

# Survey of Data Provenance Techniques

Yogesh L. Simmhan

Beth Plale

Dennis Gannon

## Data Provenance

- *Derivation History of Data starting from its original sources*
- Data: Files, tables, tuples, virtual collections
- Derivation: Process that transforms data - Script, Web service, Queries, Commands
- Lineage, Pedigree, Genealogy, Filiation, Parentage, ...

## Contents of Talk

- Background
- Taxonomy for Provenance Techniques
- Survey of Provenance Techniques
- Conclusion

2005-03-03T17:00-05:00

CS System Seminar Talk

3

## Contents of Talk

- • **Background**
- Taxonomy for Provenance Techniques
- Survey of Provenance Techniques
- Conclusion

2005-03-03T17:00-05:00

CS System Seminar Talk

4

## Importance of Provenance

- Scientific Domain
  - Publications are Provenance!
  - Many scientific datasets available online
    - Biology, Astronomy (SDSS)
  - Standard metadata describes datasets in well-known repositories
  - Lineage information usually missing, but vital
  - GIS: Fitness for use
  - Material Engineering: Pedigree, Auditing
  - Biology: Citation & copyright, trust
  - Astronomy: Context information

2005-03-03T17:00-05:00

CS System Seminar Talk

5

## Importance of Provenance

- Business Domain
  - Data warehousing: Integrated view over historical data from multiple sources
  - Complex transformations to generate normalized view
  - Business analytics and intelligence (OLAP queries)
  - Lineage allows “drill-down” from view to source table
  - Allows tracing back sources of errors
  - “View deletion” problem

2005-03-03T17:00-05:00

CS System Seminar Talk

6

## Data Processing Architectures

- Service Oriented Architecture
  - Grid & Web services
  - Workflow & Service invocations
  - Data as parameters, references
- Databases
  - Update/View Queries, Stored Procedure Calls
  - Views, Tables, tuples, attributes
- Scripting, Command-line, etc.

2005-03-03T17:00-05:00

CS System Seminar Talk

7

## Contents of Talk

- Background
- • Taxonomy for Provenance Techniques
- Survey of Provenance Techniques
- Conclusion

2005-03-03T17:00-05:00

CS System Seminar Talk

8

## Taxonomy of Provenance Techniques

- Framework for analyzing & comparing provenance methodologies
- Key questions answered by taxonomy:
  - What is provenance used for?
  - What does provenance describe?
  - How is provenance represented?
  - How is provenance stored?
  - How is provenance disseminated?

## Application of Provenance

- Data Quality
  - Evaluate quality of data
  - Errors tend to inflate as they propagate
  - Trust in the source of data
  - Use provenance and metadata information to estimate data quality *for a user*
  - Assertions and Signatures for provenance guarantee
- Audit Trail
  - Error detection
  - Usage log

## Application of Provenance

- Replication Recipe
  - Provenance can be recipe for generating a dataset
  - Repeat to verify/compare
  - Recreate/replicate
  - Partial updates
- Attribution
  - Copyright, citation, check data users
- Informational
  - Discover datasets
  - Browse provenance

2005-03-03T17:00-05:00

CS System Seminar Talk

11

## Subject of Provenance

- What is provenance about?
- Data vs. Process Provenance
  - Provenance can be a graph of data & processes
  - Explicit provenance associated with data
  - Implicit provenance associated with process
  - Hybrid where all grouped together
- Granularity
  - Attribute (single pixel), tables, files, data collections (WRF experiment run)
  - Fine-grained vs. Coarse-grained
  - Trade-off with cost of collecting, storing, querying

2005-03-03T17:00-05:00

CS System Seminar Talk

12

## Representation of Provenance

- Scheme for representing provenance
  - Annotations vs. Inversion
  - Annotate data with ancestral data & the steps used to derive it e.g. a DAG
  - Store function (query) used to generate data and invert it
  - Annotation requires more storage; “Eager”
  - Not all functions are invertible; auxiliary data required; JIT computation; query optimization

2005-03-03T17:00-05:00

CS System Seminar Talk

13

## Representation of Provenance

- Contents of Provenance
  - Inversion requires stages of queries and auxiliary data
  - Minimal information provided (“Where”, “Why”)
  - Annotation can be as rich as user decides
  - Graph of steps and data sources; parameters; notes
  - Generic metadata vs. Provenance metadata
- Syntactic vs. Semantic Information
  - XML for Annotations; Implement specific for Inversion
  - Semantic information embedded in lineage metadata
  - Context; accurate searches; lineage proofs

2005-03-03T17:00-05:00

CS System Seminar Talk

14

## Provenance Storage

- With or separate from data
  - Integrity, accessibility
- Maintenance
  - Mutability, versioning
- Scalability
  - # of datasets, depth of lineage, granularity, geographical distribution, # of users
  - Inversion vs. Annotation; Distributed vs. Centralized
- Overhead
  - Collection & storage
  - Automation

2005-03-03T17:00-05:00

CS System Seminar Talk

15

## Provenance Dissemination

- Browsing Provenance as a DAG
  - Go back and forward in lineage through GUI
- Query based on lineage
  - By source data, or generating process
  - Enhanced by semantic information
  - Drill down during data mining
- Verify how data was created by reenactment or present proof statements

2005-03-03T17:00-05:00

CS System Seminar Talk

16

## Taxonomy in Brief

- **Application of Provenance**
  - Data quality                      Audit trail                      Attribution
  - Replication Recipe              Informational
- **Subject of Provenance**
  - Data vs. Process                  Granularity
- **Representation of Provenance**
  - Annotation vs. Inversion      Contents
  - Syntactic vs. Semantic
- **Provenance Storage**
  - Scalability                          Overhead
- **Provenance Dissemination**

2005-03-03T17:00-05:00

CS System Seminar Talk

## Contents of Talk

- Background
- Taxonomy for Provenance Techniques
- **Survey of Provenance Techniques**
- Conclusion

2005-03-03T17:00-05:00

CS System Seminar Talk

18

## Survey of Provenance Techniques

	Lanter, D. P.	Chimera/VDG	MyGRID	CMCS	PASOA	ESSW	Tioga	Trio
Domain	GIS	Generic	Biology	Chemical Science	Generic	Earth Science	Generic	Generic
Data Processing Framework	Command Processing	Service Oriented	Service Oriented	Service Oriented	Service Oriented	Script Based	Query Based	Query Based
Application of Provenance	Update, Regenerate	Regenerate, Planning	Auditing, Context	Information	Information, Reenact	Information	Track Errors	Information
Data/Process Oriented	Data	Both	Process	Data	Process	Both	Data	Data
Granularity	Spatial Layer	Abstract Datasets	Flexible	Files	WF Params	Files	Attribute in Database	Tuples in Database
Provenance Format	Annotation (Commands)	VDL Annotation	XML/RDF Annotation	DC XML Annotation	XML Annotation	XML/RDF Annotation	Inverse Functions	Relational Table
Semantic Information	No	No	Yes	Partial	No	No (Proposed)	No	No
Storage	Meta-Database	RDBMS	mIR Repository	WebDAV	RDBMS	RDBMS	RDBMS	RDBMS
Automated Recording	Store user commands	Workflow Trace	Workflow Trace	No	No	Libraries to generate	User's Inverse Fns	Proposed
Scalability	No	Yes	No	No	No (Proposed)	No (Proposed)	No	No
Dissemination	Queries	Queries	Semantic Browser	Browser; RDF	Browser; Queries	Browser	Queries	TriQL Queries

2005-03-03T17:00-05:00

CS System Seminar Talk

19

## Contents of Talk

- Background
- Taxonomy for Provenance Techniques
- Survey of Provenance Techniques
- Conclusion

2005-03-03T17:00-05:00

CS System Seminar Talk

20

## Discussion

- Convergence of Web service & Data warehouse provenance
- Trust, security
- Quality metrics
- Standardization

2005-03-03T17:00-05:00

CS System Seminar Talk

21

## Conclusion

- We motivate the necessity for recording and managing provenance information in the scientific and business communities
- We define a taxonomy of provenance techniques that have been adopted for managing provenance
- We use the taxonomy to compare and contrast nine key provenance systems and concepts that have been proposed or are under active research

2005-03-03T17:00-05:00

CS System Seminar Talk

22

## Current & Future Work

- Exploring provenance management as part of LEAD Project
- Research issues in using provenance for data quality estimation

2005-03-03T17:00-05:00

CS System Seminar Talk

23

*Thank you!*

Questions | Comments