APAN PS5400: Managing Data

Week 6: Data Quality, Data Governance

Lecturer: Shekhar Pradhan

COLUMBIA UNIVERSITY
School of Professional Studies

Recap of last week

- Data Warehouses
- Schema-on-write
- Data Lakes
- Schema-on-read
- Data warehouses vs Data Lakes

This week

- Data Quality
- Data governance
- Data Policies

Data Quality

- It is commonly recognized that data is a resource as well as a product
  - Information and knowledge are higher level products
  - Many operations on data take “raw” data and create a data product
- As with any resource and product, it makes sense to talk about the quality of data
- Poor/inaccurate data can lead to major disasters
  - Discussion question: What are some actual ways in which poor data has led to bad decisions?
  - Discussion question: What are some possible ways in which poor data can lead to bad decisions?

What is Data/Information Quality

Some common definitions

- IQ is information that is ***fit for use*** by information consumers.
- IQ is the characteristic of information to consistently ***meet or exceed customer expectations***.
- Quality information is information that ***conforms to specifications*** **or** ***requirements*** of its consumers, producers, administrators.
- IQ is the characteristic of information to be of ***high value*** to its users.

DQ Dimension and Definition (Pipino 2002)

Dimension	Definition
Accessibility	Extent data is available/retrievable
Appropriate Amount of Data	Data volume is appropriate for its uses
Believability	Data is credible and believed true
Completeness	No data missing and sufficient in terms of breath & scope
Concise Representation	Data has compact representation
Consistent Representation	Data has consistent format
Ease-0f-Manipulation	Data can be applied to many tasks
Free-of-Error	Data is correct & reliable

DQ Dimensions & Definitions Continued

Dimension	Definition
Interpretability	Data is in clear definitions
Objectivity	Data is unbiased, unprejudiced & impartial
Relevancy	Data is applicable for task at hand
Reputation	Data source & content is highly regarded
Security	Access to data is appropriately restricted
Timeliness	Data is up-to-date for task at hand
Understandability	Data is easily comprehended
Value-Added	Data is beneficial and provides advantage from its use

DQ/IQ Framework

DQ Category	**DQ Dimension**
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Security, Ease of Operation
Contextual DQ	Relevancy, Valued-Added, Timeliness, Completeness, Amount of Data
Representational DQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation

Shekhar Pradhan's *Believability as an Information Quality Dimension* (https://pdfs.semanticscholar.org/e561/2a4e8ec5a7d0e1beb0800842c2385bac9179.pdf) argues that believability should be regarded as an intrinsic DQ.

DQ Evaluation

Data goes through a “data life cycle”. The four phases of this iterative cycle

1. collection, 2. organization, 3. presentation and 4. utilization

- The cycle may be repeated as desired to promote maintenance and continuous improvement of DQ.

Each stage creates is own data quality problems.

Quality problems occurring in earlier stages may cause further quality problems in later stages.

Examples of data life cycle artifacts

Definition	Applicable Examples
Data Collection	Raw data, Surveys, Observations, Recordings.
Data Organization	Data Files, Databases, Data Repositories
Data Presentation	Web pages, Reports, Account Statements
Data Application	Research Data, Medical Diagnosis Data

Each of these artifacts are subject to their own type of DQ issues.

Data Quality Roles

There are 3 main DQ roles identified by Strong et al.

- Data Producers (aka Collectors/Creators)
- Data Custodians (IT staff)
- Data Consumers

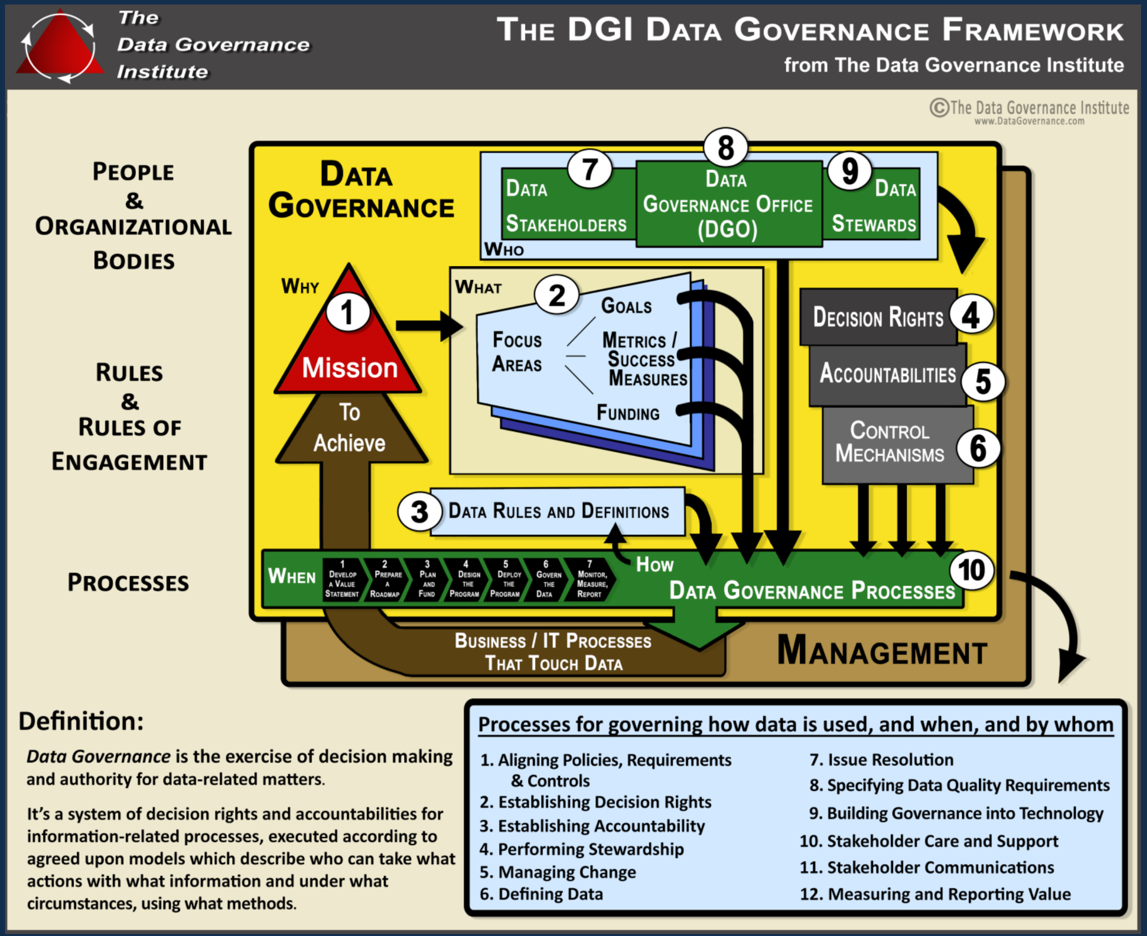
Also a Total Data Quality Manager (TDQM) is necessary

- According to Strong, Lee and Wang, the ***customers*** and not the custodians of data (such as IT departments or IT researchers or IT literature) should ***define and determine IQ***.

Data Governance

- Data quality does not just happen
- We need procedures to create and maintain data quality
  - Maybe different procedures for different stages of data life cycle
- We need roles that are assigned different rights and responsibilities with data
  - In other words, a form of data governance
- We need policies that specify which procedures are to be followed by whom for each stage of the data life cycle.

Data Governance


The Data Governance Institute
The DGI Data Governance Framework
from The Data Governance Institute
©The Data Governance Institute
www.DataGovernance.com
People & Organizational Bodies
Rules & Rules of Engagement
Processes
Data Governance
Why
1
Mission
To Achieve
What
2
Focus Areas
Goals
Metrics / Success Measures
Funding
Who
7
Data Stakeholders
8
Data Governance Office (DGO)
9
Data Stewards
Decision Rights
4
Accountabilities
5
Control Mechanisms
6
3
Data Rules and Definitions
When
1 Develop a Value Statement
2 Prepare a Roadmap
3 Plan and Fund
4 Design the Program
5 Deploy the Program
6 Govern the Data
7 Monitor, Measure, Report
How
Data Governance Processes
10
Business / IT Processes That Touch Data
Management

Definition:

Data Governance is the exercise of decision making and authority for data-related matters.

It's a system of decision rights and accountabilities for information-related processes, executed according to agreed upon models which describe who can take what actions with what information and under what circumstances, using what methods.

Processes for governing how data is used, and when, and by whom

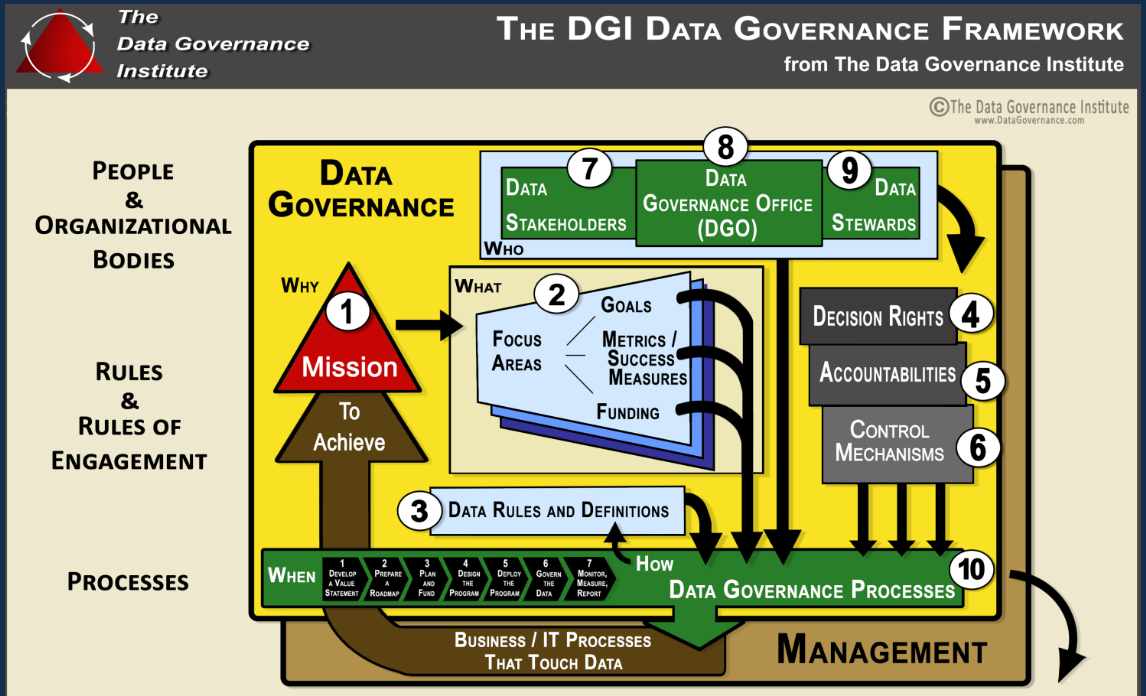
1. Aligning Policies, Requirements & Controls
2. Establishing Decision Rights
3. Establishing Accountability
4. Performing Stewardship
5. Managing Change
6. Defining Data
7. Issue Resolution
8. Specifying Data Quality Requirements
9. Building Governance into Technology
10. Stakeholder Care and Support
11. Stakeholder Communications
12. Measuring and Reporting Value

Rules and Rules of Engagement

1. Mission and Vision
2. Goals, Governance Metrics & Success Measures, & Funding Strategies
3. Data Rules and Definitions
4. Decision Rights
5. Accountabilities
6. Controls

http://www.datagovernance.com/fwk_dgi_data_governance_framework_components/

Data Governance


The Data Governance Institute
The DGI Data Governance Framework
from The Data Governance Institute
©The Data Governance Institute
www.DataGovernance.com
People & Organizational Bodies
Rules & Rules of Engagement
Processes
Data Governance
Why
1
Mission
To Achieve
What
2
Focus Areas
Goals
Metrics / Success Measures
Funding
Who
7
Data Stakeholders
8
Data Governance Office (DGO)
9
Data Stewards
Decision Rights
4
Accountabilities
5
Control Mechanisms
6
3
Data Rules and Definitions
When
1 Develop a Value Statement
2 Prepare a Roadmap
3 Plan and Fund
4 Design the Program
5 Deploy the Program
6 Govern the Data
7 Monitor, Measure, Report
How
Data Governance Processes
10
Business / IT Processes That Touch Data
Management

Definition:

Data Governance is the exercise of decision making and authority for data-related matters.

It's a system of decision rights and accountabilities for information-related processes, executed according to agreed upon models which describe who can take what actions with what information and under what circumstances, using what methods.

Processes for governing how data is used, and when, and by whom

1. Aligning Policies, Requirements & Controls
2. Establishing Decision Rights
3. Establishing Accountability
4. Performing Stewardship
5. Managing Change
6. Defining Data
7. Issue Resolution
8. Specifying Data Quality Requirements
9. Building Governance into Technology
10. Stakeholder Care and Support
11. Stakeholder Communications
12. Measuring and Reporting Value

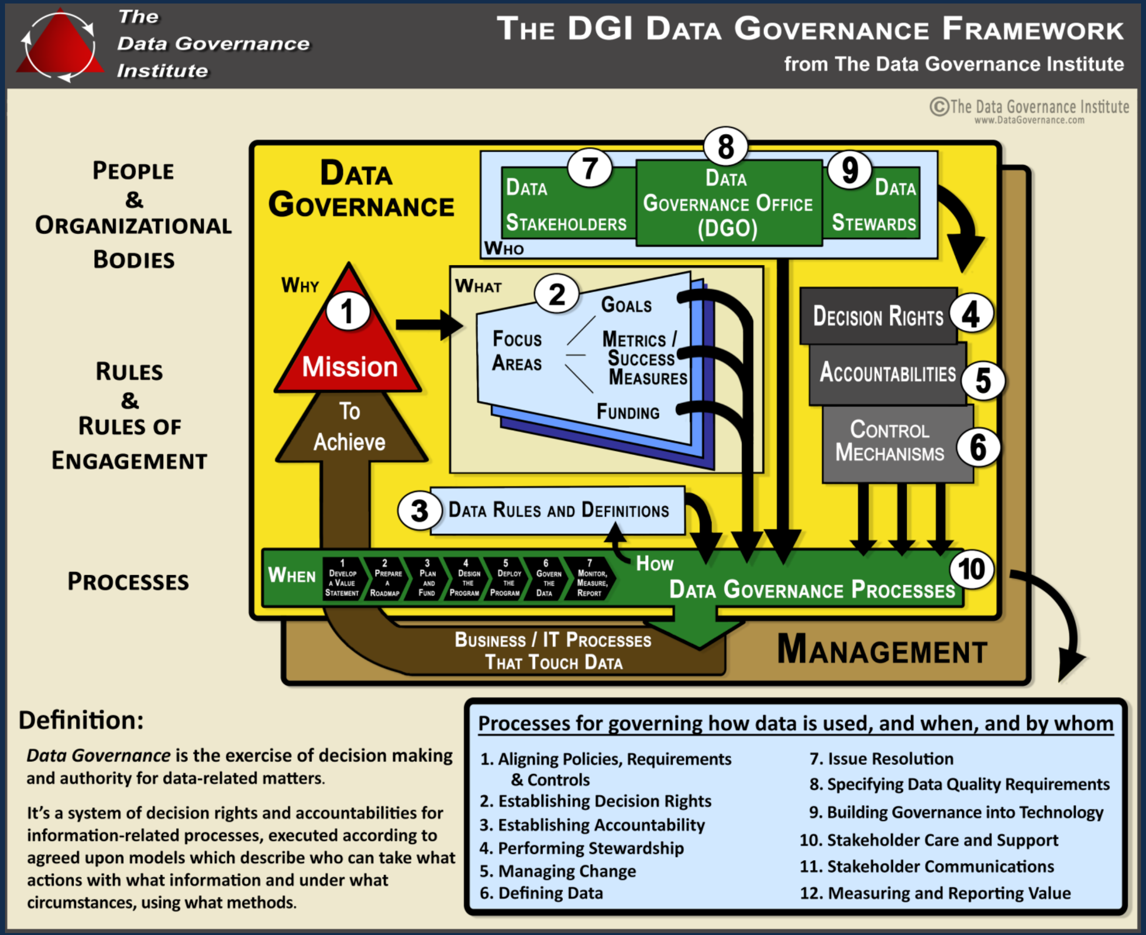
People and Organizational Bodies

7. Data Stakeholders
8. A Data Governance Office
9. Data Stewards

http://www.datagovernance.com/fwk_dgi_data_governance_framework_components/

Data Governance

Columbia University
School of Professional Studies


The Data Governance Institute
The DGI Data Governance Framework
from The Data Governance Institute
©The Data Governance Institute
www.DataGovernance.com
People & Organizational Bodies
Rules & Rules of Engagement
Processes
Data Governance
Who
7 Data Stakeholders
8 Data Governance Office (DGO)
9 Data Stewards
Why
1 Mission
To Achieve
What
2 Goals
Focus Areas
Metrics / Success Measures
Funding
4 Decision Rights
5 Accountabilities
6 Control Mechanisms
3 Data Rules and Definitions
When
1 Develop a Value Statement
2 Prepare a Roadmap
3 Plan and Fund
4 Design the Program
5 Deploy the Program
6 Govern the Data
7 Monitor, Measure, Report
How
10 Data Governance Processes
Business / IT Processes That Touch Data
Management

Definition:

Data Governance is the exercise of decision making and authority for data-related matters.

It's a system of decision rights and accountabilities for information-related processes, executed according to agreed upon models which describe who can take what actions with what information and under what circumstances, using what methods.

Processes for governing how data is used, and when, and by whom

1. Aligning Policies, Requirements & Controls
2. Establishing Decision Rights
3. Establishing Accountability
4. Performing Stewardship
5. Managing Change
6. Defining Data
7. Issue Resolution
8. Specifying Data Quality Requirements
9. Building Governance into Technology
10. Stakeholder Care and Support
11. Stakeholder Communications
12. Measuring and Reporting Value

Processes

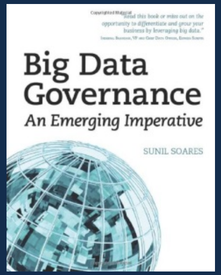
10. Proactive, Reactive, and Ongoing Data Governance Processes

http://www.datagovernance.com/fwk_dgi_data_governance_framework_components/

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



Organization for Big Data Governance

Metadata

Big Data Privacy

Big Data Quality

Business Process Integration

Master Data Integration

Managing the Big Data Lifecycle

Data Governance

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**

Process Steps

Organization for Big Data Governance

Metadata

Big Data Privacy

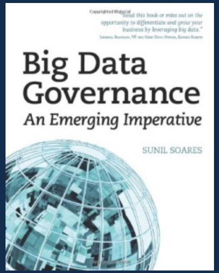
Big Data Quality

Business Process Integration

Master Data Integration

Managing the Big Data Lifecycle

Stakeholder. RACI Matrix

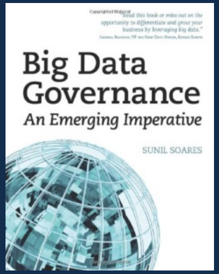


Big Data Governance
An Emerging Imperative
SUNIL SOARES

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



Organization for Big Data Governance

Stakeholder. RACI Matrix

Metadata

Glossary. Index. Search

Big Data Privacy

Big Data Quality

Business Process Integration

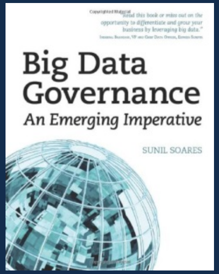
Master Data Integration

Managing the Big Data Lifecycle

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



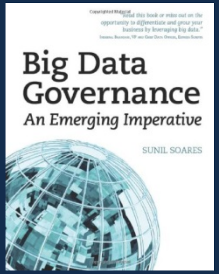
Big Data Governance
An Emerging Imperative
SUNIL SOARES

Organization for Big Data Governance	***Stakeholder. RACI Matrix***
Metadata	***Glossary. Index. Search***
Big Data Privacy	***Flag. Regulations. Monitor access***
Big Data Quality	
Business Process Integration	
Master Data Integration	
Managing the Big Data Lifecycle	

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



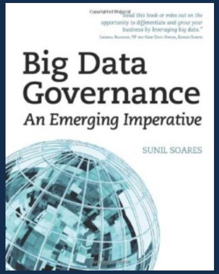
Big Data Governance
An Emerging Imperative
SUNIL SOARES

Organization for Big Data Governance	***Stakeholder. RACI Matrix***
Metadata	***Glossary. Index. Search***
Big Data Privacy	***Flag. Regulations. Monitor access***
Big Data Quality	***Business Confidence Measure***
Business Process Integration	
Master Data Integration	
Managing the Big Data Lifecycle	

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**

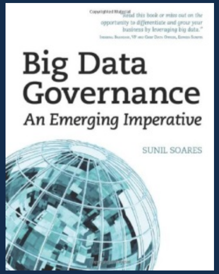


Organization for Big Data Governance	***Stakeholder. RACI Matrix***
Metadata	***Glossary. Index. Search***
Big Data Privacy	***Flag. Regulations. Monitor access***
Big Data Quality	***Business Confidence Measure***
Business Process Integration	***Policy to Process***
Master Data Integration	
Managing the Big Data Lifecycle	

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



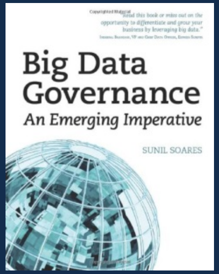
Big Data Governance
An Emerging Imperative
SUNIL SOARES

Organization for Big Data Governance	***Stakeholder. RACI Matrix***
Metadata	***Glossary. Index. Search***
Big Data Privacy	***Flag. Regulations. Monitor access***
Big Data Quality	***Business Confidence Measure***
Business Process Integration	***Policy to Process***
Master Data Integration	***Quality and Consistency***
Managing the Big Data Lifecycle	

Data Governance

Process Steps

Sunil Soars. **Big Data Governance: An Emerging Imperative. Oct 2012**



Big Data Governance
An Emerging Imperative
SUNIL SOARES

Organization for Big Data Governance	***Stakeholder. RACI Matrix***
Metadata	***Glossary. Index. Search***
Big Data Privacy	***Flag. Regulations. Monitor access***
Big Data Quality	***Business Confidence Measure***
Business Process Integration	***Policy to Process***
Master Data Integration	***Quality and Consistency***
Managing the Big Data Lifecycle	***Retention. Regulations. Usefulness. Disposal***

Data Governance

Process Steps – Short Form

Step 1: Get a governor and the right people in place to govern

Step 2: Survey your situation

Step 3: Develop a data-governance strategy

Step 4: Calculate the value of your data

Step 5: Calculate the probability of risk

Step 6: Monitor the efficacy of your controls

http://www.cio.com/article/2438861/enterprise-architecture/six-steps-to-data-governance-success.html

Data Governance

Example

Image: Mayr's

- HIPAA Privacy Rule
- HIPAA Security Rule
- HIPAA Enforcement Rule
- HIPAA Breach Notification Rule

HIPAA: Health Insurance Portability and Accountability Act
Wikipedia

GDPR

> The General Data Protection Regulation (GDPR) 2016/679 is a regulation in EU law on data protection and privacy for all individuals within the European Union and the European Economic Area. It also addresses the export of personal data outside the EU and EEA areas. Wikipedia

Every organization that collects information about any individuals that reside in the European Union must comply with GDPR or face large fines.

Do you think organizations require GDPR compliance policies and monitoring and enforcing mechanisms?

This week

- Data Quality
- Data governance
- Data Policies

Next week

- Interacting with data through
  - Web forms
  - Programs
  - APIs
  - Other channels (chat bots?)
- JSON objects as a data model