Data Governance at Stanford:

A Data Governance Overview

September 19, 2011

Contents

DG @ Stanford

What is Data Governance?

Data Governance (DG) is a cross-functional set of roles, policies and enabling technologies that work together to ensure that an organization is getting the maximum net benefit out of its data assets. To be both successful and sustainable, a DG program must be integrated with business and IT processes throughout the organization.

The fundamental purpose of all DG programs is *to improve the effectiveness and efficiency of business processes*. Critical business processes are especially sensitive to the quality of data, and the failure of such processes may have far-reaching impacts. Quality data that is "fit for use" across the organization can only be developed and maintained through the collaboration of a diverse set of data stakeholders. This group must commit to formalized responsibilities, policies and procedures around the effective management of data. All DG procedures, including data quality remediation and master data standardization, are most efficient and effective when they are understood and performed consistently throughout the institution.

> *Data Governance is formalizing behavior around the definition, production and usage of data to manage risk and improve the quality and usability of selected data.*
>
> -Robert Seiner, TDAN.

To achieve institution-wide commitment, the strategy and structure of each DG program must be designed with the unique priorities, competencies and goals of that organization in mind. This may lead to vastly different DG implementations, even among similar organizations. Additionally, as a DG program grows and technical and business environments change, its strategy and structure must continually evolve to remain effective.

High-quality data is a critical success factor across all functions of any organization. Proactive data management and a well-defined Data Governance program are required for the full value of this institutional asset to be realized.

Contact Data Governance

Matt Hoying

Data Governance Manager

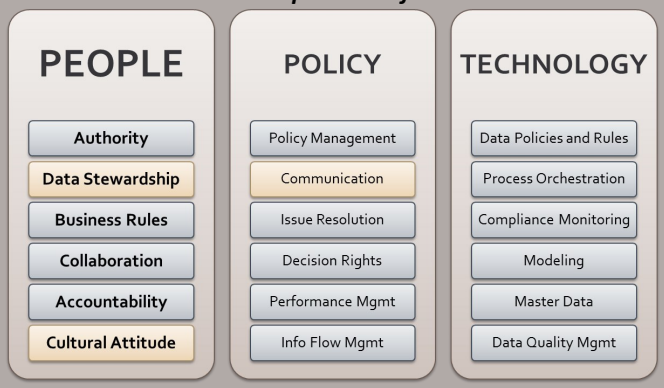
hoying@stanford.edu

(650) 736-8192

http://irds.stanford.edu/dg

Institutional Research and Decision Support

Dimensions and Components of Data Governance

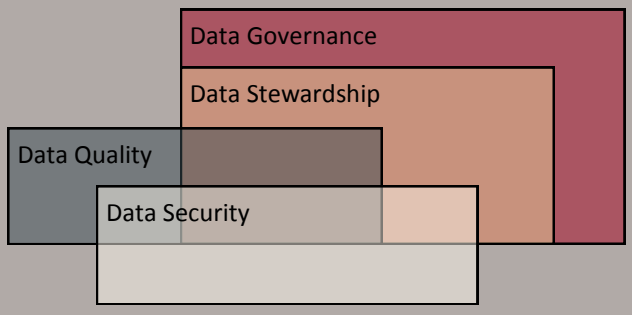


Components that we will initially focus on are highlighted: Data Stewardship, Cultural Attitude and Communication.

Benefits of Data Governance

- Reduced development time and cost
- Regulatory compliance
- Improved planning/forecasting
- Richer data environment
- More informed decision-making
- Improved operating efficiencies
- Tighter informational alignment between institutional functions

Relationships among DG, DS, DQ & DSec



Data Stewardship (DS) makes up a large part of Data Governance (DG) and is the primary enabler of Data Quality (DQ). The implementation of DQ may occur outside of the DG domain. Data Security (DSec) requirements drive DG polices and DS requirements; DSec compliance is an important aspect of DQ monitoring.

A Dual Approach to Data Governance

Many of the benefits of Data Governance (DG) are only realized in the long-term or through cross-functional projects that span the breadth of the institution and may not have any perceivable immediate benefit. Even some narrowly-focused data quality improvement efforts require a large investment of time and resources for metadata gathering and documentation. The complexity and breadth of these projects are often major obstacles in the rollout of any DG program. For this reason, we are planning to develop our DG program by performing two sets of activities in parallel: *project activities* and *foundation activities*.

Project Activities:

Data Governance will partner with projects that involve heavily shared data, explicit data integration efforts or those which have acknowledged data quality issues. We will look specifically for projects that incrementally advance our DG capabilities and expand upon available metadata. When a project has been identified, and with approval of the project management lead, DG will get involved as early as possible and perform such activities as:

- Profiling selected data
- Recommending data quality techniques
- Producing project-level information flows
- Capturing and storing metadata about key data

Foundation Activities:

In addition to project activities, DG will lead an ongoing effort to accumulate information enabling future DG, data stewardship and data quality initiatives. Beyond information about the business and technical details of selected institutional data, we will produce and maintain:

- Data Governance, Stewardship and Quality policies, procedures and best practices
- Data standards
- Institutional data dictionary
- Key data-, system- and institution-level information flows

Through this dual approach, we will be able to build a common understanding of institutional data across all levels of the organization while demonstrating tangible and quantitative value in the short-term through work with other BI and reporting projects.

DG Tools: The Information Flow

An information flow diagrams how selected data moves between people, files, applications and data stores within a defined scope. The content may additionally be stored for later use in a relational database. The scope of an information flow may be:

- A single element's lifecycle
- A specific project
- The entire institution

Contains

- Types of information moved
- Method of transport
- Frequency of transport
- Application/data store name and platform/environment
- Responsible parties

Enables

- Understanding of data dependencies
- Impact analysis
- Data quality research/ improvement
- Identification of stakeholders

DG Tools: The Data Dictionary

A data dictionary is a store of information about, and supporting the use of, an organization's data resources. The breadth and depth of this information can vary widely, depending on the organization's needs. The content (meaning) and structure (formatting) rules stored in the data dictionary should be resolved to a single institutional perspective when possible and each case where an institutional perspective cannot be established should be clearly documented.

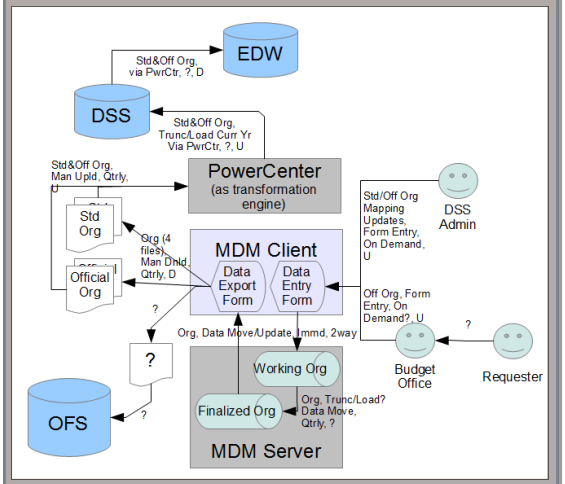
Contains

- Data element definitions
- Data element relationships
- Data element physical locations
- Data format specifications
- Recommendations for data usage

Enables

- Consistent understanding of data across the organization
- Institutional data consistency
- Shorter development cycle
- Improved integration capabilities

Information Flow



An example of an in-progress information flow that describes the movement of Stanford's Org Hierarchy Data.

Definitions

Attribute: A property of a data entity that may be used to distinguish or describe said entity.

Data governance (DG): "The formalization of behavior around the definition, production and usage of data to manage risk and improve the quality and usability of the selected data." (Robert Seiner, TDAN)

Data stewardship (DS): The formalization of accountability for the definition, usage and quality standards of specific data assets within a defined organizational scope.

Data entity: "A concept that can, or does, take on one or more values" (ISO 21961:2003). A data entity is a collection of information that represents a class of thing (such as "student" or "course"). In a relational database, this usually maps to a table and each instance of that entity maps to rows within that table.

Data element: An indivisible component of a data entity that has specific meaning. In a relational database, this usually maps to a column within a table and an attribute of an instance maps to a cell within that column.

Data profiling: Gathering statistics and information about a selected set of data or data store. Profiling data is a key activity for data quality improvement. Profiling may be done at the element, entity or system level.

Dimension: A collection of reference information that provides context to an event, object or concept. Common dimensions may be "time," "location" or "person."

Master data: Business-critical data that is highly shared across the organization. Usually codified data or data describing the structure of the organization or key data entities (such as "employee" or "course").

Metadata: Data that 1) describes other data and IT assets (such as databases, tables and applications) by relating essential business and technical information and 2) facilitates the consistent understanding of the characteristics and usage of data. Technical metadata describes data elements and other IT assets as well as their use, representation, context and interrelations. Business metadata answers the *who*, *what*, *where*, *when*, *why* and *how* for users of the data and other IT assets.

Reference data: Any type of data that is used to categorize or provide context for other data. Examples of reference data are "country" or "date."

Subject area: A grouping of information or concepts that represent a fundamental area of the organization and are logically related.