PRACTICAL DATA GOVERNANCE IN AN UNPREDICTABLE WORLD

Damon Feldman, Ph.D., Solutions Director, MarkLogic
Data governance (DG) refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise [specifying] who is accountable for accuracy, accessibility, consistency, completeness, and updating [and] how the data is to be stored, archived, backed up, and protected from mishaps, theft, or attack [and must ensure] compliance with government regulations.

http://searchdatamanagement.techtarget.com/definition/data-governance

- More process than technology
- But…
- Security
- Tracking
- Trust
- Quality
- Reconciliation
- Storage
- Auditing
- Compliance
What is *Practical* Data Governance?

- We are all performing Data Governance tasks
- But our reality may not match the Data Governance ideal for a while
Technology-enabling Data Governance – why?

- Avoid
  - Data breeches
  - Low-quality data of questionable provenance
  - Fines and censure

- Attain
  - Secure, uniform data access
  - Track and Trust
  - Simplicity and Data Governance project success
Data Governance Using the Data Hub Model
Keep Data Governance Governable

DUPLICATED ACROSS APPLICATIONS

- Call Center
- CRM
- Adjudication
- Dashboard
- Customer Portal

- OLTP
- WAREHOUSE
- DATA MARTS
- ARCHIVES
- REFERENCE DATA

- ETL
- LINEAGE
- SECURITY
- AUDIT

CENTRALIZED IN A DATA HUB / DATA LAYER

- Call Center
- CRM
- Adjudication
- Dashboard
- Customer Portal

- ETL
- LINEAGE
- SECURITY
- AUDIT

- DATA MASTERS
- DATA HUB
- DATA LAYER

- RETENTION/ARCHIVING
- QUALITY
- SECURITY
- LINEAGE
- AUDIT
- MASTERING
- COMPLIANCE
-UREMENT
- RULES
- ADJUSTMENT
Keep Data Governance Governable

**DUPLICATED ACROSS DATA STORES**

- Call Center
- CRM
- Adjudication
- Dashboard
- Customer Portal
- ETL
- OLTP
- WAREHOUSE
- DATA MARTS
- ARCHIVES
- Audit DB

**CENTRALIZED IN A DATA HUB / DATA LAYER**

- Call Center
- CRM
- Adjudication
- Dashboard
- Customer Portal
- ETL
- REFERENCE DATA
- SECURITY
- LINEAGE
- AUDIT
- SECURITY
- ETL
- Audit DB

Keep Data Governance Governable

**MIXED!**
- Call Center
- CRM
- Dashboard
- Customer Portal
- Adjudication
- ETL
- OLTP
- WAREHOUSE
- ARCHIVES
- DATA MARTS
- Audit DB
- SECURITY
- LINEAGE
- AUDIT

**CENTRALIZED IN A DATA HUB / DATA LAYER**
- Call Center
- CRM
- Adjudication
- Dashboard
- Customer Portal
- SECURITY
- LINEAGE
- AUDIT

[Diagram showing data flow and governance structures]
Data Governance Hubs

- A Data Governance Hub
  - ... vs. Data Governance within an overall silo-busting Hub.
Data Governance Functions
Data Lineage

- “I’m going to send you in for another MRI to rule out any physical abnormality”
  - Why? Imported CDA CCD data shows there an existing MRI
  - But simplistic, un-attributed data *without lineage* does not inspire confidence

MRI performed: YES
Location: HEAD
Abnormality: NO

Test Type: MRI
Sub-type: axial, T2-weighted MRI of the head and neck

Lineage:
Ordered by: Saul Yakenflaster, MD
Reviewed by: Saul Yakenflaster, MD
Imported:
LIMS Import System / August 22, 2014 / Rec no: LIMS-992-F47b
Quest Diagnostics / MRI-HN-9929331
Data Lineage Techniques in MarkLogic

- Store data as-is—keep the full incoming request
- Retain the batch control documents for batch import jobs
- Store the manifest if provided
- Store whatever “machine lint” was used to get your data into MarkLogic

```xml
<hl7:observation>
  <type>MRI</type>
  <sub-type>axial</sub-type>
  <weight>T2-weighte</weight>
<hl7:observation>
```
Data Lineage

- Clarify the source
- Conflicting information, or even basic information
- Helps read the data
- Helps trace or retract data

```
<loan loan-id="96583">
  <date>2010-02-22</date>
  <applicant>
    <name>
      <full>Robert Jones</full>
    </name>
    <address>123 Main st</address>
    <name>
      <full>Bob Jonas</full>
    </name>
    <address>PO Box 9922, Chicago, IL</address>
  </applicant>
</loan>
```
Data Lineage

- Lineage is metadata
  - Like annotations
- Per document or per field

<loan loan-id="96583">
  <date>2010-02-22</date>
  <applicant>
    <name source="loan-application" id="L-9322">
      <full>Robert Jones</full>
    </name>
    <address>123 Main st </address>
    <name source="cust-svc-phonecall">
      <full>Bob Jonas</full>
    </name>
    <address>PO Box 9922, Chicago, IL</address>
  </applicant>
</loan>
Data Lineage

- Semantic data is also a good approach
  - Requires reification of a fact to annotate it with a source

```xml
<loan loan-id="96583">
  <date>2010-02-22</date>
  <applicant>
    <name
      id="LOAN-142934">
      <full>Robert Jones</full>
    </name>
    <address>123 Main st</address>
    <name
      id="LOAN-142934">
      <full>Bob Jonas</full>
    </name>
    <address>PO Box 9922, Chicago, IL</address>
  </applicant>
  loan isA fin:loan
  hasName foaf:name#BobJones.
  foaf:name#BobJones
  hasLable "Bob Jones"
  fromSourceType fin:loanApplication.
```
Relational Data Lineage

- Metadata about batch processes
  - Build another data model repeating key elements of the batch system
- Lineage or source data per element
  - Augment the core data model with source fields
  - Promulgate through all relevant tables
  - Update all relevant inserts
- Other data about movement, sources, changes
  - More data models

**Remember** – your data has already been modeled. Just store it.
Data Lineage Usage

- Show it to the user
- Trace quality issues to a source, person or process
- Forensics
- Connect the data dots for retraction, cleanup

Test Type: MRI
Sub-type: axial, T2-weighted MRI of the head and neck

Lineage:
Ordered by: Saul Yakenflaster, MD
Reviewed by: Saul Yakenflaster, MD
Imported:
LIMS Import System / August 22, 2014
Quest Diagnostics / MRI-HN-9929331
Data Lineage for Compliance

- **Compliance**
  - US EO 12333 – delete citizen data with no intelligence value
  - EU ePrivacy directive – regulate and delete transfers of certain types
  - Data Stewards are often acting on a System of Record. Retractions must be traced to downstream storage
  - BCBS 239 Principles for Effective Risk Data Aggregation and Risk Reporting
Auditing

- Who accessed the data, and when?
- How did data change over time?
- Who changed it?
- What errors did we encounter? Fix?
Auditing

A national HMO sent an explanation of benefits to an unauthorized family member. [HHS Office of Civil Rights] found that a flaw in the health plan's computer system put the protected health information of about 2,000 families at risk and required the insurer to, **review all transactions for a six month period, and correct all corrupted patient information.**

- Who was affected?
- How can you review all transactions of this type?
HealthCare.gov Data Cleanup

- All data quality activities are implemented by a “cleanup” task
  - The old and new data is stored (as-is)
  - The cleanup ticket and date are logged
  - A cleanup library requires this data and stores it in a standard format
- Cleanups are exposed as links on internal data browsing tools
  - But not to the consumers – they generally don’t need to see that their record was corrected
Auditing – Dated Records

- Easier with Documents
  - Add a date or date range to an entire document
- MarkLogic support with Temporal Documents
  - Immutable

```json
{"loan":{
  "loan-id": "96583",
  "date": "2010-02-22",
  "applicant": {
    "name": "Bob Jones",
    "address": "123 Main St"
  },
  "status": "approved",
  "issues": [
    {"credit-risk-issue": "resolved"}
  ]
}}
```
Tracking Historic Data

- Easier with Documents and Triples
  - Add a date or date range to an entire document
- Use Temporal Documents
  - Immutable
- May be for display and user navigation, or for audit, history, reconciliation
Relational Auditing

- Build a new schema for audits (again)
- Object or service layer, possible data mappings
- Join tables and keys for any relation to some business entity

- Audits vary widely
  - Neither agile nor efficient
Auditing Key Points in MarkLogic

- Data Cleanup
  - Log changed data – immediately – regardless of size or location or type
- Audit records
  - Writing “as-is” is a huge benefit
  - Compliance and best practice requires they are written
- Logs too
  - MarkLogic audits data access, security access, changes and other activities
Data Quality

- Data Quality Rules can spot bad data before it becomes a larger problem
  - HealthCare.gov State Medicaid agencies submitted data with parents and children reversed
    - Rule: Parent age must be > Child age
  - Invalid addresses are common
- Human-in-the-loop data
  - checking and correction
Data Quality in MarkLogic

- Uniform data allows uniform rules
  - Same fields and structures
  - Uniform technology (.xsd, .sch, .js etc.)
- Manage rules in one place
- Hub approach also gathers more data together
- Good audits, history, lineage allow more checks
  - Or better forensics
- Trace issues to their source
Tracing and Lineage Use Case

- On a critical, secure Government system, MarkLogic keeps multiple copies of all data
  - **Level 0**
    - Source data in source format
    - Text messages
    - Binary Word documents
    - Binary MySQL dumps
  - **Level 1**
    - Loss-less version of source data
  - **Level 2**
    - Harmonized/canonical version of the data in the target system desired formats
Security

Obama accepted OPM head's resignation

Silk Road 2 Hacked: Entire Bitcoin Wallet Drained, $2.7 Million Stolen

Qatar National Bank allegedly hacked, data of 1,200 entities leaked
  The data breach includes the personal bank accounts and passwords for Al Jazeera employees and members of the ruling al-Thani family
Security

Recall the “re-implement everywhere” approach

- Security is a horizontal concern
- Maintenance is a nightmare
- Secure, safe data access is only possible through some application
- The data layer itself is a security risk
- Beware the “Data Lake”
Security

- MarkLogic has best in class security
  - Common Criteria Certification (as a product)
  - Multiple Intelligence Community and DoD ATOs (for specific systems)
  - FIPS 140 encryption
- Role-Based Access Control
  - Users can have many roles
  - Documents and collections have permissions configured per-role
- Compartment security
  - For checking more complex AND/OR combinations of roles
Best Practice – Reconciliation

- Key quality metric… do I have all the data?
- No data movement process is 100% reliable
  - Just as there is no software with zero defects
- Periodically check that data sets are in sync
  - Four typical approaches
    - Periodic check of total counts
    - Manifest-based checks of batch counts
    - Periodic check of per-item counts
    - More detailed check: item+hash or item+date
OTHER MARKLOGIC FEATURES FOR DATA GOVERNANCE
MarkLogic Capabilities

- We have covered:
  - Consolidated governance functions
  - Security
  - Store as-is – For audits, raw data, lineage
  - Document models for dated or historic data
- Other items to be aware of…
Temporal Documents

- Track data according to both system time and business time
- Reconstruct comprehensive temporal picture as of any timestamp
- Works with historic data and auditing
Comprehensive Data

- Structured Data – XML, JSON
  - Or versions of relational
- Text, unstructured data
- Binary data
- Semantic data
- Govern any or all data in one system
Semantics

- Data Governance employs data rules
- Compliance is complex and can be represented as ontologies and rules
- Links and associations are often needed
- Semantic updates are light and fast
  - Ideal for “annotating” existing records
Data Governance for a Data Hub

- We all perform “governance” all the time
- Governance with the integrated data
- MarkLogic is designed for data governance
  - Consolidate and Govern
  - Track history, lineage audits
  - Temporal data
  - Validation and quality
Q&A