Two Themes in Digital Curation: Archiving and Annotation

D C C

The University of Edinburgh – School of Informatics – Digital Curation Center

THE UNIVERSITY OF EDINBURGH

Archiving Scientific Data with XMLArch

Why Archive Data?

- Backup!
- Data history required for …
  - Verification of findings.
  - Citation.
  - History tracking.

How to Archive Data?

- Complete periodic snapshot.
  - High storage overhead.
- Diff-based approaches
  - Snapshot + incremental diff.
  - Snapshot + reverse diff.
- Snapshot + transaction log capture
- Some combination of the above.

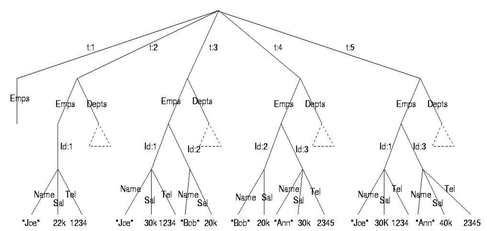
- Problems of diff-based approaches
  - Work on lines of text not data objects.
  - Know nothing about the domain.
  - Cannot track object history.
  - Sensitive to formatting/layout.
  - Retrieval is bounded by the number of diffs not data size.
  - History tracking is complex.

How do WE build Archival Databases!

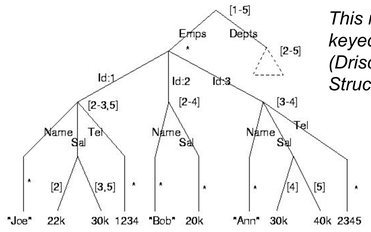
(Buneman, Khanna, Tajima, Tan, 2004)

- Hierarchical structure (XML).
  - Use unique identifiers.
- Merge all versions into a single archive.
- Benefits include …
  - Retrieval overhead is reduced.
  - History tracking is possible.
  - Reduction of storage overhead.
  - Stored in human readable format.

A Sequence of Versions



Pushing time down

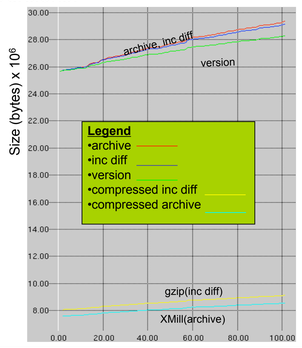


This relies on a deterministic / keyed model. (Driscoll et al. "Making Data Structures Persistent")

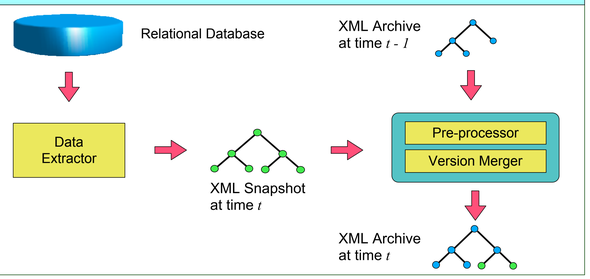
100 days of OMIM

- Recorded all OMIM versions for about 14 weeks.
- XML-ized all of them.
- Combined into XML format archive by pushing time down.
- Also recorded diffs between versions.

- Uncompressed
  - Archive size is
    - ≤ **1.01** times diff repository size.
    - ≤ **1.04** times size of largest version.
- Compressed
  - Archive size between **0.94** and **1** times compressed diff repository size.
  - gzip - unix compression tool
  - XMill - XML compression tool



Architecture

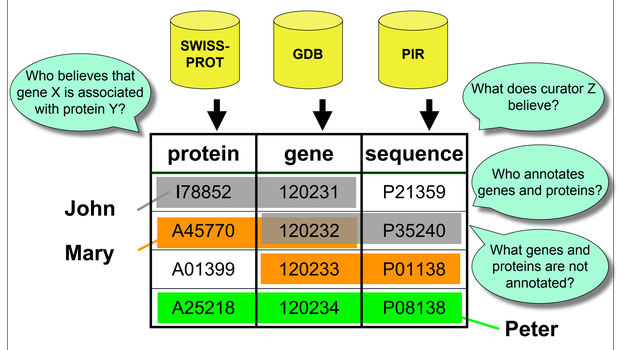


Mondrian: Annotating and Querying Databases through Colors and Blocks

Annotation in Databases

- Annotations play a central role in database curation.
- DBMS lack support for …
  - Modeling different granularity.
  - Storing and querying annotations.
- Our contributions
  - Annotation oriented data model using colored blocks to represent annotated sets of values.
  - Color query language that is minimal, sound, and complete.

Motivating Example



How to color databases?

- We need …
  - Set of colors C.
  - A coloring function x that accepts
    - a tuple t in a instance r of relation R.
    - a set of attributes $Y \subseteq sort(R)$.

    and assigns a set of colors $C' \subseteq C$ to $t[Y]$.
- A colored database is a set of blocks $(t, Y, x(t, Y))$, e.g., $x(t_1, \{protein, gene\}) = \{grey\}$.

How to query the database?

- We introduce a color algebra (*CA*)
  - selection, projection, product, renaming, union …
  - block selection (Σ), block projection (Π^L, Π^U), merge (μ), recolor (ρ).
- Prove the algebra to be **minimal**, **sound**, and **complete**.
- Benefits include …
  - appropriate level of abstraction.
  - respect semantics of colors & blocks.
  - easy to use and portable.

Block projection & Selection

Π^L gene, sequence (r):

Find tuples with annotations involving **at least** genes & sequences.

protein	gene	sequence
A45770	120232	P35240
A25218	120234	P08138

Π^U gene, sequence (r):

Find tuples with annotations involving **at most** genes & sequences.

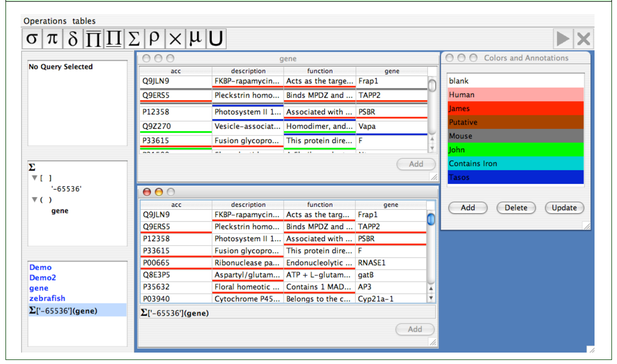
protein	gene	sequence
A45770	120232	P35240

Σ grey (r):

Find tuples with grey annotations.

protein	gene	sequence
I78852	120231	P21359
A45770	120232	P35240

Mondrian



- Other related topics
  - Data provenance.
  - Data publishing.
  - Citation for databases.
- For further information
  - www.lfcs.inf.ed.ac.uk/research/database.
  - www.dcc.ac.uk.

D C C