dbWiki – A Wiki for Structured Data

Peter Buneman, James Cheney, Heiko Müller, University of Edinburgh
Sam Lindley, Contemplate Ltd.

Background

- Sharing and publishing data is one of the most important activities of modern science.
- Well-known examples include Wikipedia, a vast store of mostly trusted knowledge that is collaboratively built and shared on the internet.
- Relational databases are widely used for building curated databases, usually capturing a great deal of human effort.

Objectives

- Why not use a wiki for curated data? Wikis are designed for text, and we need something for structured data.
- Idea: Database Wiki (dbWiki), a wiki designed for structured data.
- See also: "Structured wikis", "semantic wikis", Linked Data on the Web.
- Our goal: Combine scalability and robustness of databases with usability and convenience of Wikis.

Features

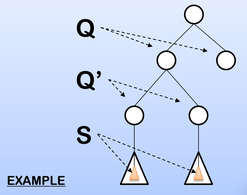
- Versioning/archiving: All previous versions of the data are stored.
- Citation: Individual data fields, records and tables can be cited and linked from other databases, not just whole databases and URLs.
- Annotation: ability to "tag" or comment upon any part of the data.
- Provenance: System provides detailed information about the history - not just snapshots of previous versions but also pointers to external sources, explanations of query results.
- Query integration.

Key challenges

- How can we make it easy for non-experts to curate, present and program with data on the Web?
- How can we support archiving, citation, annotation and provenance efficiently in databases?
- How should these advanced features be integrated with Wiki-like markup languages and database query languages?
- How can we handle evolution of data from unstructured to structured forms?

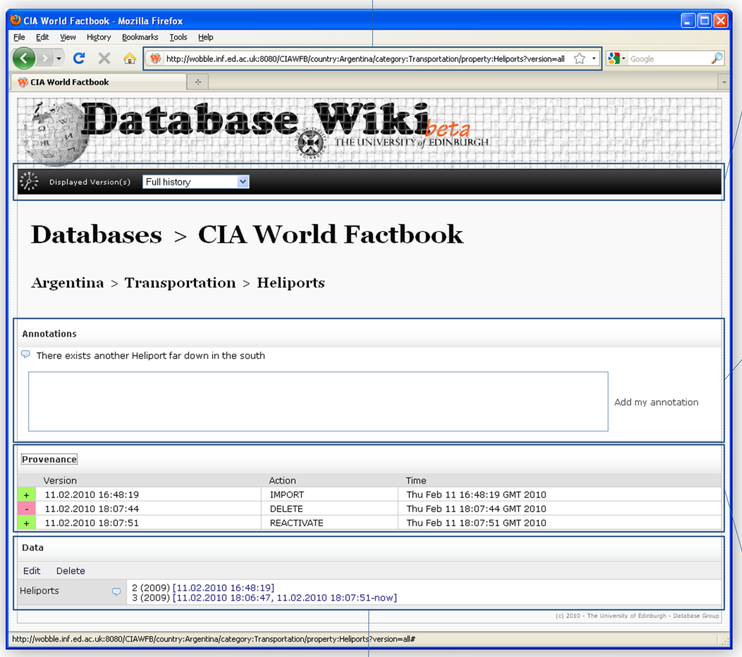
Citation

- Elements **Q'** are keyed relative to their parent **Q** by part of their subtree **S**.
- Every element has a unique absolute key in a hierarchical structured database.
- The combination of version identifier and element key forms a reliable hyperlink mechanism in which every element is identified (and thereby citable) by a unique URL.



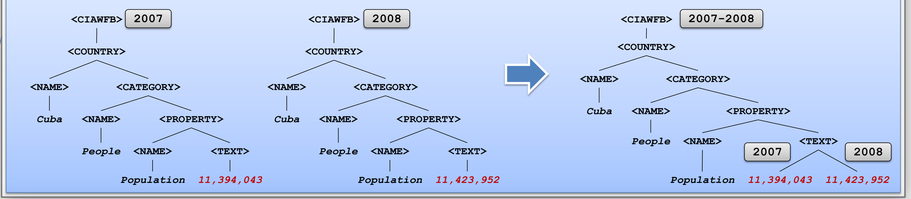
EXAMPLE

KEY /COUNTRY BY VALUES (NAME)
KEY /COUNTRY/NAME BY EXISTENCE



Versioning

- Merge database versions by "pushing time down" [Buneman et al. 2004, Müller et al. 2008].



Annotation

- Arbitrary parts of the data can be annotated.
- Enable discussions about the quality of data or alternative data values.
- Propagate annotations made on a user's view of the database back to the actual core data.
- See also dbNotes [Bhagwat et al. 2005], and MONDRIAN [Geerts et al. 2006]

R

A	B	C
1	2	2
1	2	3
2	3	4

S

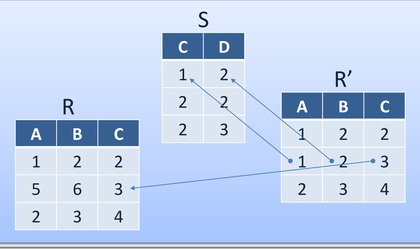
C	D
1	2
2	2
2	3

R ⋈ S

A	B	C	D
1	2	2	2
1	2	2	3

Provenance

- The majority of curated data is copied and edited from existing sources.
- Knowing the origin of data – its **provenance** – is particularly important.
- dbWiki will support rich forms of provenance tracking [Buneman et al. 2006, Cheney et al. 2007].



Modification and Views

- dbWiki will support creating custom views and forms.
- Plan to use ideas from Links Web Programming Language [Cooper et al. 2006].
- Easier than SQL / PHP / JavaScript.

```
<table>
  for (row <-- table)
  where (row.a = row.b)
    <tr>
      <td>row.a</td>
      <td>row.b</td>
      <td>row.c</td>
    </tr>
</table>
```

References

D. Bhagwat, L. Chiticariu, W-C. Tan, G. Vijayvargiya. An annotation management system for relational databases. VLDB Journal 2005.

P. Buneman, A. Chapman and J. Cheney. Provenance management in curated databases. SIGMOD 2006.

P. Buneman, S. Khanna, K. Tajima, and W.-C. Tan. Archiving scientific data. TODS 2004.

J. Cheney, U. A. Acar and A. Ahmed. Provenance as dependency analysis. DBPL 2007.

E. Cooper, S. Lindley, P. Wadler, and J. Yallop. Links: Web Programming Without Tiers. FMCO 2006.

F. Geerts, A. Kementsietsidis and D. Milano. MONDRIAN: Annotation and Querying Databases through Colors and Blocks. ICDE 2006.

H. Müller, P. Buneman, I. Koltsidas. XArch: archiving scientific and reference data. SIGMOD 2008.