE ELEMENTS	SAMPLE IMPLEMENTATIONS
Descriptive Metadata	describing and identifying information resources • at the local (system) level to enable searching and retrieving (e.g., searching an image collection to find paintings of animals) • at the Web-level, enables users to discover resources (e.g., search the Web to find digitized collections of poetry).	• unique identifiers (PURL, Handle) • physical attributes (media, dimensions condition) • bibliographic attributes (title, author/creator, language, keywords)	Handle PURL (Persistent Uniform Resource Locator) Dublin Core MARC HTML Meta Tags *controlled vocabularies such as:* Art and Architecture Thesaurus Categories for the Description of Works of Art
Structural Metadata	facilitates navigation and presentation of electronic resources • provides information about the internal structure of resources including page, section, chapter numbering, indexes, and table of contents • describes relationship among materials (e.g., photograph B was included in manuscript A) • binds the related files and scripts (e.g., File A is the JPEG format of the archival image File B)	structuring tags such as title page, table of contents, chapters, parts, errata, index, sub-object relationship (e.g., photograph from a diary)	SGML XML Encoded Archival Description (EAD) MOA2, Structural Metadata Elements Electronic Binding (Ebind)
Administrative Metadata	facilitates both short-term and long-term management and processing of digital collections • includes technical data on creation and quality control • includes rights management, access control and use requirements • preservation action information	Technical data such as scanner type and model, resolution, bit depth, color space, file format, compression, light source, owner, copyright date, copying and distribution limitations, license information, preservation activities (refreshing cycles, migration, etc.)	MOA2, Administrative Metadata Elements National Library of Australia, Preservation Metadata for Digital Collections CEDARS

In a practical sense, the subjective effort to create the unique identifiers for every digital record makes the metadata tags just as static as the tags on a book. Moreover, it is largely because of this static and subjective nature of metadata tags that we have the ‘digital information bottleneck’ (Fig. 1). It is also noteworthy that metadata represents a ‘paper paradigm’ that has been evolving in digital contexts largely since in the early 1990’s with input from the Online Computer Library Center (OCLC), which owns the Dewey

Decimal System and is the creator of Dublin-Core Metadata. This is not to say that metadata doesn't have value, just that the utility and implementation of metadata varies with different types of digital information (Fig. 5).

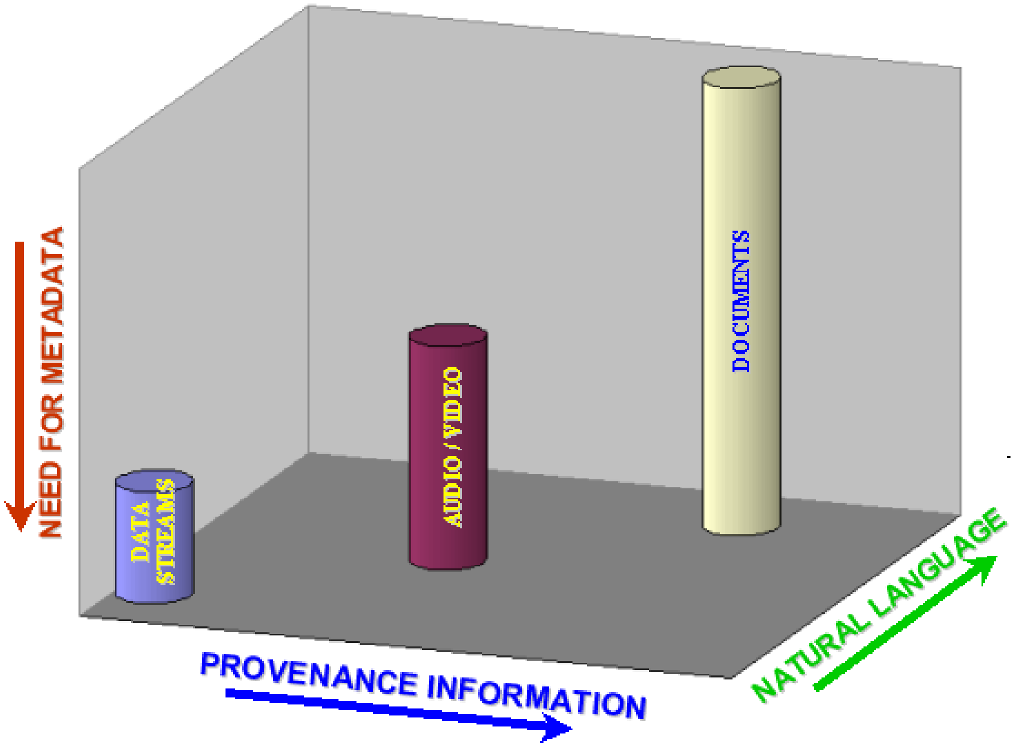


FIGURE 5: *Differential utility of metadata for different types of information in relation to the content that is in natural language formats with explicit details about their provenance (which includes the date, time, location and creator). Arrows point toward increasing magnitudes of: provenance information (e.g., creator, date and title), natural language content (as opposed to other symbolic representations); and the need for metadata descriptions.*

Tools that apply only to natural languages (like taxonomies, controlled vocabularies and semantic webs) are used to populate metadata for documents, but have no utility with binary data, human-genome sequences or other symbolic representations. In contrast, the EvREsearch® technologies (which are based on the patented *"Information Management, Retrieval and Display Systems and Associated Methods"*) can be applied to any format of digital information.

For text, which is the low-hanging fruit that account for much of the unstructured information in particular, descriptive metadata is largely redundant and incomplete. Why not let the information express itself without limiting it to the terms, phrases and concepts that are identified by the user? Every book has a title, author, publisher and date of publication. More importantly, the book has content organized in the contexts that the subject expert found to be most appropriate. Any application of subject abstracts, keywords and controlled vocabularies will be biased by the author of the metadata. Consequently, unless every substantive term in the book is included, the metadata will never be comprehensive and it will be impossible to facilitate knowledge discovery that is unbiased.

In addition, existence of structural metadata is an acknowledgement that information has pattern. However, rather than enabling the inherent structure in the information resource to express itself, digital records are marked up subjectively with tags that contaminate the authentic content. As an example (Box 1), on the website that is administrated by the Office of the Law Revision Counsel for the United States Congress (http://uscode.house.gov/usc.htm), the *United States Code* has 50 Titles with 56,847 constituent elements or granules that have an average of 5.98 ± 0.48 SGML tags per granule.

BOX 1

EXAMPLE OF STANDARD GENERALIZED MARKUP LANGUAGE (SGML) TAGS, IN RED, INSERTED INTO THE *UNITED STATES CODE*

-CITE-
1 USC TITLE 1 - GENERAL PROVISIONS 01/02/01

-EXPCITE-
TITLE 1 - GENERAL PROVISIONS

-HEAD-
TITLE 1 - GENERAL PROVISIONS

-MISC1-
THIS TITLE WAS ENACTED BY ACT JULY 30, 1947, CH. 388, SEC. 1, 61 STAT. 633

Not only are the SGML tags a form of contamination, but they represent tangible costs to tag the digital record as well as to process and store the added bytes associated with the tags.

Moreover, with the United States Congress version of the *United States Code*, there are only 50 granules – one for each title – and internal markup is added continuously throughout each of these 50 granules to create the appearance of granularity. To integrate pieces of these 50 discrete granules would require copying and pasting the selected segments into a new digital record. With DIGIN™, as part of a contract with the National Archives and Records Administration, the same 50 Titles of the *United States Code* were automatically organized into 1,042,646 discrete information granules without any mark-up tags – increasing the actual granularity by more than six orders of magnitude (Fig. 6).

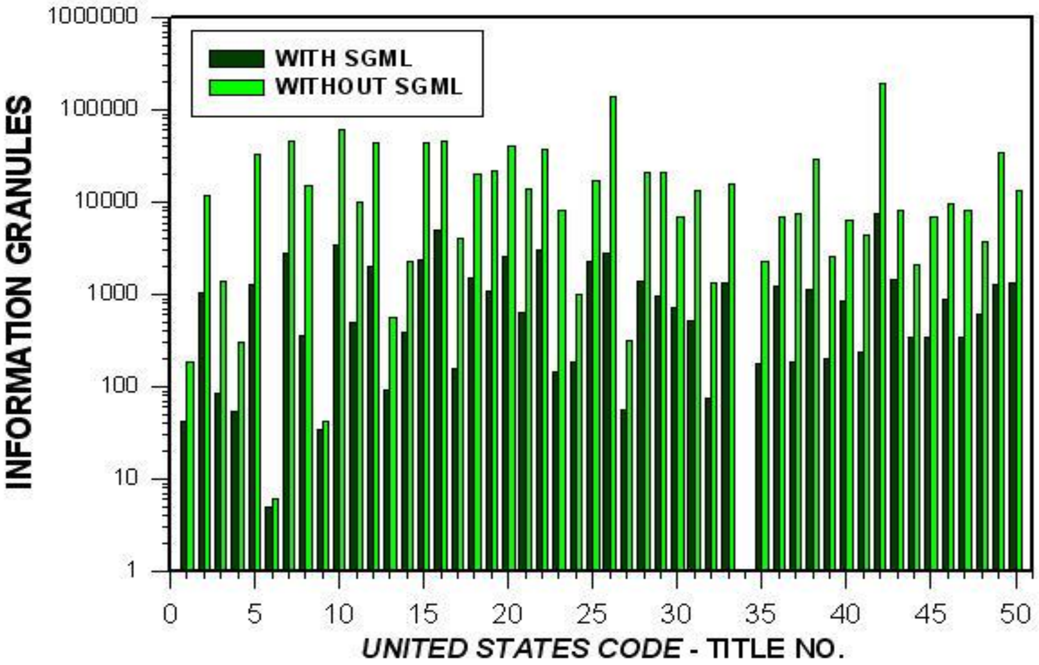


***FIGURE 6**: Granularity of the United States Code. As opposed to the 'apparent granules' produced by the United States Congress with markup tags (black), DIGIN™ automatically increased the number of 'actual granules' by nearly 2000% without imposing structural metadata (in green). From: (http://www.sdsc.edu/NARA/Publications.html).*

The fact is that structural metadata creates the illusion of granularity with segments of a digital record that are still connected. In contrast, DIGIN™ physically breaks a digital record into its constituent elements to generate actual granularity.

KNOWLEDGE DISCOVERY FROM DIGITAL RECORDS

Automated granularity itself is only part of the DIGIN™ advantage. The more meaningful utility is the DIGIN™ power of integration that translates directly into the discovery of relationships and new knowledge. Consider that two digital records have only one relationship, whereas three digital records have four intersections and four digital records have nine intersections (Fig. 7). Not only is the potential integration proportional to the number of granules, but the opportunity for knowledge discovery is geometrically related to the number of granules.

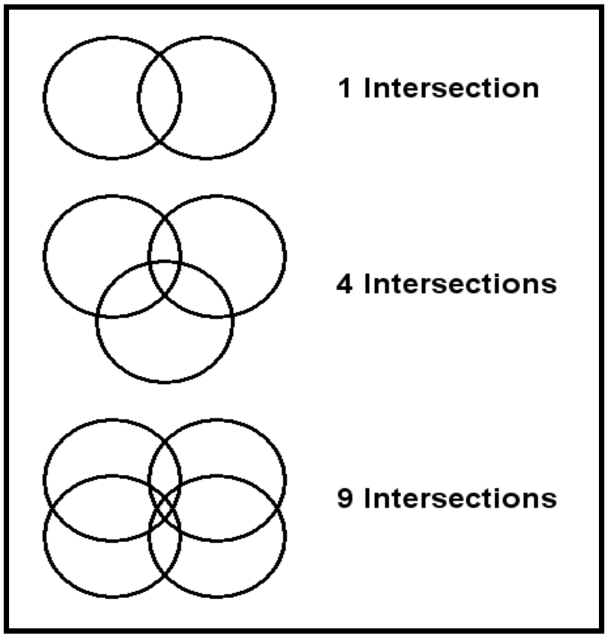


FIGURE 7: *As illustrated with 2, 3 and 4 granules, respectively - the potential integration (number of intersections) and opportunity for knowledge discovery is geometrically related to the number of information granules*.

With DIGIN™, each information granule is a unique digital record that can be integrated with any or all other digital records in a collection. For example, referring back to the earlier corporate conundrum about e-mail BLOBS, each e-mail already has information about the date and sender. Consequently, for any search query of the e-mail contents - in a single automated step - DIGIN™ could dynamically generate hierarchal displays that comprehensively and objectively identify relationships among the 120,000,000 messages based on their:

Date

Sender

E-Mail Contents

In contrast to DIGIN™, conventional technologies would require separate metadata for each e-mail message.

The problem is that metadata does not scale – which is the central bottleneck problem and reason for the existence of "unstructured" information (Fig. 1). As the granularity increases, the content volume of every granule decreases. Nonetheless, independent of granule size, the content volume of the metadata remains nearly constant. Consequently, after a few generations of increasing the granularity, the content volume of metadata exponentially expands relative to the content volume of the actual data (Fig. 8).

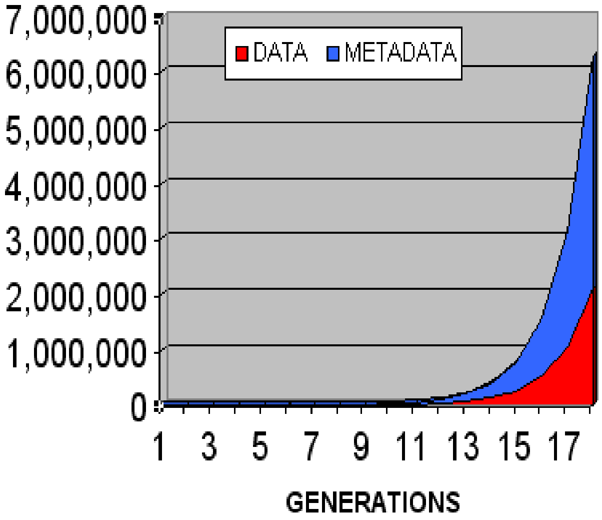
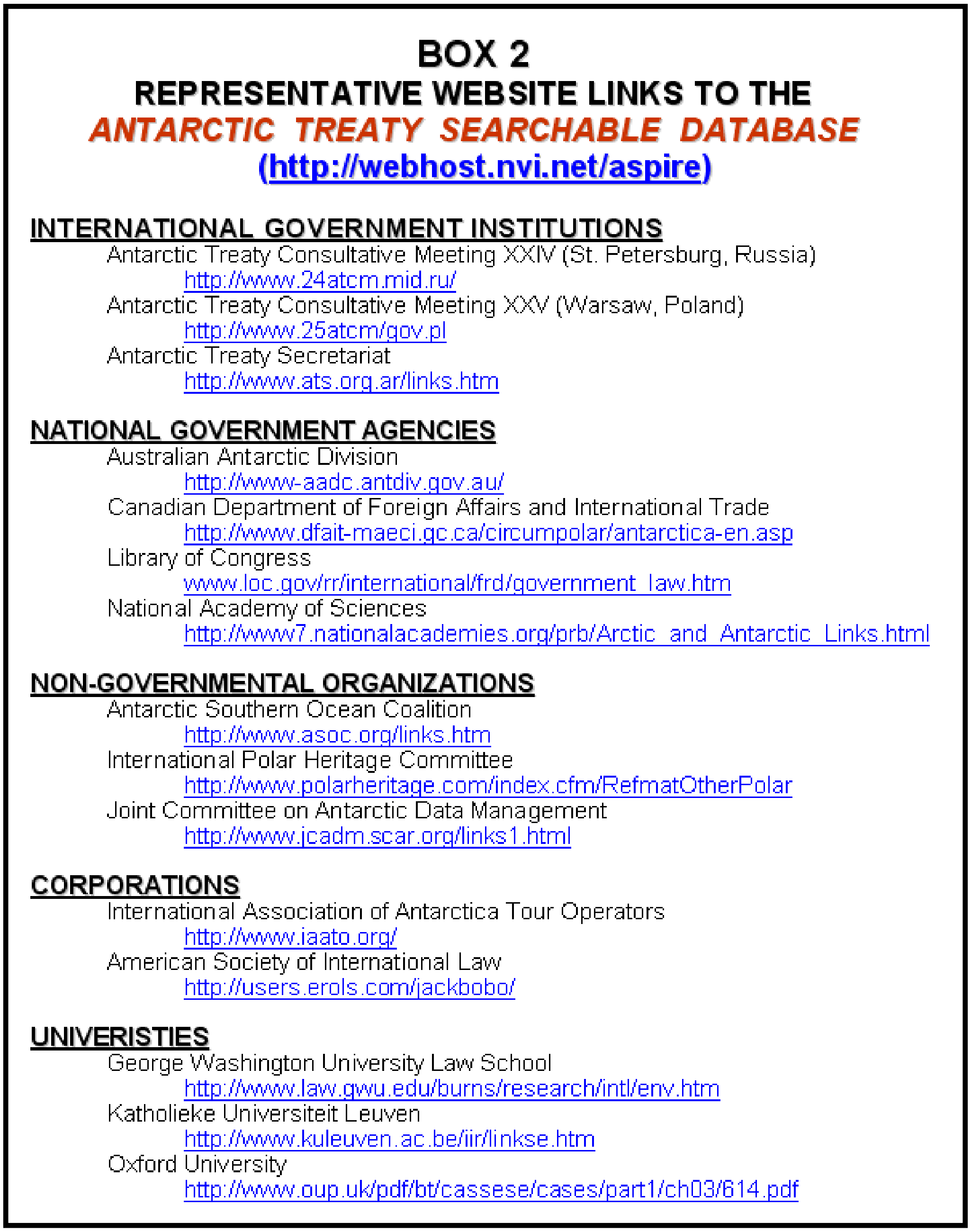


FIGURE 8: *A model of the exponentially increasing volume of metadata and/or mark-up tags by simply doubling the number of information granules (data).*

Moreover, in the same manner as markup tags, metadata becomes increasingly expensive as the granularity is increased because of the added effort and costs to store and process the additional information. Because the metadata resides in repositories that are separate from the actual digital records, there also is a risk of loosing the connection between the metadata and the digital records. Not only is metadata costly and static as well as redundant for text – but if it is ever decoupled from the digital records, those records will be effectively lost.

To further illustrate the advantages of the DIGIN™ technologies, consider the *Antarctic Treaty Searchable Database* (http://webhost.nvi.net/aspire) that is now utilized by all 45 nations in the Antarctic Treaty System as well as linked to websites of diverse institutions around the world (Box 2). The *Antarctic Treaty Searchable Database* was generated initially in 2000 with information in the *Antarctic Treaty Handbook* that was provided by the United States Department of State. Subsequently, granules have been added to the database as new measures have been adopted by the Antarctic Treaty nations. Each of these granules contains a categorical with genetic information about when and where it originated in the context of the overall history of the Antarctic Treaty System since 1959.



BOX 2
REPRESENTATIVE WEBSITE LINKS TO THE
ANTARCTIC TREATY SEARCHABLE DATABASE
(http://webhost.nvi.net/aspire)

INTERNATIONAL GOVERNMENT INSTITUTIONS

Antarctic Treaty Consultative Meeting XXIV (St. Petersburg, Russia)
http://www.24atcm.mid.ru/
Antarctic Treaty Consultative Meeting XXV (Warsaw, Poland)
http://www.25atcm/gov.pl
Antarctic Treaty Secretariat
http://www.ats.org.ar/links.htm

NATIONAL GOVERNMENT AGENCIES

Australian Antarctic Division
http://www-aadc.antdiv.gov.au/
Canadian Department of Foreign Affairs and International Trade
http://www.dfait-maeci.gc.ca/circumpolar/antarctica-en.asp
Library of Congress
www.loc.gov/rr/international/frd/government_law.htm
National Academy of Sciences
http://www7.nationalacademies.org/prb/Arctic_and_Antarctic_Links.html

NON-GOVERNMENTAL ORGANIZATIONS

Antarctic Southern Ocean Coalition
http://www.asoc.org/links.htm
International Polar Heritage Committee
http://www.polarheritage.com/index.cfm/RefmatOtherPolar
Joint Committee on Antarctic Data Management
http://www.jcadm.scar.org/links1.html

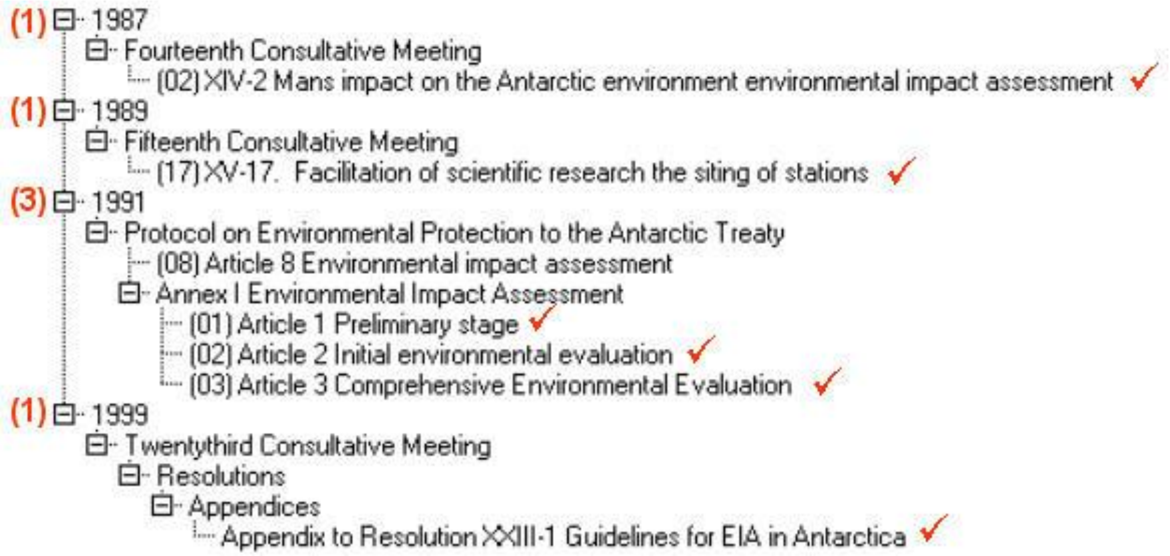
CORPORATIONS

International Association of Antarctica Tour Operators
http://www.iaato.org/
American Society of International Law
http://users.erols.com/jackbobo/

UNIVERISTIES

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
Oxford University
http://www.oup.uk/pdf/bt/cassese/cases/part1/ch03/614.pdf

In stark contrast to conventional lists that are generated with search engines, DIGIN™ dynamically generates hierarchal displays that comprehensively identify relationships among the granules (Fig. 9). With DIGIN™, it now becomes possible to automatically, comprehensively and objectively discover knowledge from digital collections.



```
(1) ⊟ 1987
      ⊟ Fourteenth Consultative Meeting
          (02) XIV-2 Mans impact on the Antarctic environment environmental impact assessment ✓
(1) ⊟ 1989
      ⊟ Fifteenth Consultative Meeting
          (17) XV-17. Facilitation of scientific research the siting of stations ✓
(3) ⊟ 1991
      ⊟ Protocol on Environmental Protection to the Antarctic Treaty
          (08) Article 8 Environmental impact assessment
        ⊟ Annex I Environmental Impact Assessment
            (01) Article 1 Preliminary stage ✓
            (02) Article 2 Initial environmental evaluation ✓
            (03) Article 3 Comprehensive Environmental Evaluation ✓
(1) ⊟ 1999
      ⊟ Twentythird Consultative Meeting
        ⊟ Resolutions
          ⊟ Appendices
              Appendix to Resolution XXIII-1 Guidelines for EIA in Antarctica ✓
```

FIGURE 9: *Hierarchal display that was dynamically generated from the Antarctic Treaty Searchable Database (http://webhost.nvi.net/aspire) for the search term: "minor or transitory." The red numbers and checks illustrate occurrences of the search term during any given year.*

In the world today, we are looking for actual relationships that facilitate meaningful decision making, which means that we need quantitative information that can be used to statistically and graphically identify trends. Because the DIGIN™ technologies comprehensively generate objective relationships (Fig. 9), it becomes possible to extract quantitative information for further interpretation as illustrated in Figures 10 and 11.

Knowledge discovery in our digital world requires the ability to integrate information at scale depending on the objectives of each individual user. Lists hide relationships within and between information resources. Moreover, the underlying metadata (Table 1) that generate the lists are neither objective nor comprehensive in terms of the original authentic content of the digital records. The "paper paradigm" limitation of these conventional approaches is their focus on content rather than structure as the primary step in managing digital information resources. Moreover, without utilizing the structure, it will never be possible to automate the granularity that is necessary to integrate information (Fig. 7), especially at the rate it is being produced in our information society (Fig. 2).

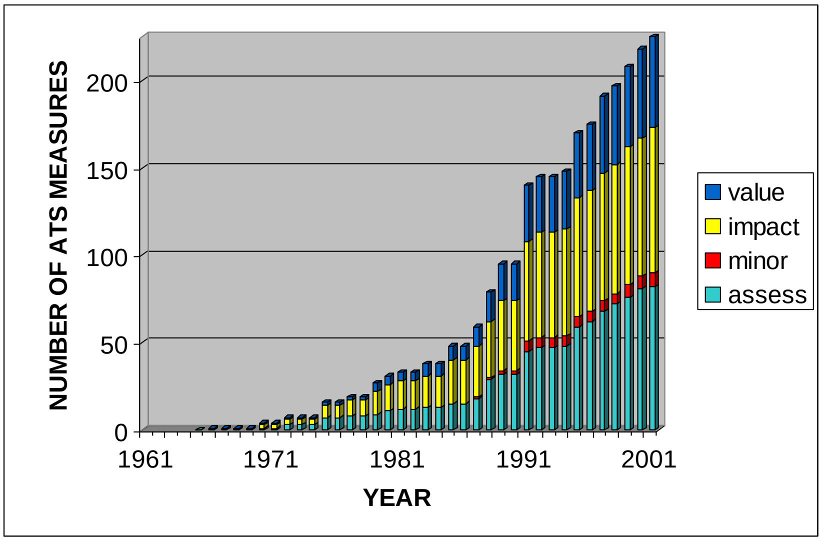


FIGURE 10: *Graphical analysis of quantitative information that was objectively extracted from the hierarchal displays in the Antarctic Treaty Searchable Database (e.g., Fig. 9) for the above search terms.*

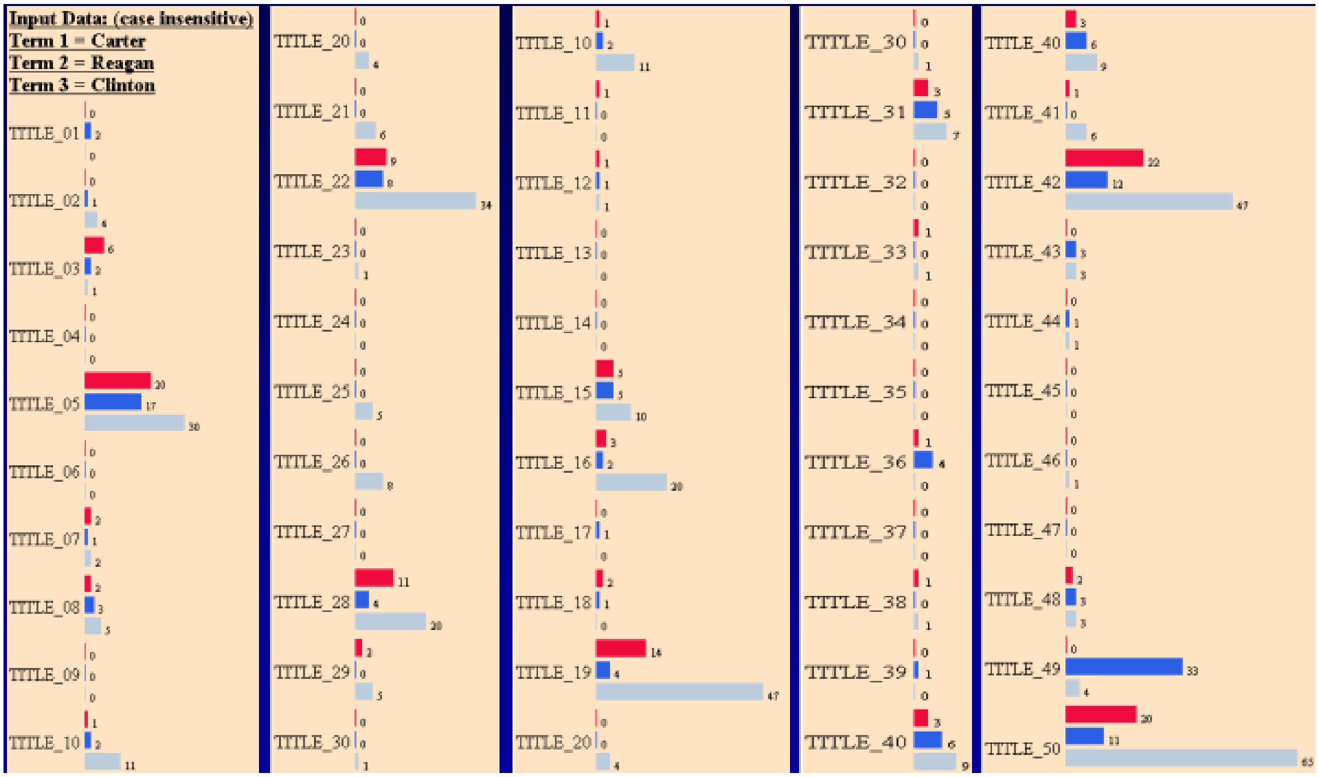


FIGURE 11: *DIGIN™ analyses of the United States Code showing the frequencies of legal documents that were signed by Presidents Carter (red), Reagan (dark blue) and Clinton (light blue) in the 1970's, 1980's and 1990's, respectively, in each of the 50 Titles.*

DIGIN™ TO DISCOVER KNOWLEDGE

Information is an asset that has value from individuals to governments (Fig. 12). For example, how can a company be certain that it is complying with all government regulations if it doesn't know what supporting evidence (and, perhaps, potential exposure) exists within its e-mail collections? How can a business accurately project business and market trends when it doesn't have a clue as to the presence or whereabouts of any major indicators that may lie hidden within its corporate BLOB collection? What is the cost of not being able to identify an information relationship that would be of strategic importance to a company?

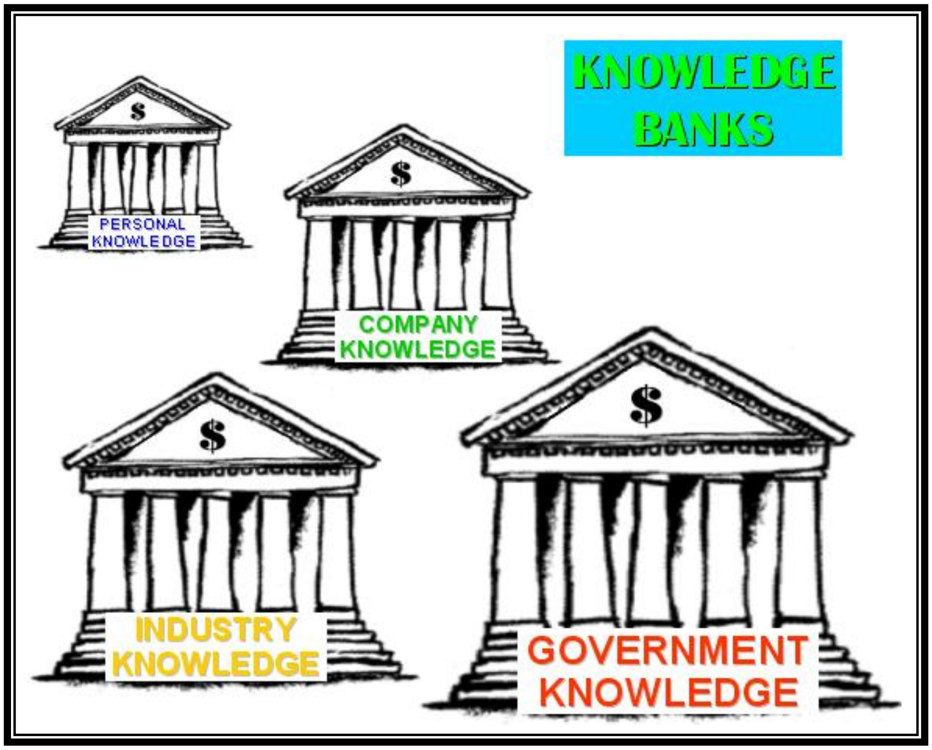


FIGURE 12: *Knowledge banks that involve the deposit, withdrawal and preservation of intellectual capital that is relevant to individuals, companies, industries and the government.*

In our information society, programmers conventionally control the architecture, access and manipulation of our digital information assets. However, at all levels – in personal, company, industry and government knowledge banks – there are five arenas of decisions that the subject experts should be able to control to manage their digital information (Box 3).

BOX 3
SUBJECT EXPERT CAPACITY
TO MANAGE DIGITAL RECORD COLLECTIONS

1. **What are the components of the digital collection?**
2. **What is the appropriate level of granularity to discover meaningful relationships in the digital collection?**
3. **What are the search queries that will be imposed on the digital collection?**
4. **How will the different tiers in the hierarchal displays be organized?**
5. **What are the appropriate statistical and graphical analyses to impose on the quantitative results from a search?**

DIGIN™ enables users, each of which is a subject expert, to mix and match information from digital resources in novel directions that they define depending on their answers to the questions in Box 3. The principal benefit of DIGIN™ is to open doors for comprehensive knowledge discovery that can be accomplished automatically based on user-defined objectives and criteria.

CONCLUSION

All signs point to radical changes in the technologies that facilitate knowledge discovery from digital records (e.g. Figs. 4, 6, 8). This is not a prediction as much as it is recognition of two marketplace forces that will compel a paradigm shift in how we organize and utilize digital records. Those forces are:

- The inexorable growth of worldwide information volume (Fig. 2); and

- The need of business and government to be able to utilize 100% of the available digital information.

The future of knowledge management involves the structure as well as the content of digital collections. Conventional solutions focus on content alone to deliver qualitative interpretations along with the illusion of granularity. However, scale is overwhelming the capabilities of metadata, mark-up languages and even databases as the volumes of structured and unstructured information explodes beyond the manpower and capabilities of these conventional technologies.

To generate actual granularity, DIGIN™ utilizes the inherent structure and patterns in information resources to automatically break digital records into logical information units. This automated granularity, which facilitates integration and enables digital content to express itself objectively, underlies the discovery of relationships within and between information resources because (Fig. 6) because:

granularity is directly proportional to the potential integration of information.

By creating the granularity and capacity to integrate digital information, there are a host of applications that become available for organizations to enhance functionality of their digital information resources in directions that influence their bottom line. The interoperability of DIGIN™ alone enables organizations (Fig. 12) to utilize this powerful integration technology in a seamless manner without interrupting standard operating procedures. There also are tangible cost savings because of reduced effort to integrate information without adding to the content of the original digital records, which further saves on storage costs and processing time. Ultimately, the benefits of DIGIN™ are that it is an automated-modular-interoperable technology that:

- ✓ **integrates digital information at scale independent of format or source;**
- ✓ **provides rapid, comprehensive and objective access to all digital information based on user-defined criteria;**
- ✓ **preserves context of information;**
- ✓ **generates expandable-collapsible hierarchies to describe relationships within and between information resources;**
- ✓ **identifies anomalies in structure and content;**
- ✓ **aggregates query results to generate new digital records; and,**
- ✓ **provides quantitative analysis of query results.**

For comprehensive integration to be possible, there must be a granularity capability that is virtually untapped today by all conventional knowledge discovery technologies. DIGIN™ offers such a paradigm shift by focusing on the structure of digital records to open doors for knowledge discovery throughout our world information society.

The DIGIN™ architecture provides an unprecedented level of user-defined control to integrate and discover knowledge from digital information independent of content, media, source or scale. In working with digital records, the richness of the electronic world is a lot broader than anyone has yet leveraged. The real knowledge management opportunity is to discover relationships among digital records and be surprised by the insights.