Effective Audit Trail of Data With PROV-O

Scott Henninger, Senior Consultant, MarkLogic
Operationalizing the Metadata

Data Governance
Quality management

Provenance Models
Metadata description

Provenance Dimensions
Technical perspective
Strategy and Execution

- **Shared**: Exchange of data between different departments is possible
- **Reliable**: Source has competence in the field of interest
- **Accurate**: All accounting events are correct in value and description
- **Current**: Data is up-to-date for the world it models
**Information Chain**

- **Create**: Generate new data entities or update their state
- **Derive**: Create value from more contributing data entities
- **Analyze**: Inspect data to discover new useful information
- **Report**: Submission of summary data as evidence of events
Data provenance documents the inputs, entities, systems, and processes that influence data of interest, providing a historical record of the data and its origins.

Data lineage includes the data's origins, what happens to it and where it moves over time.
“If the benefits of provenance are so well understood, why don’t more firms recognize it as a priority?”

What makes it difficult

- **Location**
  Data is spread across different systems in different organizational silos

- **Ownership**
  Lack of mature data governance makes the challenge of data lineage even more daunting

- **Spreadsheets**
  Business processes run outside of data management processes

Unforeseen costs

- **Compliance risk**
  The business gets exposed to difficult contract negotiations which can incur additional data costs

- **Redundant data activities**
  Duplicate controls are performed in different departments several times

- **Accuracy of analytics**
  Impossible to verify why models result in sub-optimal outcomes
Operationalizing the Metadata

Data Governance
Quality management

Provenance Models
Metadata description

Provenance Dimensions
Technical perspective
Metadata Repository

**PROVENANCE MODEL**

**LAZY**
Complex technique for reasoning

**EAGER**
Derived directly from output database

---

างการเพิ่มเติมเมื่อพร้อม...
PROVENANCE MODEL

Provenance Storage

**Envelope Pattern**
Provenance stored with data

```xml
<envelope>
  <provenance>
    <sem:triple>
      <sem:subject>/doc/id_a12a3.xml</sem:subject>
      <sem:predicate>http://www.w3.org/ns/prov#wasGeneratedBy</sem:predicate>
      <sem:object>/xform2016-07-20</sem:object>
    </sem:triple>
    <sem:triple>
      <sem:subject>/CanonicalTransform2016-07-20</sem:subject>
      <sem:predicate>http://www.w3.org/ns/prov#endedAtTime</sem:predicate>
      <sem:object datatype="http://www.w3.org/2001/XMLSchema#dateTime">2016-07-20T12:01:42.987</sem:object>
    </sem:triple>
    ...
  </provenance>
  <content>
    <doc-id>a12a3</doc-id>
    <workflowStatus>Draft</workflowStatus>
    <version>2.3</version>
    ...
  </content>
</envelope>
```

**Separate Database**
Large provenance payloads stored with reference to data

**Content Database**
```
<doc-id>a12a3</doc-id>
<workflowStatus>Draft</workflowStatus>
<version>2.3</version>
...
```

**Provenance Database**
```
<provenance>
  <sem:subject>/doc/id_a12a3.xml</sem:subject>
  <sem:predicate>wasGeneratedBy</sem:predicate>
  <sem:object>/xform2016-07-20</sem:object>
</provenance>
```
**Entity**: a trade, order, document, or other kind of entity, physical, digital or conceptual with some fixed aspects

**Activity**: something that occurs over a period of time and acts upon or with entities, such as creating, consuming, transforming, modifying, etc.

**Agent**: the business line responsible for an activity taking place, for the existence of an entity

W3C standard, circa 2013: https://www.w3.org/TR/prov-o/
PROVENARIO MODEL

Encoding Specification

**PROV-O**
- Reason on provenance data
- Specialized properties
- Model-based extensions of the standard

**PROV-XML**
- Types and elements are reusable

```xml
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix ex: <http://example.com/> .
:geneSequencing a prov:Activity;
  prov:startedAtTime "2012-04-25T01:30:00Z";
  prov:used :drosophilaSample-84;
  prov:wasAssociatedWith :lab-technician-GH-32.
:drosophilaSample-84 a prov:Entity;
  prov:wasAttributedTo :lab-technician-FE-56.
:lab-technician-GH-32 a prov:Agent .
```
EFFECTIVE AUDIT TRAIL WITH PROV-O

Operationalizing the Metadata

Data Governance
Quality management

Provenance Models
Metadata description

Provenance Dimensions
Technical perspective
Provenance Dimensions

Content, Use, and Management

- Non mutually exclusive
- Who endorses the information
- How a decision is made
- User consumption of provenance
- What was considered for that decision
Scenario

…an investment bank is implementing new regulatory reporting defined by CFTC, that will provide more information on their trading activities (extended type of financial products and pieces of data) in a shorter time frame (near real-time publication) with higher complexity of the rules determining who has the obligation to deliver the information.
### PROVENANCE CONTENT

**Attribution**

- What the provenance is about
- Sources used to create new result
- Process that yielded the artifact

```xml
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix : <http://example.com#> .

:TransactionReport a prov:Entity;
    prov:generatedAtTime "2017-04-12T12:12:12" ;
    prov:wasDerivedFrom :TransactionA ;
    prov:wasGeneratedBy :ReportGen .

:ReportGen a prov:Activity;
    prov:used prov:TransactionA ;
    prov:used :Venue1 ;
    prov:wasAssociatedWith :Msma .

:TransactionA a prov:Entity
    prov:wasAttributedTo :Murex

:Venue1 a prov:Entity .

:Murex a prov:Agent, prov:SoftwareAgent .
:Msma a prov:Agent, prov:Organization ;
```
Amendments are incorporated in the trade

Different aspects of the same trade linked together

@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix : <http://example.com#> .

:Transaction1 a prov:Entity.

:Transaction2 a prov:Entity;
  prov:wasRevisionOf :Transaction1.

:TransactionReport1 a prov:Entity;
  prov:wasDerivedFrom :Transaction1.

:TransactionReport2 a prov:Entity;
  prov:wasDerivedFrom :Transaction2.

:PostTradeReport a prov:Entity;
  prov:generatedAtTime "2017-04-12T12:12:14";
  prov:wasDerivedFrom :Transaction2;
Bitemporal Timelines

- Two temporal axes to maintain the business valid and the system times

```json
{ "transaction":
  {
    "system-start": "2014-11-19T11:00:00",
    "valid-start": "2014-11-21T12:00:00",
    "trader": "12XL9A",
    "price": 12
  }
}

{ "trader":
  {
    "system-start": "2014-11-19T11:00:00",
    "valid-start": "2014-11-20T12:00:00",
    "id": "12XL9A",
    "name": "John"
  }
}
```
PROVENANCE DIMENSIONS

Content
PROVENANCE USE

Understanding

- ... what was the trading and reference data used to generate this transaction report ...
- ... why is there a difference between the transaction report and the post trade report ...
- ... were there any changes in reference data at the time the correction was sent ...
Compliance

- ...which department provided the trade data and when was the booking done ...
- ... what transactions were not reported in the time required, and for what reasons ...
- ... are any transactions that should have been reported for new versions of rules ...
- ... have traders complied with rules ...
PROVENANCE USE

Debugging

- ... where did an error occur in a specific data field ...
- ... was the notification for the post trade publisher sent for that specific trade ...
- ... what version of data extraction rules were used when the transaction report was created ...
- ... what is the percent of reportable transactions from the daily volume ...
Trusting Data Sources

- Forward-Looking provenance
  - Anticipating problems given provenance information from other systems

- Analysis may find some sources/transformations/ETLs are troublesome
  - ...sometimes in specific contexts, such as high load rates, etc.

- Look for alternatives when designing future efforts
- Target troublesome processes for future refactoring efforts
PROVENANCE MANAGEMENT

Publication

HOW TO CONSUME PROVENANCE?

- Access
- Locate
- Query
Dissemination

- **Security**: secure HTTP should be used across unsecured networks; authentication should be enforced
- **Access control**: provenance information should follow the same access control rules as the resources
- **Bundle**: care is needed to ensure that the integrity of provenance is maintained
solid compliance architecture

- ALL of your data and metadata
- Complete track of data changes
- Full query composability
- Security, publishing, monitoring, etc.
- Fewer tools and processes to manage
Questions?