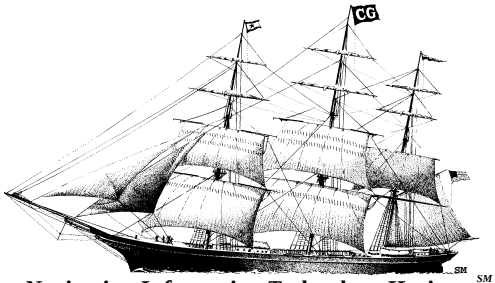
Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE FOR THE INFORMATION TECHNOLOGY BUYER



Navigating Information Technology Horizons℠

New in 2007 Report #TCG2007089 Updated September 14, 2007

Reference Data Concepts — What IT Folks Should Learn from Libraries and Archives

Analyst: Anne MacFarland

Management Summary

Business places great value on experience. It gives us the knowledge of what not to do that makes strategies effective and brings products to market with the right features, at the right time, at the right price. And, until recently, if we needed to refresh our memory of what happened over the years to some strategy, there were paper files to give us the overview that was hidden in the welter of day-to-day events. Digitizing business processes has given us huge productivity gains – but there is a trade-off. Because digitized information is generated by, and interpreted by, software applications, our access to that broad, historic view is stymied by application stovepipes. **We can search or browse within an application, but it is a lot of work to get a comprehensive view of the business experience of an enterprise.**

This need to use IT information for reference is not just a whim. As people take on new responsibilities, they need some sense of history to get a handle on the hidden issues they face. The most seasoned external consultants cannot match a company-specific, comprehensive, long-term overview of actual events. As people collaborate, there is a need to assure that documentation of the knowledge gained through that collaboration persist beyond the completion of the project or initiative. More pragmatically, there is a need to control the domain of fishing expeditions. Moreover, there is a need to do it with IT, because those systems are where the data is. This means managing the data as information and not just as a presentation or a part of a process. **It is time to look beyond IT to the information access systems that have been honed over centuries – libraries and archives. Both pertain to what IT is trying to do, in slightly different ways**.

Library reference materials, like Web pages, are purpose built for reference and bristle with helpful features like indexes. Archives, on the other hand, are records of an organization or a person that are deemed to have long-term value – very much like enterprise IT data. Archives are minimally processed to support occasional use. Both traditions focus on connecting requests from multiple patrons to the right information, first by amassing information sources, and then by filtering out what is not relevant, and brokering a connection to what remains. Few patrons look at one thousandth of the material in a library, and fewer still over the age of twelve set out to read entire fiction collections, which is how IT systems often look at data. **Setting up systems to filter as well as flood is what IT can learn from libraries and archives.** This is more than just a front-end capability. It comes from the way the system is organized and built.

In this bulletin, we will take the concept of provenance, on which the traditions of libraries and archives rest, and see how it can help bridge the gap between the old and new traditions of libraries and archives and IT. **Provenance is the history of a piece of information, both the *who, where, when, why* and *how* it was created, and how it has been transformed, through the filtering over time of selection, editing and transformations into its present incarnation.** Provenance can provide key elements needed to transform information for use as a reference resource. Read on for more details.

IN THIS ISSUE

So Alike – And Yet So Different

Many IT folks have been in a library and feel that they know what a library is all about. Hey, it's the quiet, the books, and that old Dewey Decimal System, not to mention the cozy chairs for reading and computers for Internet access, right? Yet these same people bridle when end users dismiss an application or technology in general based on a few bad experiences. **Most of the value of information access systems lies in decisions about taxonomy and metadata (or classification) elements that library patrons take as second nature and IT users don't see at all.**

Think of libraries and archives as the somewhat-estranged great granddaddies of information technology systems. Both the geezer and young-whippersnapper methods support use of information, and they manage it as a commodity – either by blocks, volumes, and files, or as books and pamphlets. But, the talk they talk is very different. It is only when you understand the difference in attitude that you can get beyond the cultural mismatch.

While IT, libraries and archives are all systems concerned with information preservation and access, there is a *cultural gap* in both vocabularies[1] and, more importantly, in approach**.** Generally, **libraries and archives address information control from the top down.** Think of it as a paternalistic approach. Reference collection development is a matter of surveying user population needs, and then populating broad knowledge segments with instances of knowledge (quality stuff, that – not merely information) in the form of dictionaries, encyclopedias, directories, manuals and other mostly-published, bound materials.

By contrast, **IT information systems work from foundations upwards, with the control of the information usually split between a gang of applications that proper management keeps from competing too vigorously for available resources.** IT just inherits the data as a byproduct of an application. Until recently, many in IT did not consider the data as a significant asset in its own right. Access to information independent of the generating application is a recent development and still not universal.

Libraries have been satisfying (and dissatisfying) patrons in more or less the same way for centuries, not just decades. They are filled with information produced to be used by some closely or loosely defined set of users. Whole industries have grown up producing material for reference[2]. Moreover, library science methodology is changed only by a usually very slow process of consensus. **IT moves at a very different pace**. IT systems started out to support massive amounts of calculations. Quickly, at least by library standards, their use for communication and content management has made them ubiquitous in the enterprise.[3]

Now, these enterprises are thinking about how to get more use out of the heaps of information they have amassed through their myriad of business applications, including e-mail. Libraries and archives have something to tell them.

To achieve what librarians call *intellectual control*[4] over a huge mass of data (as opposed to tracking a particular value in a particular field in a particular database), they define the scope of the inquiry carefully, dismiss all the information that is irrelevant, and concentrate their attention on what is left. While technology has given us prodigious search and indexing and classification capability, looking everywhere is usually constrained by privacy and confidentiality, as well as by time and money. **The librarian's methodology of using the *metadata* of cataloging and other finding aids to find the relevant and dismiss the irrelevant can be applied to using IT data for reference. It all starts with a traditional concept called *provenance.***

Provenance[5]

Libraries have most frequently used subject as the organizing criteria[6], and have spent a great deal of administrative time dealing with the ramifications (multiple subjects, new subjects, defining

[1] For example, the uses of the words *archive* and *cleanse* are egregiously disparate. In libraries, archives are cherished as unique and valuable. In IT, they are stores of data often considered "dead", but reluctantly kept just in case. Recently, this attitude has changed with the threats of Sarbanes-Oxley, and other laws. IT's cleansing refers to abnormalities in format and regularizing of names. Library and archival cleansing refers to lower life forms, mold and mildew, if you're lucky.

[2] Textbooks are usually not included in reference collections because of their limited target audience. Their content is a carefully honed subset of information rather than a comprehensive compendium.

[3] In an ironic twist, the spread of the use of computers coincided with (and precipitated) the demise of many company libraries in the late '70s and early'80s.

[4] This is knowing – in general – what the information says, in contrast to *bibliographic control*, which is being able to find it (more like a file system).

[5] No, not the place in France – or Rhode Island.

[6] There are as many approaches to classification, just as there are computing languages. Most of them are just as moribund.

Copyright © 2007 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

the nature of subject-to-subject relationships, and figuring out where to put material with no clear subject) of that choice.

Archives and government document repositories have taken the clearer path of assigning materials to collections by provenance. **Provenance, in brief, is the documentation (again, think *metadata*) about where the information came from.** In business records, provenance usually reflects the organizational structure. When that structure is reorganized, new groupings emerge. In archives, provenance is often the name of the person who amassed a body of information, like the president in each presidential library. Provenance is something that rumors conspicuously lack, but all other forms of information have. In museums, provenance is the litany of ownership of an object, and a key to proving it genuine. For information, it is a key to knowing the contexts in which the information may be relevant.

As an example, a Scottish book in a French collection of cookbooks may have informational value both as a source of a recipe for haggis, and as an instance of how the recipe of haggis has changed in the last 50 years, and even for documentation that Scottish cuisine was represented in a French collection of cookbooks. This instance may seem farfetched, but a similar hierarchy of context is extremely important in research fields such as drug research. It is why *tiered taxonomies* are useful. Provenance gives a richer, more dimensional context than you get from a search result. **In statistical and other analyses, knowing the provenance of data is key to ensuring that the results are meaningful.**

In digital libraries, provenance takes a new twist. Digital collections used for reference have some kind of fingerprint to attest to the integrity and quality of the data. But, these days, that ability to annotate can accessorize the record with useful information.[7] If the data is exported and annotated, it comes in as a new record – but the aggregate of annotations become a useful enhancement to the original – a post-production provenance[8].

Context

Provenance is useful not just as an identity marker, but because the *who*, *where*, *when*, *why* and *how* of the **context** derived from provenance also act as primary filters of relevance – and, more importantly, irrelevance.

- ***Who*** said it, if the answer is inadequate, may be grounds for disregarding the information. Every source has a bias.
- ***Where*** determines whether the situation documented by the information is congruent with the situation for which the information is being sought.
- ***When*** is a quick sort, if one is looking for current information, and critical if one is looking for information to analyze changes over time.
- ***Why*** also can reveal bias or incompleteness. Depending on the reasons for producing the information, there may be factors that were omitted.

Exception logs do not tell you about normal operations, even though the who, when, and where may be right. No information is universal. **Except for mathematics and perhaps for the chart of elements, all information will be seen by some audiences as having left something out.**

How is also important. In library systems, the *how* is almost synonymous with the *who*, as it is expressed in terms of status and methodology. Information gleaned from interviews will have a different relevance from poll results, where the polling structure constrains the answers to *yes* and *no*. In information systems, the *how* is more closely collected with the *why*. Is the information gleaned as part of a commercial transaction or in the course of a contracted relationship? This *how/why* will determine the contexts in which the information is relevant[9] and also will have a bearing on its *confidentiality*.

Confidentiality

If relevance is one constraint on the reuse of information for research, confidentiality and its twin, privacy, is another. This does not arise in

[7] Think of the recipe and the hotel reviews that clarify which to choose.

[8] Even this is not entirely new. In museums, the exhibit history of a piece of art, while not strictly provenance, is an important part of its documentation.

[9] Recently, libraries, museums, and IT organizations have done a lot of work standardizing the definition of entities relevant to their operations and developing *ontologies* to capture the full semantics of bibliographic information (OWL, FRBRoo, and many others). Basically, this information has gone far beyond the "Place of publication, Publisher, date, etc.," format that served to identify tangible assets for centuries. With digital assets and the Internet, it is not so simple. You need a Unique Resource Identifier (URI) (think: unique name + URL). Approaches such as 3WC's OWL (Web Ontology Language) come in different strengths that allow organizations to address this problem at different levels of complexity.

Copyright © 2007 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

public libraries, though it did in private libraries centuries ago, and it does in archives and in traditional business records management systems. The customs of confidentiality vary from enterprise to enterprise. Often they are unwritten – but violating them can be disastrous. In the past, the default was the need-to-know and exceptions were handled by human discretion, or lack thereof. Now authentication and permissions limits what we can access, and informal sharing is often done by e-mail.

There is a need to foster connection in enterprises of increasingly isolated, even out-sourced, parts. The willingness of departments to share information with peers may depend on trust and competitive pressures within the organization. Portals and collaborative software give new paradigms of securable data sharing and joint access. These paradigms also change what users expect from their IT systems.

The Path Forward

All reference collections are built to fulfill the expectations of the people who will use them. If the re-use plans for IT information are clear as to these expectations, it may be efficient to build the provenance and context into appropriate tags, indexes and other metadata structures as the data designated for reference passes through an initial assessment. It may even be worth revisiting recently used data and adding metadata elements. Certainly, that is what we were taught in library school.

However, in my life as an archivist, I dealt with records groups that measured tens of linear feet and were described in under 50 words. This was sufficient to give the provenance elements, date range, access restrictions (confidentiality), and a few other key words that would determine whether they were worth looking through.[10] With unique items of vast quantity and sporadic relevance, and a minimal budget, this was the way to go. It should be possible to document the applications used by departments and lines of business, and pull together a *snapshot* view of information with the same provenance, and to generate an archival-style collection description that will allow selection or de-selection of the material to be processed by text search or other knowledge management methods. With digital information, such collections do not have to be different copies of the data – though if the data is transformed by indexing, that event should be documented as part of the overhead of the reference transformation process.

As both a librarian and as an archivist, I have been surprised by the strange but not unreasonable ways that patrons use information. As there are trends in academia, so there are trends in business. How a certain kind of information will be used ten years from now – or even two years from now – is hard for us systems-focused folks to see.

One way to move from one state to another when the migration path is not clear, is to create a *threshold*, like a rite of passage. In repurposing enterprise information for reference, such a threshold could be an assessment of a business unit's records, which must be made in any case to exclude both confidential records and those of low relevance beyond the business unit from use in a widely-accessed reference knowledge base.[11]

This assessment is part of the *information usage lifecycle* (as opposed to the ILM information storage lifecycle of IT storage systems[12]). Provided the information is in an application-neutral, open format, it can then be left in its original state of organization, searched, indexed, and cataloged by the wide variety of analysis and business intelligence tools on the market. Inevitably, some information, like database information, will need to be endowed with the contexts given by the table structure and relationships. Many formats are not vendor-neutral, and inconsistencies in the taxonomies used by different parts of the organization will crop up.

Currently, there is a penchant for recasting information as objects[13]. The traditional argument of storing metadata with the object (promoting scalability of the whole) versus putting the metadata on a separate tier (enhancing performance) is pertinent here, for although scalability is obviously desirable, the performance to accomplish vast queries will also be needed. Enterprise digital rights systems may give the paradigm for controlling and measuring use of archived information at a very granular level.

Thus, we come to resemble the surprisingly messy, argumentative world of library science –

[10] You may be sure, however, that Thomas Jefferson's letters are cataloged in detail.

[11] Compliance issues, of course, surpass these boundaries.

[12] See the August 8, 2007, issue of **Clipper Notes** entitled *ILM Stage 2 - Full Spectrum Information Lifecycle Management*, available at http://www.clipper.com/research/TCG2007080.pdf.

[13] There are standards for object attributes in many industries, and increasing attention is being paid to Schema flexibility.

Copyright © 2007 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

but with one, huge, important difference**. With digital data, the ability to re-classify and re-index, and the ability to structure queries with multiple factors to reveal not just facts, but patterns gives an IT-based reference collection a potency far beyond what the largest library of paperbound information could ever provide.** Such a collection would transform how business is done as dramatically as technology has transformed drug development and health care systems.

Reference use tends not to be a sudden burning need. Targeted time-of-use indexing and classification of well-organized data sources, with documented provenance, may be a more efficient approach than trying to catalog each object completely up front. With technology's ability to crawl through vast quantities of data, preemptive classification up front is not necessary, and not always useful.

Repurposing information for reference is an effort, but not an insurmountable effort. Start with data that has obvious reuse. Talk to users, and see what they wish they had access to. Their quick response will probably be things that you cannot provide, like the confidential files of competitors. However, their subsequent responses will be a basis on which to build.

Conclusion

Looking at other information control systems, particularly those developed over time to optimize use of information by many people of many purposes can move the focus of concern from *what we do now* to *what might we be able to do*. Using IT information for reference is, at present, a considerable leap. **The need for better business management demands that the leap be made. The traditions of archives and libraries demonstrate how that leap can be made.**

Copyright © 2007 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Enterprise Architectures and Infrastructure Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 28. (Please dial "1-28" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **Clipper Notes,** and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.

Copyright © 2007 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.