THE Open GROUP
Making standards work™

Boundaryless Information Flow™

Business Scenario: Measurement of Quality of Digital Information

The Open Group would like to thank all those that have contributed to this Business Scenario through participation in the Workshop that was held on 21 July 2004 at The Open Group Conference in Boston, MA ("the Boston Workshop"), or through submission of comments. The views expressed in this Business Scenario are, however, not necessarily those of any particular member of The Open Group or any particular contributor.

Business Scenario: Measurement of Quality of Digital Information

Document No.: K042

Published by The Open Group, September 2004

Any comments relating to the material contained in this document may be submitted to:

The Open Group
44 Montgomery St. #960
San Francisco, CA 94104

or by email to:

ogpubs@opengroup.org

Contents

Business Scenario Problem Description 1

Detailed Objectives 3

Views of Environments and Processes 5

Actors and their Roles and Responsibilities 20

Information Quality Measurement Methods 22

Requirements 24

Next Steps 27

Glossary of Terms and Abbreviations 28

References 29

List of Figures

Figure 1: Shared Problems Across Industries 5

Figure 2: Hardware Components 11

Figure 3: Information Distribution 12

Figure 4: Information Creation, Management, and Display Systems 13

Figure 5: Example of Corporate Information Flows 14

Figure 6: Information Tagging 18

List of Tables

Table 1: Human Actors and their Roles 20

Table 2: Computer Actors and their Roles 20

Table 3: Information Quality Dimensions (Lee, Pipino, and Wang) 22

Management Summary

This Business Scenario was developed by The Open Group in order to ascertain the requirements for measuring information quality.

Enterprises want to improve the quality of their information in order to improve their efficiency, effectiveness, and reputation, and to comply with the law. Measurement of information quality is a means to this end.

Information technology developments have put massive storage and processing capacity into the hands of enterprises. Most enterprises have a diverse mix of data processing and management products. These products generally use different information formats.

Most enterprise information exists in unstructured form. This makes it more difficult to measure and manage. Against this background, enterprises are finding that:

- Executives waste time because they can't find information.
- Time is spent re-creating information that already exists.
- Mistakes due to wrong information cost money.
- Decisions based on poor information lead to disaster.
- Poor quality information makes an enterprise look like a low-quality outfit.
- Too much effort is spent making different products work together.
- Increasingly, the law requires higher data quality standards.

There are existing methods for measuring information quality, such as Total Data Quality Management (TDQM). There is broad consensus on an overall approach, but the dimensions and metrics are not standardized. The methods are oriented more towards structured than to unstructured information. They are not applied uniformly, and their use is the exception, rather than the rule.

This Business Scenario explores the business and technical environment and processes, and the human and computer actors involved in information quality, and identifies the following as its most important dimensions:

- Ability to find information or to determine that it does not exist
- Accuracy and timeliness
- Trustworthiness
- Meta-information
- Standardization of format

There are **specific requirements for standardization of metadata, tagging, and metrics** to enable enterprises to measure and improve information in these dimensions. The next steps are:

- To review the suitability of existing standards to meet the requirements
- To identify gaps
- To put in hand the integration of existing standards and the development of new ones
- To communicate to vendors the need to implement the standards

A Forum of The Open Group should be chartered to carry out this work.

Business Scenario

Measurement of Quality of Digital Information

Business Scenario Problem Description

Everyone concerned with running an organization – including CIOs, IT directors, and development managers – wants to run their business better. They want to find ways to improve the operational effectiveness of the organization. In most cases they are looking for ways to make these improvements through the business processes, looking at internal processes, and those processes that extend to their suppliers, partners, and customers.

The main challenge to doing this is that information in the typical organization today is not easily integrated, and is not provided to those who need it in an integrated way. Everyone feels that they could run their business better if they could gain operational efficiencies by improving the many different business processes of the enterprise, using integrated information, and improving access to that information.

This is the key requirements statement for Boundaryless Information Flow.[1]

This Business Scenario is concerned with the measurement of information quality, as a tool for improvement of information quality within the context of Boundaryless Information Flow.

The Business Scenario is concerned only with digital information. Other forms of information – such as engineering drawings, typescript, and hand-written notes – exist in many organizations, and impact on their overall information quality, but are beyond the scope of this document.

What is information? It is a concept that is not easy to define. A working definition for many IT professionals is “data in context”. This Business Scenario assumes a broad community of understanding in its readership on what information is, and does not attempt to define it. In fact, the Scenario does not systematically distinguish between “information” and “data”.

The distinction that is important for this Scenario is that between *structured* and *unstructured* information (or data). Structured information – such as is found in database management systems – is relatively easy to manage. Coping with unstructured information – such as documents, email, and spreadsheets – is much more difficult. Most information held by enterprises today is unstructured. While the quality of structured information is not easy to measure and improve, the task for unstructured information is much harder.

[1] Boundaryless Information Flow is a significant approach to tackle a major problem of enterprises today. It is shorthand for “Access to integrated information to support business process improvements”. Boundaryless Information Flow is a desired state for an enterprise’s infrastructure and is specific to the business needs of the organization.
An infrastructure that provides Boundaryless Information Flow has open standard components that provide services in a customer's extended enterprise that:

- Combine multiple sources of information
- Deliver information to the places where that information is needed
- Deliver it in the right context for the people or systems that use it

For more about Boundaryless Information Flow, see www.opengroup.org.

The problem, then, is to find ways to measure and improve the quality of information that is largely unstructured.

This Business Scenario explores the business and technical environment and processes, and the human and computer actors involved. It briefly reviews existing information quality measurement methods, and then presents a set of requirements for improving and measuring the quality of information. These cover both structured and unstructured information, but apply particularly to unstructured information.

Detailed Objectives

Why do enterprises want to improve information quality?

Poor quality information affects enterprises in many ways:

- Executives waste time because they can't find information.
- Time is spent re-creating information that already exists.
- Mistakes due to wrong information cost money.
- Decisions based on poor information lead to disaster.
- Poor quality information makes an enterprise look like a low-quality outfit.
- Too much effort is spent making different products work together.
- Increasingly, the law requires higher data quality standards.

Enterprises want to improve the quality of their information in order to improve their efficiency, effectiveness, and reputation, and to comply with the law.

Reducing the cost of managing IT is not a part of this objective. An enterprise will spend more on its IT department if this will lead to greater efficiency and effectiveness in the organization as a whole. Improving information quality may not be cheap!

Each enterprise will want to achieve one or more of the following primary objectives by improving information quality:

- Improve project times
- Improve production costs
- Improve production quality
- Increase customer satisfaction
- Reduce the time taken to find a piece of information
- Reduce the time taken to build an interface
- Reduce the time taken to perform an information operation (e.g., indexing)
- Reduce time taken to get data updated
- Increase probability that decision-makers have the information they need

Second-order objectives are:

- Lower storage costs
- Lower discovery costs
- Avoidance of risks

Each organization will have specific percentages or amounts that apply to each objective. For example, "*Improve production costs by 10%*".

- A typical target for the time taken to build an interface is "less than a day".
- A typical target for the time taken to perform an information operation is "less than 10 seconds".
- A typical target for the time it takes to find an item of unstructured data (a document, email, etc. rather than a database record) is to cut it from 2 hours today to under an hour.

Given the desire to improve information quality, enterprises will need to measure that quality. Without metrics, it will not be possible to judge how far the measures taken are successful. Measurement enables improvement through a process of action, feedback, correction, and further action.

The objectives for developing metrics are to enable the objectives for improving quality to be achieved.

Views of Environments and Processes

Business Environment

Quality of information is an issue that applies across the whole of industry, commerce, and government.

There is commonality of areas and problems across industries, as illustrated in Figure 1. This is important because, given this commonality, it is likely that standard solutions can play a role.

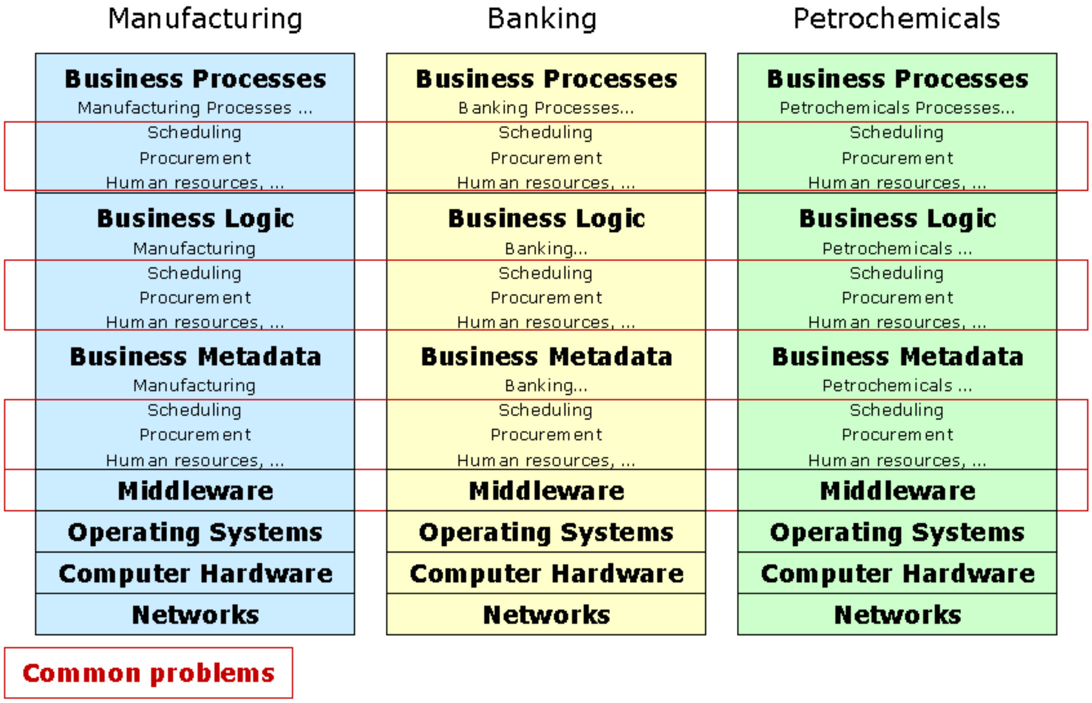

Figure 1: Shared Problems Across Industries

Information plays a key part in all of the processes shown in Figure 1. The quality of that information is a problem that is common to all of them.

Business Drivers

The driver for measuring information quality is:

- To improve the quality of the information

The main reasons for doing this are:

- To improve operational efficiency
- To cut development costs
- To improve the quality of decisions
- To comply with legislation

Operational Efficiency

The three most important ways of improving operational efficiency are:

- Saving time
- Reducing mistakes
- Increasing information re-use

Saving Time

Lost time is the most serious problem. It translates to cost. The larger the business, the larger the information base, and the greater the proportion of time lost. This is a fact of corporate life.

Time is lost when information cannot be found quickly, or at all. This problem increases as the corporate information becomes less structured and the search criteria become less precise. It results in lower staff productivity.

Chevron Texaco estimate that their executives spend one hour a day just looking for information. Of the time spent finding and using information, 60% is spent finding it and 40% is spent using it. Reducing the time to find information so that people will be more productive is a key management objective.

Reducing Mistakes

There can also be costs that are directly due to poor information. For example, there can be high marketing costs due to duplicate mailing addresses.

Information Re-Use

Information re-use, like software re-use, can have a huge effect in improving productivity.

If software is built well and fulfills a purpose, then it should be re-used, but this requires good documentation of the software system. The same applies to information. Information re-use is a major issue that affects decision-making processes across the board. It also affects business partners.

Failure to re-use means the extra cost of redeveloping information, plus the likelihood of increased mistakes. The potential dollar impact is enormous.

A significant part of the problem is the lack of ability to re-use information that is maintained by particular employees when those employees leave the corporation. This information is often kept on personal computers, or stored on shared drives using idiosyncratic filenames that no-one left in the organization understands. Nevertheless, this data is corporate knowledge, and the corporation can ill afford to lose it.

Chevron Texaco estimate that, in their asset utilization processes, they spend $100 million per year looking for plans and other information that support prior asset utilization decisions.

Development Costs

Building interfaces between applications is a pain point that results in excessive costs.

This is not an issue that just affects individual organizations. Enterprises wish to work with customers and suppliers. To do so, they must handle information provided by those organizations. For example, a major customer organization can dictate information standards to small suppliers.

Often, information is poorly integrated into an application, and its use following the original design can result in very high cost to the thousands of suppliers that use it.

The situation at Lockheed Martin shows the scale of the problem for a large corporation. With over 30,000 suppliers, and hundreds of large customers (particularly the US DoD), and a need to interface to those suppliers and customers, Lockheed Martin has information in a large number of different formats, and needs to translate between them. A substantial number of people spend their time building interfaces. And people just add applications as a new processing or conversion need arises. Often, because information on existing transformations cannot be found, a new interface system is built when an increment to an existing system would meet the need.

The problem of integrating applications that use different information formats causes:

- Problems in the supply chain through misunderstanding of the nature of the information
- Added cost
- Added time to build the interfaces

This issue relates, not to the quality of the information content, but to the quality of its form and structure.

Decision Quality

Business decisions are taken on the basis of information. Poor quality information leads to poor quality decisions.

In some contexts – such as the military context – poor information can lead to loss of life. In industry, health and safety can be compromised through misinformation on dangers and safeguards.

In the business context, opportunities can be lost through misunderstanding of the situation. Losses can be made because marketing markets things that customers aren't looking for. Business can be lost through poor knowledge of the customer base.

Quality decisions need complete, timely, and accurate information. And they need the right information: presented with a mass of data, it can be difficult to identify what is important. And sometimes it is important, not to find information, but to establish that it is not there. Organization and structure are important qualities in information used for decision-making.

Being able to find information is not always enough. Information without pedigree is not trusted. Suppose you are researching the market need for a new widget, and find a note that says: "80% of automobiles manufactured in the USA include widgets that cost $5 or more". This looks good, because you can produce widgets at $1 each. But do you believe it? Not unless you know who wrote the note, when they wrote it, and how they obtained the information. You must do further research.

A geologist talking about selecting the point where a well will be drilled is likely to be asked two questions: "Have you considered all the relevant data in determining the recommendations?" and "Can you rely on the integrity of the data?" Even if he has all the data about the data and can say yes to the first question, it is unlikely that he will be able to answer the second one.

Some information is highly subjective. Whether it can be trusted depends on how it was produced, and by whom. In Chevron Texaco, for example, they do not try to audit the content of seismic data; they audit the processes by which it was produced and maintained.

Legislation

There is an increasing amount of legislation relating to information quality. Failure to comply with legislation can lead to financial penalties for a corporation, and in some cases to its officers being imprisoned. Mitigation of the risk of non-compliance is a major driver for improving information quality.

In addition to the criminal law, corporations are concerned to avoid civil suits.

The law can be broken through failure to keep accurate records. It can also be broken through having information that you should not have or do not need.

Structuring data aids compliance. It makes the information more manageable. It also makes it easier to determine categorically that a corporation does not have certain information. Legal departments see unstructured data as a threat.

Some of the relevant legislation is summarized below, including examples of laws in particular US states.

Utah Government Records and Management Act

Subject to certain conditions, every person has the right to inspect a public record free-of-charge, and the right to take a copy of a public record during normal working hours. All records are public unless otherwise expressly provided by statute.

UK Data Protection Act

Personal data must be obtained for specific purposes, and must then be processed only for those purposes, and not kept longer than necessary. It must be adequate, relevant, and not excessive. It must be accurate and up-to-date. Its security must be maintained. It must be processed fairly and lawfully. Cross-border transfers are restricted. Officers can be held liable for offences committed by their institutions.

California Records Management Act

The director shall establish and administer in the executive branch of state government a records management program, which will apply efficient and economical management methods to the creation, utilization, maintenance, retention, preservation, and disposal of state records.

US Data Quality Law

The Office of Management and Budget (OMB) must develop government-wide standards for the quality of information used and disseminated by the federal government, and issue guidelines for data quality which define four key terms: quality, objectivity, utility, and integrity. Other federal agencies must issue their own conforming guidelines and must report periodically to the Director of OMB on the data quality complaints that they have received.

The OMB issued guidelines in 2001 and updated them in 2002. They require federal agencies to issue information quality guidelines for the information that they disseminate, to establish mechanisms by which affected people can seek correction of information, and to report annually to the OMB.

Individual agencies have issued their guidelines in response to this. For example, the Surface Transportation Board has issued a set of guidelines which define utility, objectivity, and integrity as the main aspects of information quality, define an information correction request procedure, and list a number of documents with specific quality criteria. (For example, "Railroad Rate

Studies" to be published on the website within 10 days of production and to be reachable within five clicks from the home page.)

HIPAA

The US Health Insurance Portability and Accountability Act of 1996 (HIPAA) is mainly concerned with general improvements to the health insurance system, but includes provisions relating to electronic information. As President Clinton said: "It will modernize, streamline, and cut the cost of insurance paperwork by devising a uniform electronic system for paying health care claims. It will provide steps to protect the privacy of people in the system as it does so." It requires use of a single set of national standards and identifiers, and imposes security standards.

Sarbanes Oxley

The US Sarbanes Oxley act was passed in response to a number of highly publicized failings in corporate accounting and auditing procedures. Amongst other things, it makes the officers of a corporation responsible for the quality of financial information.

Enterprise Quality

The quality of enterprise information can significantly influence the perceived quality of the enterprise itself.

For some organizations, information is the end product. For example, a major role of the Federal Highways Agency is to provide information such as highway safety numbers, injuries, and working zone information to state and local agencies. The information supports decisions such as selection of material for specific roads under specific conditions. Getting this wrong could lead to loss of life. For organizations like the Highways Agency, improving the quality of their information means improving their product.

This quality is not always completely under the control of the organization concerned, or of its IT department. For example, the Federal Highways Agency has experienced delays of two years in fatality data trickling up to them. Human stewards do not necessarily share their data readily.

Even where information delivery is not part of an organization's core function, poor information quality can impact the organization's standing and reputation. Poor accounts information leading to incorrect or duplicate invoicing reflects on a company's competence. Poor information about customers can lead to embarrassing mistakes (as, for example, when a hospital sends an appointment letter to a patient who has just died).

These things lower the confidence of business partners and customers in an organization's general ability to perform. As a result, business relationships can suffer, and corporate reputation can suffer.

At its worst, inability to present correct and relevant information may prejudice relations with business partners and affect contracts. Large as well as small contracts can be affected by poor information.

Damage to reputation can be significant, and can be very hard to overcome. (Who can forget the Mars Climate Orbiter, lost because of an error in a transfer of information between two teams that used US and metric measurements respectively?)

The damage to business relations and reputation resulting from poor quality information can reach billions of dollars.

Business Processes

Most business processes depend on quality of information:

- Manufacturing processes rely on process specifications, blueprints, supplier details, component catalogs, and quality control information.
- Sales processes rely on information about customer and prospect contact details, customer preferences, product features and pricing, and available stock levels. Decisions are taken on the basis of customer credit histories.
- The marketing function uses market breakdown information, historical sales data, etc.
- HR departments keep records of employee contact details, work assignments, and career progress.
- Corporate finance needs records of bank transactions and balances, customer and supplier account details, and employee salary levels and payment methods.

These are some obvious examples. The complete list is a long one. In fact, it is hard to imagine a business process today that does not have some dependency on information.

Technical Environment

Ever since their earliest use in the 1950s, the capabilities of computing devices to store and process information have grown at a rate that would be regarded as unbelievable in most industries, but that in the IT industry is taken for granted.

Gordon Moore, co-founder of Intel, noted in 1965 that the number of transistors per square inch on integrated circuits had doubled every year since the integrated circuit was invented, and predicted that this trend would continue for the foreseeable future. In fact, data density has doubled approximately every 18 months, and this phenomenon is known as Moore's Law. Intel expects that it will continue at least through to the end of this decade.

This massive fundamental technical improvement has led to an explosion in the amount of data owned and maintained by enterprises. Enterprise data stores are now measured in terabytes. It has also led to an explosion in the power of computing platforms, and in the sophistication of information processing applications.

Hardware Components

The main hardware components of the technical environment are illustrated in Figure 2.

Figure 2: Hardware Components

Terminals

"Dumb" terminals enable users to view information, and to interact with applications, but do not include local information storage or processing capabilities.

Personal Computers

Personal computers include both desktop and laptop devices. Today, they can be quite powerful machines, with significant information storage and processing capability.

PDAs

Personal Digital Assistants (PDAs) are small personal computing devices that can be carried in the pocket. They have only limited information storage and processing capability. Some mobile telephones have sufficient intelligence to be included in this category.

Multimedia Devices

There is an increasing number of multimedia devices, including digital cameras, digital audio devices, and document scanners, that are able to create information that can be processed in and stored by computer systems.

Servers

A server is a shared computing resource that provides information processing and storage capabilities to multiple users. Within an enterprise, a server may be accessible for use by the whole enterprise, or may be dedicated for use within a single division or department.

Information Storage

Information storage systems include disk-based network storage systems, tape backup systems, and departmental shared drives. They are typically controlled and managed by the enterprise or department, rather than by the individual. They are often used to share information, and for security.

Printers

Printers are used to make information available in “hard copy” form.

Networks

Networks within enterprises, and the global Internet, form the basic transport system by which digital information is conveyed from person to person and from place to place.

Information Distribution

Figure 3 illustrates how information of different kinds can be distributed over systems of different kinds in an enterprise.

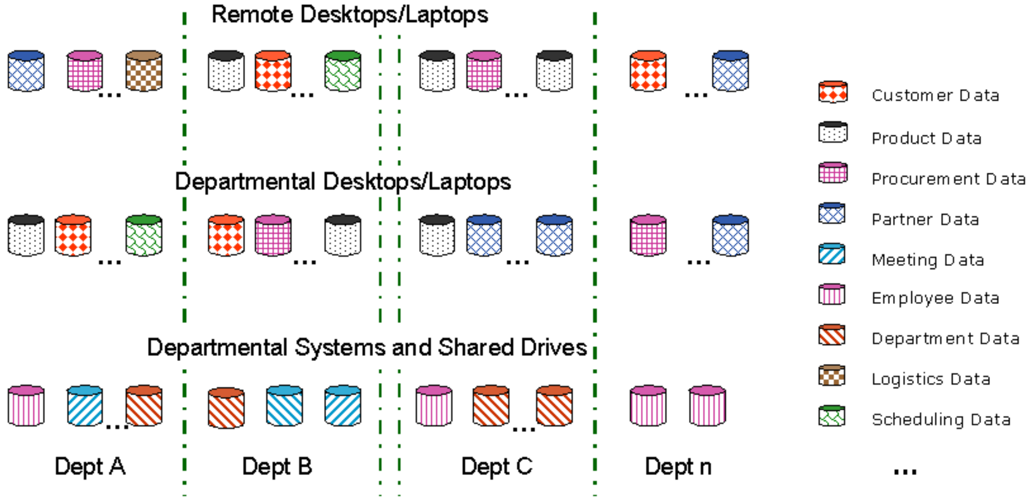

Figure 3: Information Distribution

Each server or information store can contain information of many different kinds. And related information of the same kind can be spread over many different servers and information stores. For example, customer data may be held on the laptops of members of the sales force, in the sales departmental server, and in the technical support department database.

Some information may be held on portable memory stores, such as CDs.

Some data is only is accessible through specific applications.

Information Structure

Information can exist in structured or unstructured form, and this is a key distinction.

If all the information is in the corporate database, and search requirements can be expressed as simple SQL queries, then the time to find information depends simply on the efficiency of the DBMS, and is generally measured in seconds. “Find all customers that live in New York State” is a simple matter when there is a well-organized customer database with a “State” field in the contact address record. “Find everyone who lives near New York and has sent us email in the past two years” is a tall order if all you have are the mail archives.

Today, only about 20% of corporate information exists in structured form. The remaining 80% is in mail archives, spreadsheets, documents, note files, etc. Most of it is organized (if at all) according to the whims of individual employees, on their PCs, and on departmental servers and shared drives.

For example, Chevron Texaco has about 90 terabytes of corporate information, growing at about 30% per year. They estimate that 85% of it is unstructured – emails, contracts, terms, studies, etc. About 15% of it is on laptops, and more than that is held on shared drives; this information is managed by the end-users using shared drives and shared services.

Generally, there is a lack of processes for handling unstructured information. Communication is not controlled. Quality checks are by-passed. Structured data is much easier to manage and control.

Information Creation, Management, and Display Systems

Running on the hardware components shown in Figure 2 are software systems for creation, management, and display of information. These are illustrated in Figure 4.

Figure 4: Information Creation, Management, and Display Systems

These systems process information to meet the needs of the business. Often, there are business rules that they must follow.

There is considerable overlap between some of the information transmission and processing systems shown in Figure 4, notably ERP Systems, Middleware, ETL Tools, Content Managers, Data Warehouses, and PDM systems. They offer different approaches to a common set of problems: the need to organize information and enable integrated use of it by different applications.

Figure 5 shows the information flows in a department of one major corporation, and illustrates some of the problems that arise. They have a content management system, but it does not solve all of the problems.

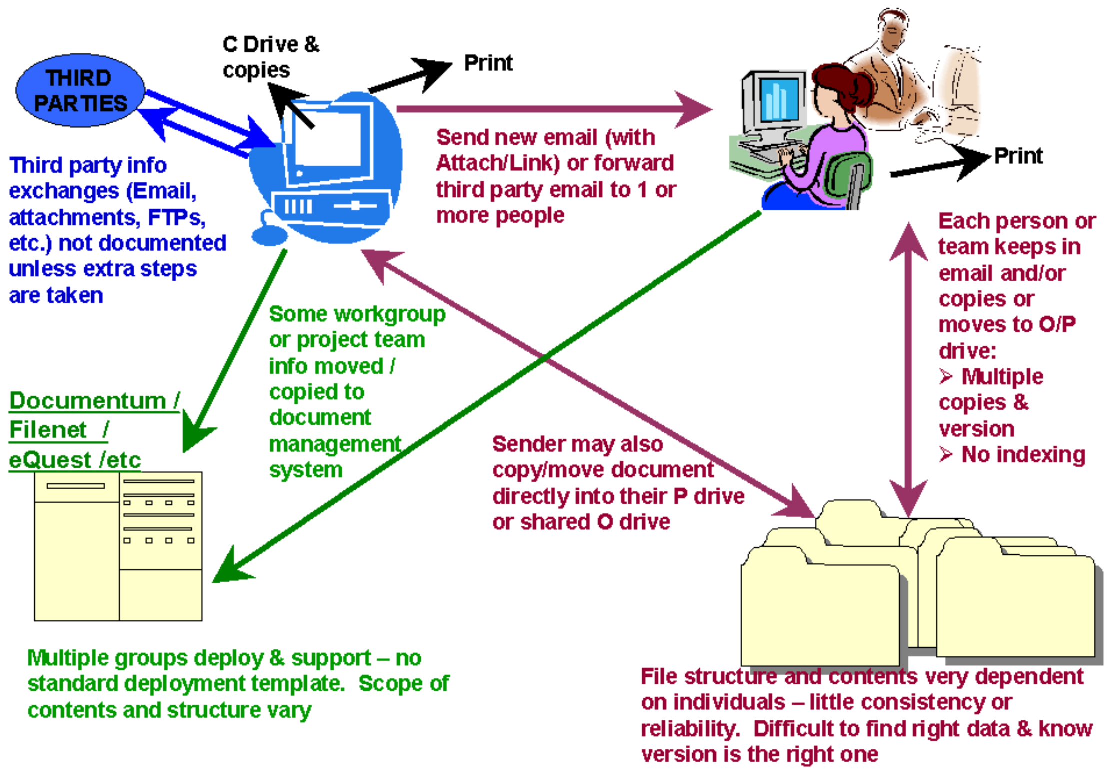
THIRD PARTIES
C Drive & copies
Print
Third party info exchanges (Email, attachments, FTPs, etc.) not documented unless extra steps are taken
Send new email (with Attach/Link) or forward third party email to 1 or more people
Print
Each person or team keeps in email and/or copies or moves to O/P drive:
➢ Multiple copies & version
➢ No indexing
Some workgroup or project team info moved / copied to document management system
Documentum / Filenet / eQuest /etc
Sender may also copy/move document directly into their P drive or shared O drive
Multiple groups deploy & support – no standard deployment template. Scope of contents and structure vary
File structure and contents very dependent on individuals – little consistency or reliability. Difficult to find right data & know version is the right one

Figure 5: Example of Corporate Information Flows

An organization may seek to follow a single approach; for example, Content Management or Data Warehousing. But circumstances often lead to proliferation of systems and development of a "spaghetti network" of applications and information management and transformation tools. Then new business requirements create a need to reach information through a mess of interfaces.

Legacy systems often remain crucial to business operation. Legacy systems need to share information with new systems, and with each other.

Data Managers

Relational database management systems are the most important products in this category. Other data management systems include specialist databases such as directories.

ERP Systems

Enterprise Resource Planning (ERP) software attempts to integrate all departments and functions across a company onto a single computer system that can serve all those different departments' particular needs. It is important for handling structured information. It is a solution that is most often used within the manufacturing environment.

An ERP system typically contains several modules, such as a financial module, a distribution module, and a production module. The key characteristic of ERP is that all of these modules are integrated, and share information that is housed within a single common database.

Middleware

Middleware is a layer of software between the operating system/network layer and the applications. This software can provide services of many kinds, including:

- Inter-application messaging (often as a connectionless, asynchronous transactional store-and-forward capability)
- Information transformation (including field reordering, data validation and filtering, and rules-based translation)

ETL Tools

Extract, Transform, and Load (ETL) tools enable enterprises to take data from multiple sources, reformat and cleanse it, and load it into another system for analysis or to support a business process. The data can come from sources such as corporate databases or applications, flat files, and spreadsheets. It may just be re-formatted, or it can also be cleansed; for example, to remove duplicates. Re-formatting and cleansing follows custom-defined business-specific rules. The treated data may be stored in a central database or fed into a business application.

Content Managers

Enterprise content management software provides a set of tools and processes for managing all types of content, from textual documentation to video, throughout its lifecycle. It includes Web Content Management, Document Management, Catalog Management, Product Data Management, XML Publishing, and Digital Asset Management.

Categorization and tagging of information to facilitate retrieval is a key aspect of ECM.

Data Warehouses

A data warehouse is a repository storing integrated information for efficient querying and analysis. As information is created or updated by different applications, it is translated into a common data model and integrated with existing data at the warehouse. This makes it easier to find and use information that was generated by heterogeneous sources.

PDM Systems

Product Data Management (PDM) systems address management of product data, and also management of the processes by which that data is created and used. They are important for management of unstructured information.

Master data is held in a secure "vault" where its integrity can be assured and all changes to it monitored, controlled, and recorded. Duplicate copies can be distributed freely, to users in various departments for design, analysis, and approval. When a change is made, a modified copy of the data, signed and dated, is stored in the vault alongside the old data, which remains in its original form as a permanent record.

Classification of data is fundamental to PDM. Information is grouped in named classes, and attributes can be used to describe the essential characteristics of each component in a given class. This enables retrieval by browsing classification trees and searching for attributes.

Process management is also fundamental to PDM. It is concerned with the procedures by which:

- People create and modify data
- Data flows between people

Process management functions keep track of all the information creation and modification events, and of all the data flows that occur, during the history of a project.

Individual PDM systems vary widely in how they perform these functions.

Workflow Managers

Workflow managers provide automation of business processes in which documents, information, and tasks are passed between human and computer participants for action, according to procedural rules.

Personal Applications

Personal applications include word processors, spreadsheets, and other office applications. They also include task-related applications such as CAD. They are a primary means by which digital information is created (or information is put into digital form).

Most of these applications create information in unstructured form. Movement of information between them is a manual and error-prone process. For example, information may be copied from an Excel spreadsheet to a Word document and then to a database, with loss of information at each translation.

Corporate Applications

Corporate applications include those used internally, such as HR and Payroll, and those that form part of the organization’s external interface, such as web services and transaction processing. Some organizations have multiple applications of the same kind. For example, an enterprise might have several HR systems.

As well as new applications, most enterprises have legacy applications that support old business requirements.

Email

Email is the primary tool used by people to send information to each other in enterprises today. The information is sent in the form of unstructured text messages, to which files of various kinds may be attached. Many email systems provide archival storage for messages that an individual has sent and received.

A particular issue is that email generates multiple versions of attachments, rather than a single master for reference.

The Web

The World Wide Web is a massive source of information. An enterprise may use it as an information resource or as a means of delivering information internally to its employees and externally to business partners and customers. Many enterprises have their own webs on corporate intranets for internal dissemination of information.

Technical Processes

Creation

Information is created in digital form when people write documents, produce spreadsheets, enter information into databases, enter information into applications, and so on. It can also be created by digitizing drawings, digital audio recording, digital photography and video, and other multimedia capture processes.

From the point of view of an organization, information imported from outside (for example, when an email is received or a parts list downloaded) is effectively created.

Modification

Information is modified by applications with editing capabilities. Often, each such application is dedicated to a particular kind of information. For example, word processors are used to modify textual documents.

A distinction is made here between *modification* and *transformation*: modification changes the information content; transformation changes its form but preserves its content.

Deletion

Information is deleted when it is removed from computer storage. Generally, this is done by an editing application, by the computer operating system or storage management system, or by an information management system.

Viewing

Viewing is the process by which information is displayed to people. Viewing information enables people to use it.

Classification

Classification is the process by which classes, attributes, or types are associated with information items. These classes, attributes, or types may be represented by *tags,* which are stored with the information. They often have *codes* associated with them to remove ambiguity; these are unique, possibly arbitrary, alphanumeric identifiers. For example, in the oil industry tags are organized by domains – finance, legal, engineering, seismic, etc.

Definition of information classifications is typically a difficult matter. Problems can be found in clashes between semantics of similar value sets in times of integration. Inconsistent tagging is often encountered.

Code lists should be corporate data. Governance should enforce common standards. It is a bad idea to create a code list based on a single departmental view; a broad constituency should be involved in code set standardization. For example, well status code lists are used in the oil industry, but it is often found that what works for a department does not work for the company. It can be disruptive for departments to change what works for them to support a company standard, and persuading them to do so may not be easy.

The tagging process is illustrated in Figure 6.

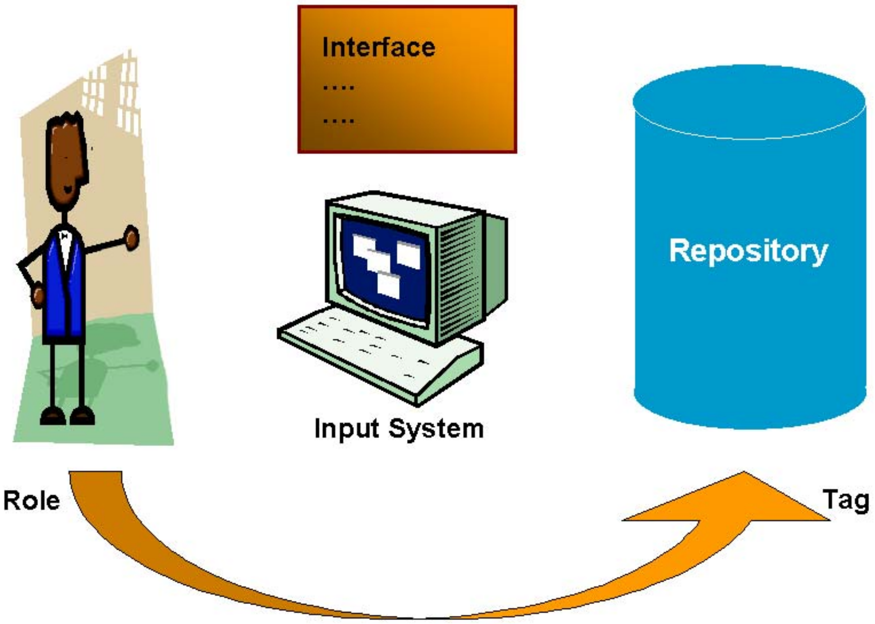

Figure 6: Information Tagging

Tags should ideally be associated with information at the point of creation. Not all information needs to be tagged. A key question is: "Under what circumstances should a piece of information be tagged?"

The appropriate tag will generally depend on the role of the person creating the information. Different classifications are relevant to different people.

Storage

This is the holding of information in some kind of digital memory so that it can be retrieved for viewing.

Retrieval

Retrieval is the process by which users obtain information for viewing. It includes:

- Navigation by Index: The information store has an index to its content, and the user first finds an index entry and then de-references it to find the information. Indexes can be multi-level.
- Browsing: The user scrolls through a set of information until the desired information is found.
- Searching: The user enters criteria that describe the desired information, and the system searches for it based on those criteria.

Transmission

Transmission is the process of moving information between users, applications, and information stores.

Transformation

Transformation is the process of changing the form of information while preserving its content. (And is thus distinguished from *modification*, which changes the content.)

Different applications often use different information formats. This means that, if one application is to use information created by another, transformation of the information is required.

Understanding of the information semantics is most important for building interfaces between applications. In order to transform the information correctly, it is necessary to be able to map the applications' semantics. For example, in interfacing order management and procurement systems, it is necessary to align their product description semantics.

Process Management

Management of the processes for creating and modifying information is an important factor in information quality.

In the past, when information handling was paper-based, many organizations had formal procedures for cataloguing and storing documents. These procedures do not apply in the e-world, but they have typically not been replaced by new ones. For example, in Chevron Texaco, 85% of real business use is now bypassing the old records management process.

The library services provided by records managers and secretaries that helped to make information re-usable have been lost. Procedures that helped assure information quality are no longer followed. Audit trails, providing evidence that process has been followed, no longer exist.

Actors and their Roles and Responsibilities

Human Actors and Roles

Everyone with access to the enterprise network plays a role in the use of enterprise information. This includes partners, contractors, etc., as well as employees. It is often impossible to tell them apart.

The human actors and their roles are listed in Table 1.

Table 1: Human Actors and their Roles

Human Actor	Roles
Creator	Creates information, or imports it into the enterprise.
Updater	Modifies information
Handler	Handles movement of information between members of enterprise and applications and information stores.
Receiver/Consumer	Retrieves and uses information.
Lifecycle Manager	Has authority to archive or delete information.
Manager	Member of enterprise management structure, which can be complex. Responsible for processes that create or use information.
Governance Group	Develops and monitors application of enterprise information management standards and processes. Rules and governance impacts on productivity and decision quality by identifying the really relevant information – and can help to determine when something does not exist.
IT Manager	Manages enterprise information infrastructure.
Developer	Develops extensions to enterprise information infrastructure, including special-purpose interfaces between applications.

Computer Actors and Roles

Computer actors include the hardware components illustrated in Figure 2 and the Information Creation, Management, and Display Systems illustrated in Figure 4. The complete set of computer actors, and their roles, is shown in Table 2.

Table 2: Computer Actors and their Roles

Computer Actor	Roles
Terminals	Interface to corporate applications.
PCs	Interface to corporate applications, personal application platform, email platform, and storage.
PDAs	Personal application platform, email platform, and storage.
Multimedia Devices	Creation and viewing.
Servers	Personal application platform, email platform, and storage.
Information Storage	Storage.
Printers	Viewing.

Computer Actor	Roles
Networks	Transmission.
Data Managers	Storage and retrieval.
ERP Systems	Storage and retrieval.
Middleware	Transmission and transformation.
ETL Tools	Classification, storage, and retrieval.
Data Warehouses	Transformation, storage, and retrieval.
PDM Systems	Classification, storage, retrieval, and process management.
Workflow Managers	Process management.
Personal Applications	Creation, modification, deletion, and viewing.
Corporate Applications	Creation, modification, deletion, viewing, storage, and retrieval.
Email	Creation, viewing, deletion, and transmission.
The Web	Viewing, storage, and retrieval.

Information Quality Measurement Methods

The state of the art in information quality measurement theory is represented by the Total Data Quality Management (TDQM) approach.

The TDQM program is a sponsored research program at MIT, founded in 1991 to:

- Establish a solid theoretical foundation
- Devise practical methods for organizations to improve the quality of their information

The TDQM program defines data quality as follows:

> Although the notion of "data quality" may seem intuitively obvious, data quality is not well defined in current practice. Our studies have revealed that data quality has a number of dimensions for data users, including accuracy, believability, relevancy, and timeliness. A clear and uniform articulation of data quality metrics is needed. In fact, even a relatively obvious dimension, such as accuracy, does not have a sufficiently robust definition to make techniques apparent as to how to measure the accuracy of data.This leads to an approach where:

- A set of information quality dimensions is identified.
- A set of metrics is defined for each dimension.

While there is consensus on this general approach, there is no agreement on a standard set of dimensions, or on standard metrics for the dimensions. Different sets of metrics can be found in the literature. There are a number of published metric definitions. They include excellent and effective methods, but they are not standardized.

A typical set of dimensions, due to Pipino, Lee, and Wang, is shown in Table 3. Note that the dimensions do not all allow objective measurement; some are quite subjective.

Table 3: Information Quality Dimensions (Lee, Pipino, and Wang)

Accessibility	Interpretability
Appropriate amount of data	Objectivity
Believability	Relevancy
Completeness	Reputation
Concise Representation	Security
Consistent Representation	Timeliness
Ease of Manipulation	Understandability
Free of Error	Value-added

Many information quality practitioners, using variations on the core method, implement the TDQM approach. They can:

- Model an organization's use of information
- Identify appropriate dimensions and metrics
- Determine values of metrics at key stagesThis enables them to:
- Produce a Financial Impact Assessment by:
  - — Determining the cost of poor quality information at each stage, due to increased costs, reduced revenue, delay, customer dissatisfaction, etc.

    - Aggregating to obtain an overall picture of information quality (and justification for doing something about it)

- Apply statistical process control methods to monitor and control information quality:
    - Data is sampled regularly at key stages of the organization's information model.
    - Control charts and other statistical techniques are used to determine when something is wrong (unusual number of invalid records, unusual number of particular kind of valid record, breach of business rules, etc.).
    - Investigation can then determine the cause of the unusual data, and the problem can be fixed.

Methods such as this are in place in some organizations, particularly for highly important information. Some data management tools support them. But they are not applied uniformly, and they are the exception, rather than the rule.

Requirements

This section discusses the key information quality dimensions, methods for improving information quality, and methods for measuring information quality. These discussions expose requirements for standardization of metadata, tagging, and metrics. These requirements are summarized at the end of the section.

Key Information Quality Dimensions

The business drivers for information quality identified in this Business Scenario are:

- Operational efficiency, through:
  - Time saving
  - Reduced mistakes
  - Information re-use
- Reduced development costs
- Improved decision-making
- Improved enterprise quality
- Conformance to legislation

For this Business Scenario, the important dimensions of information quality are:

- Ability to find information or to determine that it does not exist

  Contributes to operational efficiency through time-saving, to decision-making, and to conformance to legislation.
- Accuracy and timeliness

  Contributes to operational efficiency through avoidance of mistakes, to decision-making, to enterprise quality, and to conformance to legislation.
- Trustworthiness

  Contributes to decision-making.
- Meta-Information

  Having information about the information contributes to operational efficiency through information re-use, and to conformance to legislation. Meta-information can also record the information's "pedigree", and contribute to trustworthiness.
- Standardization of format

  Contributes to reduced development costs.

Improving Information Quality

Measurement is a means to the end of improving information quality. The end is more important than the means. There are techniques for improving quality other than measurement, which may be more effective. In particular, characterizing content is more important than measuring quality.

Each enterprise should develop an overall information management strategy that takes a holistic view, and defines infrastructure metadata management services connected to business processes that make sure data is captured appropriately.

The strategy should be reflected in information quality policies, and backed by a governance structure. It will be necessary to change behavior. Policies should cover information handling processes and toolsets, including those for email and attachments. They must be usable and kept up-to-date. They should reduce information duplication and redundancy.

Meta-information is a key tool. This should cover:

- Purpose and usage of information, to facilitate re-use
- Creators, modifiers, and circumstances of creation and modification
- Quality processes that have been applied, to indicate trustworthiness

It must be able to handle compound documents, such as an email plus its attachments. It must be applicable to large-grain collections of information, as well as fine-grain ones.

Information tagging should be supported in information management product interfaces. This should be sensitive to the roles of the users, so that appropriate choices are presented.

Industry standards are needed for metadata and indexing. A naming standard is a critical component. It should not be enforced on everyone, but should be a turntable standard that enables communication, and facilitates automated transformations; for example, using XSLT.

Application interfacing is currently very labor-intensive. It is an art form, but development of standard semantics could turn it into a science.

Solutions should be global, applying across national and cultural boundaries.

Measuring Information Quality

The following information quality metrics are needed:

- Time taken to find information, or to determine that it does not exist
- Accuracy
- Time taken to perform an operation, such as indexing, on information
- Time taken to build interfaces

Measurement of the first three of these quantities is not straightforward. There would be benefit from development of standard metrics, and for implementing support for them in products. This is, however, not the case for the time taken to build interfaces. This is an important metric, but it can be measured by enterprises in a straightforward manner.

The time taken to find information is particularly important where unstructured information is concerned.

Accuracy can be very hard to measure, due to dynamics of information and external events.

Accuracy is easiest to measure for structured data. Conventional data quality methods provide for tools that ran against the database to test data against business rules. This concept remains valuable and important.

Other quality metrics indicated by this Business Scenario are for:

- Trustworthiness
- Amount of meta-information
- Standardization of format

The value of these metrics is less clear. They are likely to be most useful in the context of a particular information quality policy, in showing how far the policy has actually been applied in practice.

The following requirements for the measurement process were identified during the Boston Workshop:

- It must require minimal internal development (ideally none). This implies that it is supported by vendor solutions, not user-developed code.
- Measurement must not influence the result.

Finally, measurement is not a process that should be applied blindly to information of all kinds. Every enterprise should consider what information is important enough to build metrics for.

Summary of Specific Requirements for Standards

Note that a statement of requirement in this section does not imply that no standard exists. Existing standards should be reviewed to determine how far they meet these requirements. New standards should not be developed unless it is clear that no existing standards meet the need.

Metadata

Standards are required for metadata. They may be industry-specific. They should include standards for names for information types, and enable the following to be represented:

- Purpose and usage of information
- Creators, modifiers, and circumstances of creation and modification
- Quality processes that have been applied

They must be applicable to information formats of all kinds, and be applicable to collections and structures of information, as well as to single items.

Tagging

Standards are required for information tagging in information management product interfaces. They should enable information to be tagged with meta-information in a uniform way, when it is created or modified.

The tagging process should be sensitive to the roles of the users, so that appropriate choices are presented.

Metrics

Standard metrics should be developed for:

- Time taken to find information, or to determine that it does not exist
- Accuracy
- Time taken to perform an operation, such as indexing, on information

The development of products that enable users to measure these quantities should be encouraged. These products should be usable with minimal or no configuration and coding by the users. Their use to take measurements must not affect the values of those measurements.

Next Steps

The next steps should be:

- Review the suitability of existing standards.
- Adopt, adapt, and integrate them to meet the requirements.
- Put in hand the development of any new standards needed to fill gaps.
- Communicate to vendors the need to implement the standards.

These steps should be taken by a new Forum of The Open Group. The immediate actions are to develop its charter and recruit members.

Glossary of Terms and Abbreviations

CAD	Computer-Aided Design
CIO	Chief Information Officer
Code	A unique, possibly arbitrary, alphanumeric identifier that represents the class or type of an item of information.
DBMS	DataBase Management System
DoD	Department of Defense (of the United States of America)
ECM	Enterprise Content Management
ERP	Enterprise Resource Planning
ETL	Extract, Transform, and Load
HR	Human Resources
IT	Information Technology
MIT	Massachusetts Institute of Technology
OMB	Office of Management and Budgets (US Government Department)
PC	Personal Computer
PDA	Personal Digital Assistant ("pocket computer")
PDM	Product Data Management
SQL	Structured Query Language
Tag	An indication, stored with an item of information, of the information's class, attribute, or type
TDQM	Total Data Quality Management
US	United States (of America)
XML	eXtensible Mark-up Language (defined by the Organization for Advancement of Structured Information Standards)
XSLT	eXtensible Stylesheet Language Transformation (World Wide Web Consortium standard for specifying transformations of information represented in XML)

References

Business Scenario: Interoperable Enterprise (K022), published by The Open Group, October 2002.

Data Protection Act 1998 (ISBN: 0105429988), HMSO (UK).

Data Quality Guidelines by Agency, Center for Regulatory Effectiveness (thecre.com/quality/agency-database.html).

Federal Data Quality Legislation, US Public Law 106-554, Section 515.

Government Records Access and Management Act (GRAMA), Utah Code 63-2 (2), 1992.

Health Insurance Portability and Accountability Act (HIPAA), US Public Law 104-191, 1996.

Kon H.B., Reddy M.P., Wang R.Y., Towards Data Quality: An Attributes-Based Approach, Decision Support Systems 13, published by Elsevier, 1995.

Lee Y.W., Pipino L.L., Wang R.Y., Data Quality Assessment, Communications of the ACM, Volume 45, No. 4, April 2002.

Sarbanes Oxley Act, US Public Law 107-204, 2002.

State Records Management Act, California Government Code, Sections 14740 *et seq*.

Loshin D., Kaufmann M., Enterprise Knowledge Management: The Data Quality Approach (ISBN: 0124558402), 2001.

White Paper: An Introduction to Boundaryless Information Flow (W201), published by The Open Group, July 2002.