# 'UNSTRUCTURED DATA' PRACTICES IN POLAR INSTITUTIONS AND NETWORKS: A CASE STUDY WITH THE ARCTIC OPTIONS PROJECT

***Paul Arthur Berkman*[1,2 *]**

[1] *Bren School of Environmental Science and Management / Marine Science Institute, University of California, Santa Barbara, CA 93106, USA*
**Email:* berkman@bren.ucsb.edu
[2] *DigIn (Digital Integration Technology Limited), 6 Caxton House, Broad Street, Great Cambourne, Cambridge, CB23 6JN, UK*
*Email: paul.berkman@digin.co*

## *ABSTRACT*

*Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability is a new three-year research project to assess future infrastructure associated with the Arctic Ocean regarding: (1) natural and living environment; (2) built environment; (3) natural resource development; and (4) governance. For the assessments, Arctic Options will generate objective relational schema from numeric data as well as textual data. This paper will focus on the 'long tail of smaller, heterogeneous, and often unstructured datasets' that 'usually receive minimal data management consideration', as observed in the 2013 Communiqué from the International Forum on Polar Data Activities in Global Data Systems.*

**Keywords:** Big Data, Unstructured data, Relational schema, Infrastructure, Integration

## 1 INTRODUCTION

Interests are awakening globally to take advantage of extensive energy, shipping, fishing, and tourism opportunities associated with diminishing sea ice in the Arctic Ocean (Berkman & Vylegzhanin, 2013). Because of these diverse interests, there is urgency to develop infrastructure so the commercial activities can proceed in a sustainable manner, balancing:

- National interests and common interests
- Environmental protection, social equity, and economic prosperity
- Needs of present and future generations

Infrastructure in the Arctic Ocean will include port facilities, sea lanes, emergency response assets, communication systems, and observing networks as well as regulatory and policy systems. Cross-cutting all aspects of the infrastructure will be information management and knowledge discovery using disparate data. The *Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability* (Arctic Options, 2014) project, which is being funded by the National Science Foundation in the United States and Centre Nationale de la Recherche Scientifique in France from 2013–2016, provides a case study for international, interdisciplinary, and inclusive data practices.

As part of the *Arctic Science, Engineering and Education for Sustainability* programme (ArcSEES, 2012), *Arctic Options* will consider data for:

1. Natural and living environment
2. Built environment
3. Natural resource development
4. Governance

To enhance its cost-effectiveness, *Arctic Options* also has established links to the *Study of Environmental Arctic Change* (SEARCH, 2013) and *Arctic Climate Change, Economy and Society* (ACCESS, 2013) projects that are supported extensively within the United States and Europe, respectively.

These data for Arctic Ocean infrastructure will be generated from sensor and transactional systems from observing networks, as well as experiments, in numeric formats that are ***structured*** (i.e., managed) with databases, which can be analyzed statistically and graphically with various relational approaches. Geographic

Information Systems (GIS) will be particularly powerful for marine spatial planning, ecosystem-based management and integrated ocean management in the Arctic Ocean (Håkon Hoel, 2010; Ehler, 2011; Clement Bengston, & Kelly, 2013; PAME, 2013).

In addition, the data in the *Arctic Options* project will involve digital resources in natural language formats (e.g., papers, reports, and agreements), which commonly are considered to be ***unstructured*** (i.e., unmanaged) because they *cannot be decomposed into standard components* or relational schema (Oracle, 2002). The unstructured data will be aggregated from diverse institutions that have Arctic remits (Berkman & Vylegzhanin, 2013), such as the Arctic Council (2013).

Within the popular framework of 'Big Data' (Lohr, 2012), structured and unstructured data together reflect the full complement of digital information that we produce as a global society, with the volume of unstructured data accounting for upwards of 85% of the information and growing twice as fast as structured data (Figure 1). In this paper, innovations with unstructured data will expand on earlier developments through the National Science Digital Library (NSDL, 2013) and International Research on Permanent Authentic Records in Electronic Systems project (InterPARES, 2013), as summarized in a publication through the Committee on Data for Science and Technology (Berkman, Morgan, Moore, & Hamidzadeh, 2006).

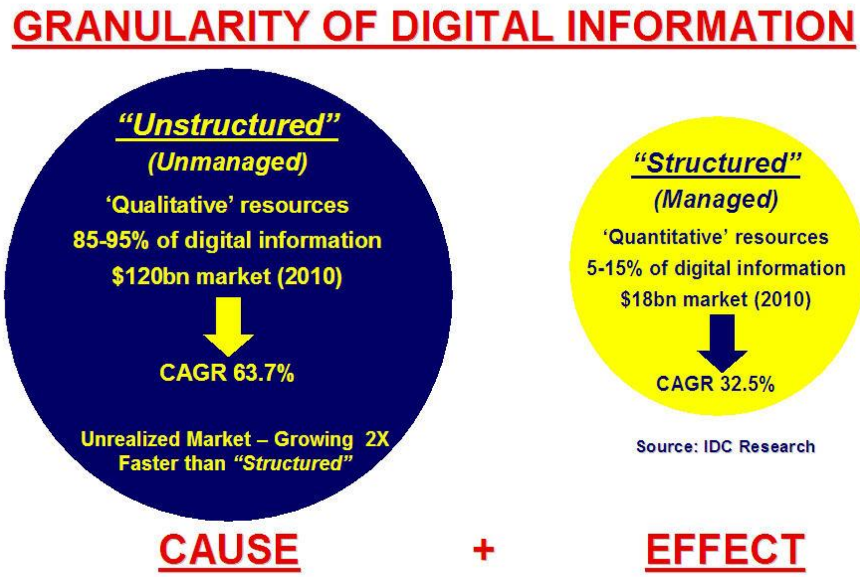


**Figure 1.** Big Data defined in terms of 'structured' and 'unstructured' data, both of which relate to granularity of the information resources. Compound annual growth rates (CAGR) and estimated market sizes are from Gantz & Reinsel (2010).

This paper also will focus on manipulation of unstructured data in view of the following observation in the *Communiqué* from the *International Forum on Polar Data Activities in Global Data Systems* (Polar Data Forum, 2013):

> *It is the long tail of smaller, heterogeneous and often unstructured datasets (those without metadata, mark-up and not in databases) that receive the least data management attention by scientific repositories. However, utilizing the inherent structure of any digital resources provides an objective framework to discover their relationships in a manner that complements existing content and context management solutions.*

In particular, this paper is intended to provoke discussion about applying the inherent structure of digital information, which is a fundamental opportunity with digital information and a distinction compared with all hardcopy resources.

## 2 ELEMENTS OF MEANING

Each era of global communication, from stone to digital (Berkman, 2008), has been accompanied by a threshold increase in human capacity to transport information. Similarly, each new communication medium has significantly increased our capacity to produce information, as indicated by the relative volumes of information that emerged. Moreover, the ability to integrate information has increased over time with tablets, folios, books, and now websites. In contrast, the most resilient medium was stone with petroglyphs and pictographs that have stood the test of time through rain, snow, wind, and even fire. Subsequent media have been much more fragile. In fact, the digital medium has been like a black hole, where most of the information produced has been lost because of limited preservation strategies and rapid obsolescence of storage devices.

Looking backward through time, information in our civilization has been managed largely through libraries and archives. While similar in their needs to facilitate information access and preservation, these two architectures possess fundamental differences. Archives manage information based on the ***context*** of records linked to specific activities and transactions, such as the housing authority that records the title of your home. Libraries largely manage information based on the ***content*** of the information resources, as with the subject categories in the Dewey Decimal System (OCLC, 2013).

Beyond content and context, the third element of information to establish meaning is its ***structure***. For example, when a message is encrypted (i.e., the structure is altered), it still has content and context but no meaning absent the key to unlock the encryption. Alternatively, 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. Similarly, meaning will be compromised by removing the context that can be used to authenticate an information resource or establish its provenance.

All information must have content, context, and structure to create meaning (Figure 2). However, with the digital medium, it is possible to utilize the *content* and *context*, as well as *structural* patterns, to manage sets, subsets, and supersets of information resources. **The capacity to utilize the inherent structure of digital resources is the distinguishing feature of digital information compared to all of its hardcopy predecessors.**

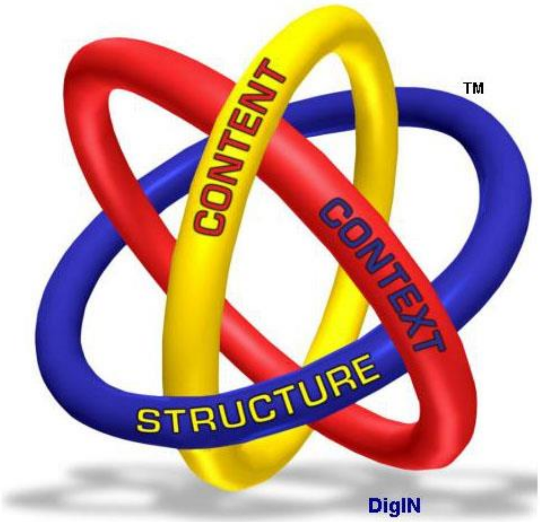


**Figure 2.** Borromean ring, illustrating the interconnected core elements of all information, both hardcopy and digital, that together create meaning (revised from Berkman, 2008)

## 3 DIGITAL INFORMATION ARCHITECTURES

In general, unstructured data are managed with metadata, markup, or databases. However, for text resources specifically, the digital resource itself contains the information content that would be summarized by metadata. Consequently, metadata incompletely and subjectively characterize digital text resources for the purposes of information access, as card catalogues did with hardcopy resources (OCLC, 2013).

Nonetheless, ubiquitous use of metadata, which originated with card catalogues for hardcopy libraries (Dublin Core, 2003), has become a *de facto* approach for digital information management around the world with diverse ‘standards’ through the International Organization for Standardization (e.g., ISO, 2013). These standards vary by country and require extensive effort in terms of personnel expertise, time, and cost to implement.

Although not properly quantified, back-of-the-envelope calculations further suggest that metadata production may account for more than 10% of the global expenditure on information and communications technology. Most importantly, the production of metadata does not scale with increasing granularity (Berkman et al., 2006), which largely explains the growing discrepancy between the volumes of unstructured and structured digital information (Figure 1).

Data that is unmanaged with current technologies raises the question about strategies for information management and knowledge discovery with text resources. Given the global diversity of information technology companies, Table 1 was created to compare attributes and functions for information management and knowledge discovery (Table 2) across generalized solution suites that apply the content and context as well as the structure of digital text resources.

**Table 1.** Capacity to utilize various attributes and functions (Y(es) or N(o)) in relation to generalized solution suites for digital information management and knowledge discovery

| Attributes and Functions (see Table 2) | Interconnected Core Elements of Meaning (Figure 2) and Underlying Solution Suites for Digital Information Management and Knowledge Discovery | | | |
|---|---|---|---|---|
| | Content and Context | | | Structure |
| | Search Engines | Databases / Spreadsheets | Metadata / Ontologies / Semantic Indexes | Granularity Engines |
| Language Independent | N | Y | N | Y |
| File-type Independent | Y | N | N | Y |
| Scale Independent | Y | N | N | Y |
| Classification Independent | N | Y | N | Y |
| Ranking Independent | N | Y | Y | Y |
| Markup Independent | Y | Y | N | Y |
| Metadata Independent | N | Y | N | Y |
| Re-purpose Metadata Tags | N | N | N | Y |
| Tabular Manipulation | N | Y | N | Y |
| Result Lists | Y | Y | Y | Y |
| Relational Displays | N | Y | Y | Y |
| Relational Analytics | N | Y | Y | Y |
| Preserves Authentic Record | Y | N | N | Y |
| Content-driven | Y | N | Y | Y |
| Context-driven | N | Y | Y | Y |
| Structure-driven | N | Y | N | Y |
| User-defined Rules | N | Y | Y | Y |
| Single-level Inverted Indexing | Y | Y | Y | Y |
| Multi-level Inverted Indexing | N | N | N | Y |
| $2^N$ Permutations | N | N | N | Y |
| Result-set Certainty | N | N | N | Y |
| Automated Granularity | N | N | N | Y |

Among the solutions in Table 1, search engines, databases, and spreadsheets are well known and need no elaboration. Similarly, ontologies and semantic indices are widely used (Berners-Lee, Hendler, & Lassila, 2001). The concept of a 'granularity engine', however, is being introduced herein as a fundamental solution based on the inherent structure of digital resources. In particular, granularity engines can leverage the inherent structure of digital resources to achieve functionalities beyond what is possible with solutions derived from their content or context (ie., described by the full complement of attributes and functions with 'Y' in Table 1).

Most notably, granularity engines can objectively deliver $2^N$ relationships among $N$ digital objects, overcoming a significant shortcoming with subjective content or context solutions that limit the range of relationships that can be discovered. Consider two digital objects where the four possible permutations include one, the other, both, or neither. If there are just one hundred digital objects, which is a small number, the number of possible permutations ($2^{100}$) effectively would be a googol, and we have the challenge to discover relationships across thousands and millions of digital objects.

**Table 2.** Descriptions of attributes and functions in Table 1

| Attribute or Function | Description |
|---|---|
| Language Independent | Operations not limited by symbolisms, such as different alphabets |
| File-type Independent | Operations not limited by types of digital resources, recognizing that all resources (e.g., text, images, genomes, sensor data streams, and music) have structural patterns of embedded organization that can be manipulated automatically |
| Scale Independent | Operations not limited by size of resource set, which can be applied by an individual, small business, large corporation, or government |
| Classification Independent | Operations not limited by subjective categories that are defined by individuals or programmed algorithms |
| Ranking Independent | Lists generated without programmer algorithms that arbitrarily rank relevant search results |
| Markup Independent | Navigation not limited by markup tagging, which is subjective |
| Metadata Independent | Access not limited by metadata schema because all symbols in a resource set are indexed and searchable at all levels of embedded granularity |
| Repurpose Metadata Tags | Applies existing metadata fields associated with a digital resource to manipulate the embedded levels of organization in expandable-collapsible hierarchies |
| Tabular Manipulation | Capability to combine rows, columns, and cells from multiple tables |
| Result Lists | Capability to generate linear displays of search results |
| Relational Displays | Capability to generate integrated displays of search results |
| Relational Analytics; | Capability to generate statistical displays of integrated search results |
| Preserves Authentic Record | Capability to preserve authentic digital resources, while at the same time providing the ability to search or integrate the digital resources |
| Content-driven | Based on information classification schema that are subjectively defined by individuals or programmed algorithms |
| Context-driven | Based on information provenances that are subjectively defined by individuals or programmed algorithms |
| Structure-driven | Based on information boundaries and patterns that can be manipulated within and between embedded levels of organization in digital resources |
| User-defined Rules | Contrasted with programmer-constrained rules |
| Single-level Inverted Indexing | One-to-many relationships among digital objects that is used to generate lists |
| Multi-level (Dynamic) Inverted Indexing | Many-to-many relationships within and between digital objects that is used to generate expandable-collapsible hierarchies |
| $2^N$ Permutations | Comprehensive capacity to integrate $N$ digital objects and discover all possible combinations of granules within and between digital resources |
| Result-set Certainty | Objective and comprehensive results in contrast to probabilistic solutions, which have inherent uncertainties |
| Automated Granularity | Generate subsets of embedded parent–child relationships down to finite elements at the lowest levels of granularity within and between digital resources |

## 4 INHERENT STRUCTURE AND GRANULARITY ENGINES

Effectively, we all have infinite and instantaneous access to digital data on our computers or networks and over the internet. With text resources, searching through repositories merely lists items that contain the search query. Lists generally are ranked in an arbitrary manner, commonly in terms of assumed relevance. From lists of possibly relevant results, digital resources can be selected, and it is then up to the user to hunt sequentially for the search term through each resource. If the user wants to identify content-in-context relationships within and between the resources (e.g., relevant sentences within chapters or resources within years), it is then necessary to: (a) cut the relevant pieces out of each relevant resource; (b) paste them into a new folder; and then (c) organize all of the cut-and-paste pieces. This a–b–c process to establish relationships within and between digital resources is tedious, time consuming, and subjective.

Text resources however have structure that is defined by the grammar rules of the language. For example, read left–right and top–bottom in English, books have chapters that have pages with paragraphs embedded with sentences composed of words that each contain letters. Through each book, various headings and forms of punctuation (e.g., full stop, exclamation point, or question mark at the end of a sentence) define boundaries that can be used to disaggregate embedded levels of granularity, like peeling an onion. Unlike hardcopy resources, repeating structural patterns in digital resources can be set as rules (Berkman et al., 2006) to run a granularity engine that generates and indexes discrete granules (e.g., sentences, paragraphs, pages, or chapters). Subsequently, these granules can be integrated in parent–child contexts across a collection of digital resources for any search query. Such management, discovery, and analysis of digital text documents with a granularity engine will complement the GIS manipulations of numeric data layers in the *Arctic Options* project.

A classic form of an unstructured resource is a PDF (portal document format) file, which has the advantages of being interoperable across diverse operating systems and file formats as well as serving as an archival standard (ISO, 2009). Utilizing a familiar collection to illustrate the application of a granularity engine, 53 PDF files of books written by Charles Dickens from 1836–1880 were automatically decomposed into 571,386 granules that represent all sentences, paragraphs, pages, and chapters within these books and years (Figure 3).


**Figure 3.** Granularity engine implementation with 53 PDF files of books written by Charles Dickens from 1836–1880, which were automatically decomposed by PDF KnoHow™ (DigIn, 2013) into 571,386 granules that represent all sentences, paragraphs, pages, and chapters within these books and years. The exact match

search for ‘best of times’ reveals occurrence of this famous phrase in 15 years, 16 books, 15 chapters, 18 pages, 18 paragraphs, and 18 sentences that can be expanded and collapsed with the Digital Zoom™. The relevant granules can be aggregated and further analysed to quantify parent–child frequencies comprehensively at all granularity levels in the expandable–collapsible hierarchy.

In the *Arctic Options* project, complementing the geospatial integration of data layers with GIS, granularity-engine applications will enable users to objectively zoom in and out of content layers across collections of digital text resources for any Boolean search query. As an example, to discover relationships across the entire Dickens collection, searching for ‘best of times’ in Tale of Two Cities surprisingly reveals that this famous phrase was repeated within Dickens' books throughout his career (Figure 3). Such surprises, which are at the heart of discovery, can be generated by granularity engines based on the inherent structure of digital resources without metadata, markup, or databases (Table 1).

## 5 CONCLUSIONS

There is no such thing as ‘unstructured’ data because all data must have structure as well as content and context to have meaning (Figure 2). Moreover, granularity-engine applications with PDF files (Figure 3) falsify long-standing definitions for unstructured data (e.g., Oracle, 2002) because they can be automatically *decomposed into standard components* as well as relational schema without metadata, markup, or databases. The unique advantage of digital resources over hardcopy resources is the opportunity to utilize the inherent structure of the resources as well as their content and context, for the purposes of information management and knowledge discovery.

## 6 ACKNOWLEDGEMENTS

This paper was presented as invited keynote at the International Forum on Polar Data Activities in Global Data Systems (Polar Data Forum, 2013) and has been prepared with support from *Arctic Options: Holistic Integration for Arctic Coastal-Marine Sustainability* project, supported by the National Science Foundation (NSF-PLR No. -1263819).

## 7 REFERENCES

ACCESS (2013) *Arctic Climate Change, Economy and Society.* Retrieved December 15, 2013 from the World Wide Web: http://www.access-eu.org

ArcSEES (2012) *Arctic Science, Engineering and Education for Sustainability*. Retrieved September 12, 2013 from the World Wide Web: http://nsf.gov/pubs/2012/nsf12553/nsf12553.htm

Arctic Council (2013) Retrieved January 28, 2014 from the World Wide Web: http://www.arctic-council.org

Arctic Options (2014) Retrieved October 10, 2014 from the World Wide Web: http://www.arctiptions.org

Berkman, P.A. (2008) Once in a hundred generations. In Halbert, M., & Skinner, K. (Eds.), *Strategies for Sustaining Digital Libraries*, Atlanta, Georgia: Emory University.

Berkman, P.A., & Vylegzhanin, A.N. (2013) *Environmental Security in the Arctic Ocean*, Dordrecht, Netherlands: Springer.

Berkman, P.A., Morgan, G.J., Moore, R., & Hamidzadeh, B. (2006) Automated Granularity to Integrate Digital Information: The “Antarctic Treaty Searchable Database” Case Study. *Data Science Journal 5*, pp 84–99.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001) The Semantic Web. *Scientific American* May 2001, pp 29–37.

Clement, J.P., Bengston, J.L., & Kelly, B.P. (2013) *Managing for the Future in a Rapidly Changing Arctic. A Report to the President*, Washington, D.C: Interagency Working Group on Coordination of Domestic Energy Development and Permitting in Alaska.

DigIn (2013) Dickens Christmas Present. Retrieved December 15, 2013 from the World Wide Web: http://dickens.knohow.co

Dublin Core (2003) Retrieved December 15, 2013 from the World Wide Web: http://dublincore.org/resources/faq/

Ehler, C (2011) Part II. Marine Spatial Planning in the Arctic: A first step toward ecosystem-based management. In *The Shared Future. A Report of the Aspen Institute Commission on Arctic Climate Change,* Washington, D.C.: The Aspen Institute, pp 40-82.

Gantz, J. & Reinsel, D. (2010) *A Digital Universe Decade – Are You Ready?*, Framingham, Massachusetts: IDC Corporation.

Håkon Hoel, A. (2010) Integrated Oceans Management in the Arctic: Norway and Beyond. *Arctic Review on Law and Politics 1*, pp 186–206.

InterPARES (2013) Retrieved December 15, 2013 from the World Wide Web: http://interpares.org

ISO (2009) *ISO 19005-1:2005 Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1)*. Retrieved January 29, 2014 from the World Wide Web: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

ISO (2013) *ISO/IEC JTC1 SC32 WG2 Development/Maintenance: ISO/IEC 11179, Information Technology—Metadata registries (MDR)*. Retrieved December 15, 2013 from the World Wide Web: http://metadata-standards.org/11179

Lohr, S. (2012) The Age of Big Data. *New York Times,* 11 February 2012.

NSDL (2013). Retrieved December 15, 2013 from the World Wide Web: http://www.nsdl.org

OCLC (2013) *Dewey Decimal Classification*. Retrieved January 29, 2014 from the World Wide Web: http://www.oclc.org/dewey.en.html

Oracle (2002) *Oracle9i Application Developer's Guide—Large Objects (LOBs). Release 2 (9.2)*. Retrieved December 15, 2013 from the World Wide Web: http://docs.oracle.com/cd/B10501_01/appdev.920/a96591/adl01int.htm

PAME (2013) Arctic Ocean Review. Final Report. Phase II: 2011–2013, Akureyri: Arctic Council Working Group on the Protection of the Arctic Marine Ecosystem (PAME).

Polar Data Forum (2013) *International Forum on Polar Data Activities in Global Data Systems Communiqué: Recommendations & Observations Arising From the 'International Polar Data Forum' 15–16 October 2013, Tokyo (Japan)*, Tokyo: ICSU World Data System.

SEARCH (2013) Retrieved December 15, 2013 from the World Wide Web: http://www.arcus.org/search

(Article history:Available online 10 Octorber 2014)