

Data Categories for Spatial and Attribute Data

Byline

Janzen, Jeremy; Data Administrator; Information Management Group, Ministry of Forests, Box 9527, Stn Prov Govt, Victoria, British Columbia, V8W 9C3; <http://www.for.gov.bc.ca/isb/datadmin/>

Abstract

Every organization has information that it needs for its survival. Examples might be a list of clients and receivables, or a geographic inventory of the organization's resources (e.g. timber company: forest cover information).

But, any organization collects and uses more information than just that needed for company survival. Some of this information may measure the organization or its processes, to improve efficiency. Some may be extra information about clients that helps fine-tune the business relationship. There are degrees of how important information is to the organization, depending on the information.

This paper will show the categories of corporate information now being used in the BC Ministry of Forests for spatial and attribute data, and how they can be used to promote better business understanding. The general thrust of the material has an information management focus applicable to any agency, company, or organization, and is not specifically aimed at government or the forest sector.

Data Categories for Spatial and Attribute Data

Preliminary Concept

Data quality depends on definition

In any large organization, "data" is defined, collected, transformed, used, summarized, and reported for the purpose of making business decisions. The value of such data is commensurate with the amount of resources that went into its definition, collection, and use. Data that is collected in an ad-hoc manner with no standard method is of far less value than data collected from a defined business purpose that has been thought through and agreed on by all affected business staff.

For example, send five children out to a playground to "count weeds". The five answers received will depend on what each child thinks is a weed (one child might consider grass a weed; another might consider a dandelion a flower), how well they can count, whether they cover the whole playground, what the weather was like (they might quit early), etc. The value of any one response is low, since the collection parameters were not well defined. On the other hand, it's very easy to send a child out to count the weeds — you get a number back in short order.

This example illustrates the trade-off we often make in business, in the data that we collect and use. It's very easy for an organization's employee to simply go out and collect some information, but the

value of that information to others in the organization may be questionable. It is more difficult and takes considerable resources (training, communication, auditing...) to ensure that everyone in the organization is using the same standards to collect the same information. However, information resulting from a common standard is much more useful to the organization, because it is of known quality, can be shared easily, and can be relied upon when making business decisions.

The above possibilities indicate that it is useful to know what kind of data you are dealing with when trying to make a business decision — was the data collected in an ad-hoc manner, or was the data collected to a recognized and agreed to standard. Generally, the only thing separating these two “different realities of data collection” is the amount of resources (energy) put into the decision processes around the data’s definition, collection, and ongoing management.

Fundamental Concepts

Corporate Data

I formally define the term “corporate data” to distinguish what information is important to the organization. Data is corporate if it is:

- vital to the organization (i.e. critical to the organization's business);
- of a permanent or lasting nature (i.e. kept for a significant period of time);
- within the organization’s mandate (i.e. in government, the health department would not have a mandate over drivers’ licences; in private corporations, the “mandate” would be motivated by profit and decided by the executive).

If data meets all of the above criteria, as defined by a particular organization, then it is corporate data. The implication is that if it is either not important enough, or not stored long enough, or not within the organization’s mandate, then it is by definition, not corporate data (from the organization’s perspective). Note that the boundary between corporate and non-corporate data will often be simply a human decision.

Process Maturity

The following definition of process maturity is taken from Carnegie-Mellon’s software Capability Maturity Model (CMM), developed by the Software Engineering Institute [Paulk 93a, Olson 94]. The CMM defines process maturity as having five levels. In increasing level of quality management, those levels are briefly defined as:

Initial; characterized by *ad hoc processes*; resulting in unpredictable quality throughout the organization.

Repeatable; characterized by *disciplined use of processes*; resulting in stable planning and tracking of processes so earlier successes can be repeated.

Defined; characterized by *standard, consistent processes*; resulting in both engineering and management practices that are stable and repeatable.

Managed; characterized by *predictable processes*; resulting in measured processes operating within measurable limits and producing predictably high quality products.

Optimized; characterized by *continuously improving processes*; resulting in measurable improvements that are made by both incremental advances in existing processes and by innovations in technology and methods.

Data Maturity

The following definition of data maturity is taken from “Method for Establishment of Strategic Improvement Opportunities” [Friswell 95]. The authors defined data maturity as having four levels. In increasing level of quality management, those levels are briefly defined as:

Unstructured; characterized by *local ownership and use*; resulting in local or no definition; single physical source accessed through owner; and usage being contingent on the owner’s intention.

Uncontrolled; characterized by *many independent definitions*; resulting in many physical sources and owners with access through a selected owner; many users; and usage is still contingent on each owner’s intention.

Shareable; characterized by *common definition*; resulting in one logical source and shared access but not correlated to definitions of other objects; a designated Data Custodian; many users.

Integrateable; characterized by *common definition correlated to other objects*; resulting in one logical source and shared access to the data; a designated Data Custodian with identified relationships; many users; and data that is easily correlated to other data.

The Process/Data Energy Arrow

I have so far defined:

- Corporate data being data that is a) vital to the organization's interests and operations, b) permanent, and c) owned.
- The concept of process management being defined along an increasing continuum of maturity.
- The concept of information management being defined along an increasing continuum of maturity.

Putting the process and information management ideas together, see **Figure 1, “Maturity Costs Now for Benefits Later”**. The Y-axis shows the Process Maturity levels, and the X-axis shows the Information Maturity levels. The arrow shows the energy required to define and manage both the information important to an organization, and the processes that work on it.

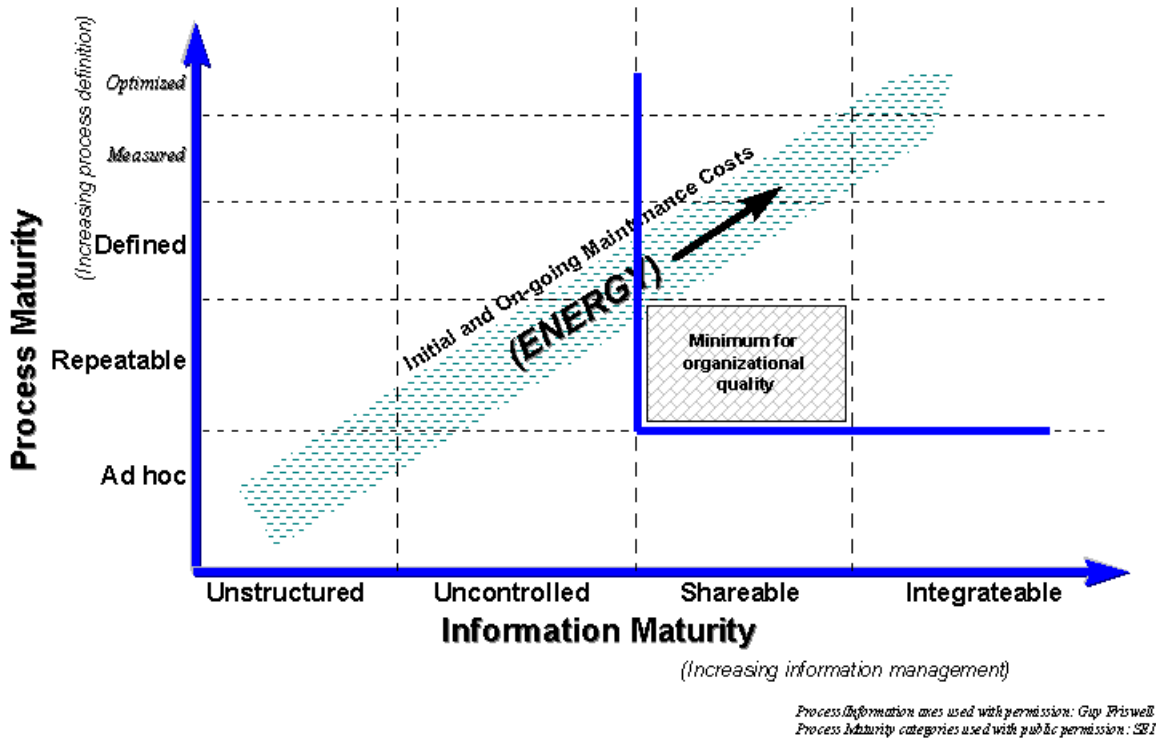


Cdcfig1.gif

[FIGURE 1: cdcfig1.gif]

FIGURE 1

Maturity Costs Now for Benefits Later



The ‘energy arrow’ goes from left to right, being highest on the right — indicating a large organizational expenditure of energy (i.e., human resources, financial resources, facilities usage, etc.) to get integrateable data and optimized processes. It takes a lot of organizational energy to come to a common understanding (and after understanding, agreement!) of a definition for something that is used organization-wide.

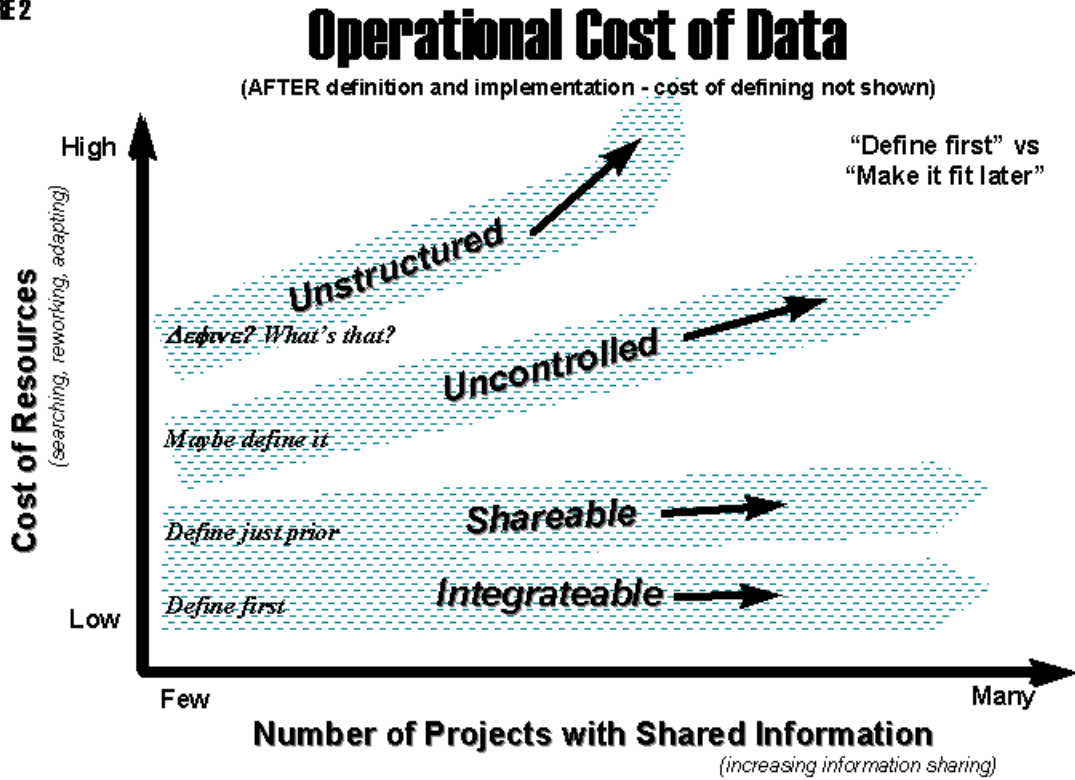
Maturity trades up-front costs for later benefits. The benefits are normally worth the cost in productivity, efficiency, and effectiveness. When the effort is put in at the front to come to common data definitions, staff from across the organization can use that data over and over in different ways, gaining multiple benefits with very little additional cost.

“Status quo” costs every organization as well. **Immaturity trades immediate gratification (“Just get me some data, now!”) for later confusion.** When common data definitions are not used, staff later spend an inordinate amount of time searching, puzzling, questioning, reworking, communicating, translating, adapting, transforming, transposing, and converting data collected by others that they think might be relevant to their use, again and again. (See **Figure 2, “Operational Cost of Data”**).



[FIGURE 2: cdcfig2.gif]

FIGURE 2



Corporate Data Categories

The point here is not that “all data must be fully defined across the organizational processes”, but rather that organizations should choose which information should be fully defined, and which can be left for individuals to start up on their own. No organization can afford to define and manage everything that is of interest completely and immediately. So choose *the most important* data and *the most important* processes to put organizational energy into.

A framework of data categories will help assess the state of an organization’s information — how much energy has been put into definition and management. Read **Figure 3, “Corporate Data Categories”** from left to right, increasing in maturity.

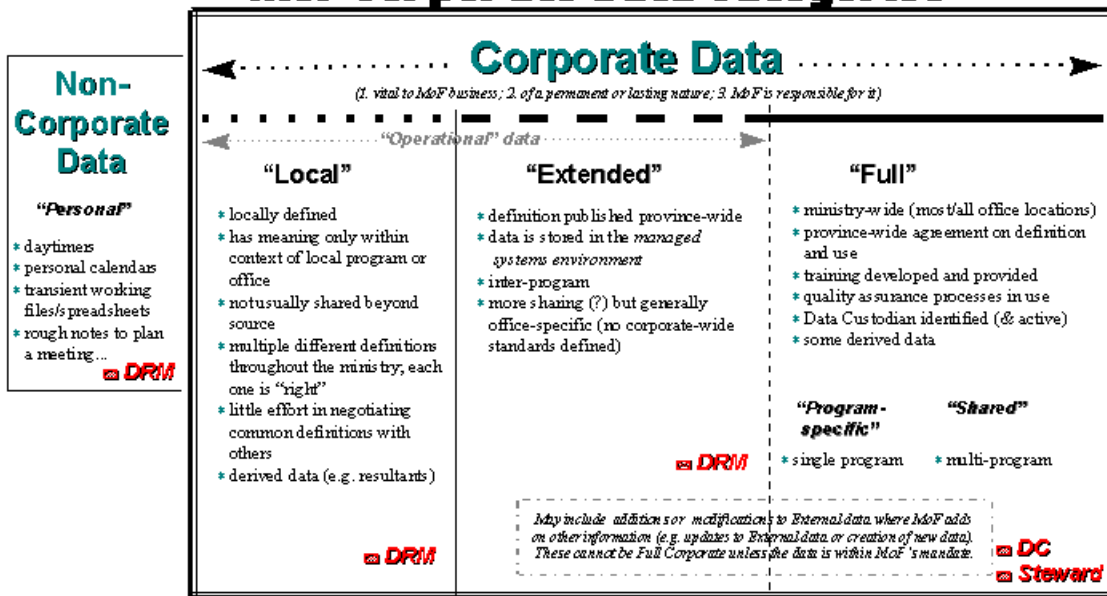


Cdcfig3a.gif

[FIGURE 3: cdcfig3.gif]

FIGURE 3

MoF Corporate Data Categories



Non-Corporate Data

- Produced by ad hoc, undocumented processes for quick results of unpredictable quality
- Examples: transient working files or spreadsheets, rough notes to plan a meeting, personal calendars

Local Corporate Data

- Locally defined and used for local (office) decision-making only
- Not measured or defined to recognised or published standards
- Similar information with different definitions or interpretations may exist elsewhere, and little effort has been expended in negotiating common definitions with others
- Produced by a repeatable, but not well defined, process
- Example: staff in an office agree to collect estimated area, location, and date for pest infestations if they happen to spot one while in the field — “10ha pest infestation 6km SW from Big Bear Creek, June 13/95” — but the accuracy of “area” and “location” will vary depending on who collects the data, and the definition is not correlated with other offices (i.e., another office may decide that “area” is not important, but that the “infestation intensity” should be collected).

Extended Corporate Data

- Data becomes Extended Corporate when its definition is published organization-wide, and the data itself is stored on a corporate platform with ongoing data management integrity (not on a personal workstation).
- Produced by a defined process (as part of the rigor of becoming Extended), but not measured or optimized; not used for organization-wide decision-making.

- This means that at least some energy has been put into an organization-wide definition. The definition may not be fully correlated to other definitions from other sources (i.e. there still may be multiple different definitions throughout the organization), but it's a start. The visibility of the definition to all staff encourages an informal negotiation process to begin (i.e. “Hey! Mine’s better than yours in this way, but you have this good idea here…”).
- An office may choose to publish their definition of a particular information need, or they may work together with another office(s) to agree on the definition. The more shared energy and negotiation that goes into the definition, the more mature and reusable the data is.
- Example: staff in an office agree to collect estimated area, damage, location, and date in the same way using the same codes — “bark beetle damage, 12ha, severity 9, location 56” — and publish their definition so that all other offices can see it (e.g. in an organization-wide discipline expert meeting; as an internally-published report; etc.). Note that storing the data on a managed platform (e.g. a Local Area Network server) is important to improve the data management integrity (regular backups, network accessibility, etc.) as well as the definition.

Full Corporate Data

- Common definition across the organization, led or authorized by a designated Data Custodian
- Collected by well-defined, standard, formal, and measurable processes, such that the data can be easily, reliably, and confidently linked to other Full Corporate data.
- Defined primary source, with normally shared access; consistently collected and used across the organization; ongoing training, support, and data management is provided.
- Recognised as permanent and vital, for both operational and strategic decision making — relied on to be accurate, meaningful, and available.
- Example: most organizations’ client information is a good example of organization-wide understanding and rules around definition, collection, and use.

Roles

There are two key roles that must be active in each organization to properly exploit the Corporate Data Categories. They are the Data Custodian, who defines and sets standards for data within their organizational mandate; and the Data Resource Manager, the generic title for anyone who collects or manages data within the organization. The specific titles are not important as long as their defined role is being carried out.

Data Custodian

- The senior executive or manager in an organization who establishes organization-wide policy, definitions, requirements, and rules for business information within their mandate, to enable the organization to gain maximum value out of Full Corporate information. The results must meet the entire organization’s needs, not just the Data Custodian’s own departmental needs. There may be multiple Data Custodians in an organization, but their individual information mandates must not overlap.

Data Resource Manager

- A generic title for someone who is responsible for collecting and/or managing corporate data (to the standards set by the Data Custodian). The most senior manager in each office is ultimately accountable for ensuring corporate data collection and management is done properly throughout the office, to enable effective business decisions. They are accountable and responsible for the quantity and quality of information utilized in their day to day administrative practices. The Data Resource Manager is a key role for the organization: where Full Corporate data does not exist, they will determine what data is important enough to define as Local Corporate, and when to begin moving it to Extended Corporate.

Final Message

The quality of data in any organization is directly dependent on the effort that is put into its definition and management. If the work is done up front across the organization to gain agreement on definitions and use of certain data, all staff can be confident about using and sharing that data. There is still some value in information that is defined by each office or department for their own use, as long as it is recognized that during cross-department operational projects, a huge amount of discussion, negotiation, translation, and conversion will be required before that data can be shared. The key for an organization is to choose which data is most important for company-wide sharing, then to define **that** Full Corporate data from a shared perspective while still using Local Corporate or Extended Corporate data as appropriate.

Acknowledgements

The author gratefully acknowledges the contributions of the following people. Without their insightful thought and theories, the concept of the Corporate Data Categories would never have made any sense.

The Capability-Maturity Model from the Software Engineering Institute at Carnegie Mellon University (see References) was invented specifically for software processes; I have taken the liberty of expanding the meaning of the CMM from its focus on software, to the general process design of an organization.

Information maturity was developed and graphed (together with process maturity) on X-Y axes by Guy Friswell and Gerry Moore (see References).

Richard Dzobia brought both these works to my attention and explained them, our conversations thus triggering ideas for the Corporate Data Categories.

And finally, John Ellis' original suggestion to develop an "Extended data" concept provided the impetus to start the whole thing rolling.

References

[Olson 94]

Olson, Timothy; Reizer, Neal; and Over, James. A Software Process Framework for the Capability Maturity Model (CMU/SEI-94-HB-01, ADA 285595). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1994. Available WWW:
<http://www.sei.cmu.edu/pub/documents/94.reports/pdf/hb01.94.pdf>.

[Paulk 93a]

Paulk, Mark C.; Curtis, Bill M.; and Chrissis, Mary Beth. Capability Maturity Model for Software, Version 1.1 (CMU/SEI-93-TR-024, ADA 2634034). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1993. Available WWW:
<http://www.sei.cmu.edu/pub/documents/93.reports/pdf/tr24.93.pdf>.

[Friswell 95]

Friswell, Guy; and Moore, Gerry. Method for Establishment of Strategic Improvement Opportunities. British Columbia Ministry of Transportation & Highways, 1995.

Biography

Jeremy Janzen, a designated Information Systems Professional of Canada (I.S.P.), has been actively working in the systems field for twenty-one years, the past sixteen concentrating on developing a data administration function and promoting information management within the British Columbia Forest Service (BC Ministry of Forests) in Canada. He has chaired the BC government-wide Data Administration Forum since its inception in 1996. Jeremy feels strongly about the importance of managing information as a corporate business resource, and that it is necessary to re-build how people think about information (a.k.a. "culture change") for organizations to be successful. Jeremy brings a practical rather than academic approach to spatial and attribute data management, with a focus on realizing short- and long-term gains and "implementing goodness"!