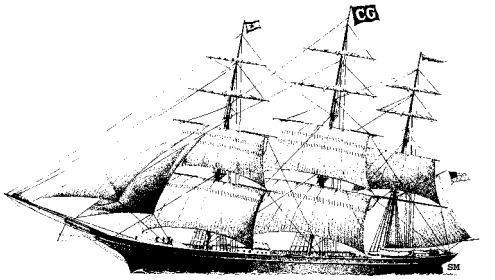
THE CLIPPER GROUP
Explorer™



Navigating Information Technology Horizons℠

Published Since 1996 Report #TCG2004057 July 8, 2004

Building a Fixed Data Repository, Not an Enterprise Junk Drawer

Analyst: Anne MacFarland

Management Summary

Business information can be divided into two flavors: the changeable - *you are here* - real-time, current-state information of inventories, and the mature, fixed - *way we were* - documentation of past events and finished processes. Both kinds of information are important to the enterprise, but the difference in the IT system requirements to protect and provide access to the two types of information is profound. Real-time information demands immediate replication to assure that no data is lost and that applications using the information can be rebuilt or restarted promptly, preferably transparently. Low system latency and high refresh rates are needed to synchronize data for multiple users. Locking and versioning are needed to maintain the integrity of the dataset or file. Nevertheless, when all is said and done, **the value of real-time information is bound to the applications that use it.**

Data as documentation has a broader value. It gives context to current events and reveals a business event as part of a vector, not just a point in time. This is obviously critical to business planning and strategy. **The IT requirements of fixed data differ significantly from those of real-time data. Once replicated, fixed data does not need refresh or back-up because *it does not and should not change.*** Quick access to data is still important, but this access is a matter of latency, not of version control. Fixed data can be accessed and used by multiple people and applications in read-only mode (again, because documentation should not change). If it is transformed (by analysis, or some other process), it will be saved as a separate entity.

Traditionally, information in IT systems has not been segregated by its maturity. Information was saved in the data-dumps of backup, from which events could be reconstructed, but not quickly. Then, a rash of corporate malfeasances highlighted the vulnerabilities of electronic data to revision and data systems to fraud. To address these situations, government regulations were extended beyond the traditional focus on financial records. Now regulations mandate the capture, retention, and retrieval of certain sorts of information (such as medical records) and comprehensive documentation of certain kinds of business activity (such as customer contacts). This requirement has extended the use of data replication from an assist for faulty hardware and media (backup) and a way to get higher throughput (striping) to include saving an inviolable and accurate record of business events in a way that fosters easy retrieval and insures the completeness of *finding*. This changes archiving from a shelve-and-forget-it process to a fully featured, but differently demanding, data access system.

Assuring capture and retention is the first step, but stopping there will not serve the need for quick retrieval as the repository grows, and leaves a lot of potential enterprise benefit on the table. With a little thought, and the information your enterprise may already have in its business process models, the fixed data repository you must establish can be an enterprise asset. For more details, read on.

IN THIS ISSUE

- From IT to Information Management 2
- Some Preliminary Considerations 2
- Dimensions of a Fixed Data Repository .. 3
- A Plan of Action 4
- Conclusion 5

From IT to Information Management

Compliance with government regulations is but a slim slice of the benefits of a good fixed data repository. **Good corporate governance relies not just on comprehensive knowledge of what is going on, but an understanding of how and why it got that way.** In these days of partnering and organizational complexity, gaining the kind of comprehensive, in-depth knowledge of enterprise operations that ERP and CRM applications give to an enterprise's knowledge of its resources and customers is constrained by the fragmentation and obscurity of enterprise information. **This in-depth knowledge of the enterprise is further compromised if information about past events is unavailable or subject to revision.**

This enterprise need to know translates to a need for *information management* to support broad-scale, opportunistic access to data (with appropriate controls), and security to protect both the data and the enterprise. This is not the same thing as *information technology or IT management*, which focuses on devices and systems, yet many IT elements can be involved if the system is to be comprehensive and fast enough for business use.

Long-term preservation of and access to data is a new use for the business computing crowd. Because fixed data is less frequently used, and less relevant to the immediacy of revenue generation, its use cannot generate justification for high expenditures. The once-only backup and elimination of the locking strategies needed to support write access simplify some processes, but the need to support prompt retrieval will add new challenges as a massive sea of assets is generated by multiple applications. It is important to think about the nature of your enterprise. Consider the following questions.

Some Preliminary Considerations

How Does Your Enterprise Generate Information?

The shape, extent and organization of a fixed data repository depends on the nature of the organization whose records it contains. Information may come in text, in forms, and in graphics, as well as in spreadsheets. Some enterprises may focus on the streamlined exchange of goods, services, and money, while others, like law firms and health care, have significant information in often-unstructured notes, correspondence, and conversations. E-mail may be the problem driving the establishment of fixed information archives today, but it is far from the only source of information - or liability - in the enterprise.

Some enterprises live in the present. Others reuse information, and even build their business on that reuse (such the movie industry and other repeatable forms of entertainment).

How Does Your Enterprise Use Its History?

Even within an industry, there are opportunists and there are planners. The need in an enterprise for effective fixed information management will depend on the characteristics of their business processes and the value to those processes of historic information for trending, modeling, and planning. It will depend on their tolerance for waste – the time and money wasted replicating work already done by others or the time wasted simply in finding information.

Of course, there are some routine processes are simply routine processes and from which no nuggets of institutional knowledge can be gleaned. It is important to winnow this chaff out of the wheat, and not spend effort facilitating long-term access to it.

It should also be noted that information can also be a liability, and unfettered access to it a short road to enterprise chaos. The establishment and use of a fixed data repository may call for a fresh consideration of the needs for corporate confidentiality weighed against the need for *big-picture* thinking. Some information may need to be closed to all access save governmental investigations for a period of time – or for its entire retention period.

How Does Your Enterprise Use IT?

IT systems generate the data that eventually becomes the fixed data in a repository, and the nature and demographics of those systems also will shape your repository. The variety of and rate of change in an enterprise application inventory will differ between bleeding-edge enthusiasts and more conservative IT shops.

Why does this matter? Technology generations, like dog years, come quicker than calendar years. Mandated retention periods and corporate prudence dictate enterprises keep data a very long time, when considered in technology years. **Retention of data may exceed the life of the application that generated it, or the hardware that supported that application.** Long-term support of outmoded applications and hardware they run on becomes both difficult and expensive. Depending on the retention requirements of the information and the lifespan of the application that generated or captured it, it may be prudent to transform[1] the data (not generating the contents) into a neutral format.

[1] Most of this transformation extracts metadata from the application or adds index terms from relevant taxonomies to make information more broadly findable.

Copyright © 2004 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

Many of the answers to these questions may be derived from business process modeling and management initiative, and from IT system characterizations, once you change the focus from real-time to residual value.

Dimensions of a Fixed Data Repository

The following four dimensions are basic to all fixed data repositories, whether they are to be used just for compliance or for business self-knowledge as well. They will vary for every enterprise – and may vary for each enterprise over time. They are:

TIME: The Effect of Time on TCO

Long-term data retention and recovery skews the economics, penalizing high ongoing environmental costs and technology turns for the elements involved in storing the data (media, drives) and the elements accessing it (transport protocols, messaging, the gateways of NAS and applications). Consider the following:

- Retention will also surpass the expected lifespan of a disk drive – perhaps more than once.
- Proprietary approaches that depend on the viability of the company take on new liabilities when the period of use is measured in decades. Open standards and neutral formats become invaluable.
- Performance-based acquisition rubrics may not be as relevant as considerations of extensibility. Non-rotating media may become attractive.
- Search and indexing applications must be evaluated in terms of their long-term limitations as well as for their capabilities. The extensibility of XML variants become a comfort, not just a source of confusion.

SIZE: The Need to Scale

Once it is established, is it highly probable that the repository will grow faster than expected, and will increase in scope in ways not anticipated, as was the case when records management programs were instituted in corporations during the 20th Century. An enterprise needs to think how it will structure this growth.

- Some enterprises with diverse and separate operations may set up separate repositories.
- Others will implement fixed data retention as another tier of storage.
- For mid-sized enterprises, or enterprise, or ones with segregated processes that must be documented, a large repository appliance may suffice.

The need to scale out can be met by clusters, federation, single name spaces, open standards, and/or grid. Distributed autonomic capabilities become desirable in any massive entity like a large-scale enterprise fixed data repository. And if your repository will be used for business analysis and short query response time is desirable, distributing the intelligence to parallelize the queries becomes essential.

MANAGEMENT and CONTROL

This large and less intensely used body of information should be as self-managing as possible. The following capabilities are part of fixed data self-management.

- *WORM (Write Once, Read Many) Format*

WORM technologies satisfy a "compliance" requirement, but are also useful to maintain the validity of corporate governance. WORM comes as a hardware or software capability. In hardware, the WORM is a simple-to-manage, permanent restriction. Software WORM may be a part of a file system or an object structure. It is achieved by the combination of a unique identifier containing a timestamp (sometimes called a *fingerprint*) and an enforceable retention period. This retention period needs to be invulnerable to administrative alteration or to accelerating the clock. Once the retention period has passed, with software WORM the storage capacity can be reclaimed for reuse, though if the retention period is long, the physical media may have little residual value.

- *RAID or RAIN*

These striping and failover processes are used to guard against media corruption. Replication can also be used to meet high concurrent demand, and to achieve higher throughput for massive files such as video feeds. Usually remote replication (for disaster recovery) is done once, as the data becomes part of the repository.

- *Access controls*

If information is extracted from the application into a more widely usable format, some kind of enterprise digital rights system will be needed to control who has access to what information. If revenue is to be derived from assets in the repository, the digital rights component should be one structured to support billing.

- *Destruction*

Few enterprises are without some form of real or perceived sin, particularly if they have endured long enough for business practices and rules to change. Information, like any asset, is also a source of liability. At the end of the retention period, data assets can be flagged for destruction or automatically destroyed.

- *Usage metrics*

Repositories should have an audit log of use of

Copyright © 2004 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.

repository information, particularly is the enterprise heavily reuses existing information assets.

TRANSPARENCY: The Ability to Find.

Transparency runs a gamut from highly organized spaces to classic *junk drawers*. Physical IT storage, particularly virtualized pools, may resemble a junk drawer but this is not a problem since the applications that are the gateways to their data know where they have stored it, and what it is that they have stored. In an ideal fixed data repository, applications would expose these assets to some UDDI-style enterprise registry – but the variety and vintage of the applications, the variety of data formats, and security considerations can make this a complex approach. There are various other approaches, many of which may already being used in your enterprise IT environment.

- *A Comprehensive Content Management Gateway*

Content Management regulates the push and pull of enterprise information assets, regulates their life cycle, and can endow their metadata with taxonomy elements and search capabilities to make them more easily found. Records Management applications manage the often structured workflows of an enterprise. Obviously, content and records management applications provide many of the capabilities needed by a fixed data repository. By contrast, document management is a sequencer of a business process and not part of the find process – but it can play a large role in the capture strategy.[2]

- *Standardization and Posting*

In some industries, formats are standard, like DICOM in health care. In other Industries, formats vary by application. Where information is of broad interest or is saleable, it may be worth the effort to transform it using open standards such as XML and all its variants.

- *Seek and Ye Shall Find*

Search, indexing, and knowledge management products targeted at the enterprise may also be appropriate for a fixed data repository. It all depends on how the information is going to be used, and how quickly taxonomies evolve in the industry concerned.

But transparency is more than just a matter of find – it also involves the need to find all. This is obvious in the case of medical records, where a comprehensive view of the patient's health is obviously needed. A comprehensive view of a customer and the usefulness of screening for similar enterprise initiatives before reinventing the wheel are also clear. More subtle is the need to *find all* in order to be able to attest that there is *none*. To do this, the *all* must be managed as a whole.

This last challenge is the most difficult – and (of course) is a mandate of SOX, BASEL, and many other government regulations. Meeting this challenge can only be managed by organizing your enterprise information as an addressable pool of information.

A Plan of Action

First

- Characterize your lodes of information in terms of their residual value and their use in meeting governmental compliance.
- Consider the applications that generate that information you wish to retain for a long time. Determine whether the information should be accessed through the application or extracted into objects. Expected patterns of access to the information by other applications should be a part of this consideration.

Then

- Determine usage needs and restrictions.
- Determine access requirements and restrictions.

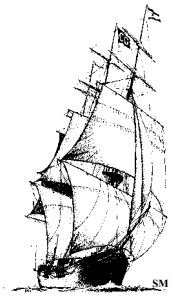
Only then

- Determine the nature of the requisite data services to support the repository (metadata extraction, digital rights, knowledge management) as well as the traditional storage data services (replication, migration).

At this point, you have a profile of requirements for vendors and distributors to meet with a long-term, low-TCO solution.

Conclusion

Government regulation is only one of the imperatives demanding that the seas of enterprise information become more navigable. The different needs and roles of fixed data and real-time data allow an enterprise to control costs by delivering differentiated data services to fixed and real-time data. And, **by treating fixed data as an enterprise asset to be optimized, the enterprise can make a virtue of what government regulations impose as a necessity.**



[2] The capture strategy for a fixed content repository is not covered in this bulletin, but the preliminary considerations listed here also are relevant to the capture strategy.

Copyright © 2004 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*, 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 © 2004 by The Clipper Group, Inc. Reproduction prohibited without advance written permission. All rights reserved.