Toward a Functional Model of Data Provenance

George Komatsoulis

25-October-2004

Data Provenance: What do we mean?

- Dictionary Definition
  - Place of origin; derivation.
  - Proof of authenticity or of past ownership. Used of art works and antiques.
- The “6 W’s” Plus
  - Who, What, When, Where, Why, How
  - Chain of custody

Data Provenance: Characteristics

- Provenance attaches to individual assertions in a record.
  - Many value added databases (GenBank, SwissProt, etc) and objects that model this data agglomerate data that are derived from different sources.
- A single assertion may have more than one provenance associated with it.
  - Consider the assertion that a sequence is expressed in a particular tissue. This could result from a northern blot, EST tissue determinations, or a combination of the above.
- Provenance is metadata
  - But, since provenance is concerned with how data ‘migrated’ from one form to another; the old adage about ‘my data is your metadata’ is particularly true.

Why Does Data Provenance Matter?

Data Reuse:

- One of the primary purposes of large scale databases and repositories is the ability to leverage information to answer questions not posed by the person who originally collected the data.
- To evaluate data’s suitability for reuse, it is necessary to understand the details of its collection.
- Concrete example: Relative Expression Measurements
  - To reuse the expression levels it is essential to know most of the data contained in the MIAME model.

Data Reliability

- Data produced from different sources and by different methods vary in the degree of real (or perceived) reliability
- Data that has been transformed multiple times is more likely to have been incorrectly transformed (the Fax machine problem)
- Data that has been transformed many times is more likely to lose an important context element
  - Part of the problem identified in the decision making process for the reentry of the Space Shuttle Columbia
  - As data moved up the chain of command, important caveats to the analysis results were lost
- Concrete Example: Gene-to-Genome Location mapping
  - Source of a genomic location might be Golden Path, Affymetrix or other source.
  - Determination of source essential to determining confidence in that location, and correcting errors if one source found to be incorrect.

Data Confidence

- A quantitative measure of how reliable ‘we’ think that any arbitrary data is.
- Could be related to provenance information, or determined from other data properties.
- Ideally, use this information to select or exclude certain data from analyses
- Examples:
  - Restrict searches to expression data where chips had suitable gross statistical properties
  - Only use SNPs that have been independently identified by multiple methods
- Needs to be attached to primary record for searching (i.e. it cannot be buried in a hierarchical stack).

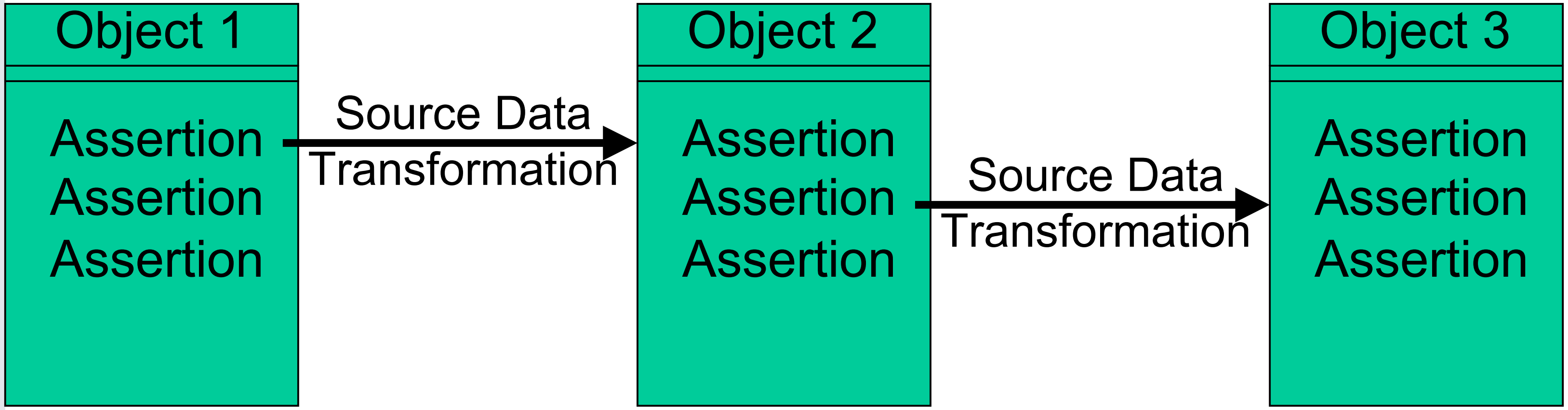
Provenance Models

Possible Provenance Models:

- Data Specific Provenance Model
  - Each data type has its own provenance model, carrying forward information covering the complete path of the data.
  - Advantage:
    - All provenance metadata comes with each result.
  - Disadvantages:
    - All provenance metadata comes with each result.
    - Each type of data/service has its own provenance model
- Generic Complete Provenance Model
  - Provenance Model consists of retaining provenance information in the form of prior data sets and transformations (CHIMERA is a Grid Instance)
  - Advantages:
    - All provenance metadata comes with each result.
    - Model is generic.
  - Disadvantages:
    - Requires storage of intermediate results
    - Model is sufficiently generic that it does not lend itself to simple visualization or analysis.

Hierarchical Provenance Model

- An alternative would be a hierarchical provenance model. In a hierarchical model, a result would only have provenance information that covered the previous transformation.
- An option would be to return a 'heavyweight' provenance object that recursively returns all provenance information. This should be user selectable since it will require more time



Example: EBI Protein Record

- A protein record from the EBI asserts that a turn exists from residues 102-105.
- EBI obtained this information from PDB
- Provenance object lists original source as Protein Data Bank, with links to original data (a protein structure file) at PDB.
- EBI does not supply original data; only information on how to get to the original data and how they used that original data.
- If additional information is needed, retrieve original information from PDB and study its provenance metadata.

Data Provenance: A Straw Man Model Proposal

- Unique Identifier: An identification uniquely associated with this data object and assertion
- Generating Source: The original source of an assertion
- Immediate Source: Where the information actually came from
- Number of Transformations: i.e. How many hops from Generating Source to this instance.
- Transformation: How was the data manipulated between the the immediate source and the current data object
- Reference: A reference to an electronic means of obtaining the original information (where possible) from the immediate source. Might be a URI, an RMI call, a Grid call, etc. Evaluating the reference should return a domain object of some kind; either a physical object or an XML representation of a domain object.
- Evidence Code: A controlled vocabulary term describing the type of evidence for the assertion.

Structure of the provenance metadata

- For data retrieved as XML (SOAP, HTTP) the provenance metadata should be returned as an XML provenance object that contains instructions for retrieving the original data with its provenance metadata.
- For data retrieved through an RMI method, provenance information should be returned as one or more provenance objects that contain references that would allow instantiation of domain objects.
- In this model, there should be no difficulty consuming the returned metadata because it is in the form of domain objects that have (hopefully) already been registered in the caDSR.
- The end of the trail is a provenance object that contains no references to additional data, either because it is the original source or because there is no additional provenance information.

Example 2: An Expression Change

```
<expressionRatio>
   <uniqueID>NCICB-20041005-1234-ABC</uniqueID>
   <foldChange assertion=1>5.6</foldChange>
   <basalTissue assertion=2>Normal Brain</basalTissue>
   <testTissue assertion=3>Glioblastoma</testTissue>
   <basalExpression assertion=4>1.0</basalExpression>
   <testExpression assertion=5>5.6</testExpression>
   <provenanceRecord>
       <assertion>2,4</assertion>
       <generatingSource>Caltech</generatingSource>
       <immediateSource>NCICB</immediateSource>
       <transformation>Normalization</transformation>
       <reference>http://someurl.cgi?id=NCICB-20041005-123</reference>
       <evidence>EV-Exp-TAS</evidence>
   </provenanceRecord>
   <provenanceRecord>
       <assertion>3,5</assertion>
       <generatingSource>Cornell</generatingSource>
       <immediateSource>NCICB</immediateSource>
       <transformation>Normalization</transformation>
       <reference>http://someurl.cgi?id=NCICB-20041005-124</reference>
       <evidence>EV-Exp-TAS</evidence>
   </provenanceRecord>
</expressionRatio>
```

Example 2: Continued

```
<arrayRecord>
    <uniqueID>NCICB-20041005-123</uniqueID>
    <tissueSource assertion=1>Glioblastoma</tissueSource>
    <patientAge assertion=2>17</patientAge>
    <prepMethod assertion=3>total polyA mRNA</prepMethod>
    <rawExpressionLevel assertion=4>2345.2</rawExpressionLevel>
    <provenanceRecord>
       <assertion>1-4</assertion>
       <generatingSource>Caltech</gereratingSource>
        <immediateSource>Caltech</immediateSource>
       <transformation>Original Record</transformation>
       <evidence>EV-AS-TAS</evidence>
    </provenanceRecord>
<arrayRecord>
```

Evidence Ontology: GO Proposal

- IC: Inferred by Curator
- IEA: Inferred by Electronic Annotation
- IEP: Inferred by Expression Pattern
- IGI: Inferred from Genetic Interaction
- IMP: Inferred from Mutant Phenotype
- IPI: Inferred from Physical Interaction
- ISS: Inferred from Sequence or Structural Similarity
- NAS: Non-traceable Author Statement
- TAS: Traceable Author Statement
- ND: No Data (for ‘Unknown’ Annotations
- NR: Not Recorded (for ‘Legacy’ Annotations)

Evidence Codes: Karp Ontology

- EV-Comp: Inferred from Computational Analysis
  - EV-Comp-HInf: Inferred by Human based on Computational Inference
  - EV-Comp-AInf: Inferred Computationally Without Human Oversight (Automated Inference)
- EV-Exp: Inferred from Experiment
  - EV-Exp-IPI: Inferred from Physical Interaction
  - EV-Exp-IMP: Inferred from Mutant Phenotype
  - EV-Exp-IGI: Inferred from Genetic Interaction
  - EV-Exp-IEP: Inferred from Expression Analysis
  - EV-Exp-IDA: Inferred from Direct Assay
- EV-IC: Inferred by Curator
- EV-AS: Author Statement
  - EV-AS-TAS: Traceable Author Statement
  - EV-AS-NAS: Non-traceable Author Statement