A curation interface for temporal databases

Vashti Galpin[1], Simon Fowler[2], James Cheney[1]

University of Edinburgh[1]
University of Glasgow[2]

International Data Curation Conference 2021 (IDCC21)
19 April 2021

THE UNIVERSITY *of* EDINBURGH
informatics

University *of* Glasgow

Introduction: Curated scientific databases

Curated databases:

- → "databases that are populated and updated with a great deal of human effort" (Buneman et al., 2008)
- → Examples: IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb), CIA World Factbook, Manually Curated Database of Rice Proteins (MCDRP)

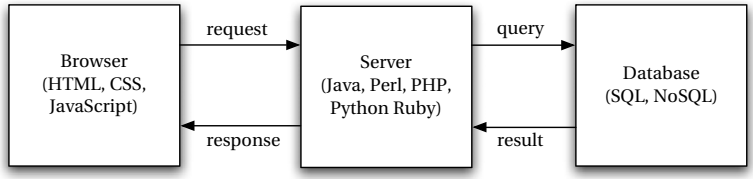
Curation challenges:

- → Provenance tracking, citation, archiving, annotation propogation ...
- → Digital tools have to be hand-crafted for every curated DB — often with very stretched resources

Ultimate goal: programming language support for curation functionality

- → Cross-tier web-programming using the **Links** language
- → Previous work: GtoPdb re-implementation (Fowler et al, 2020)
- → This research: **prototype** curation interface for Covid-19 data

Background: Cross-tier web programming



Curated databases normally a **collection of applications**:

→ Database, web frontend, curation application

Links: cross-tier programming language with language-integrated query:

→ Client, server, and database code written in same language

→ Database queries written in the same language providing well-formed and efficient queries

→ Language developed at UoE since 2006, with temporal database features recently added

Provenance and curation

Definition of provenance:

- → "Essentially, provenance can be seen as meta-data that, instead of describing data, describes a production process." (Herschel et al, 2017)
- → Here our focus is on data provenance in the context of **data updates**

Provenance facilitates curation:

- → Captures the process of data creation and allows analysis of data change
- → Records the human effort and decisions involved in curation
- → Allows for assessment of data quality and integrity

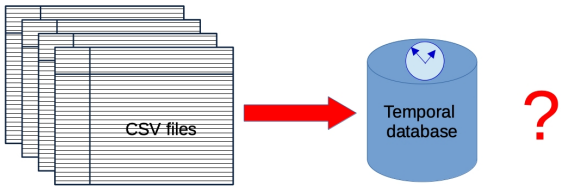
Provenance supports data sharing:

- → Availability of provenance information can build trust for data sharing and reuse
- → Provides information for data versioning via time-slicing

Temporal databases and provenance queries

How many deaths were reported in the week of 5 July?

How were those deaths distributed across the health boards?



How many deaths were first reported in week of July 5?

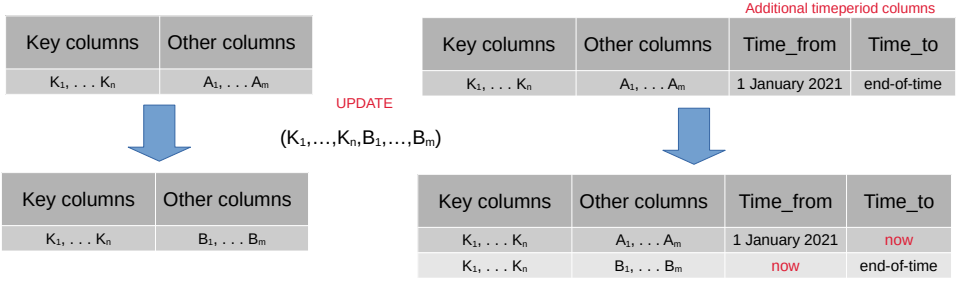
How does that differ from the most recently reported figure?

Which health boards reported the most changes?

Temporal databases and update provenance

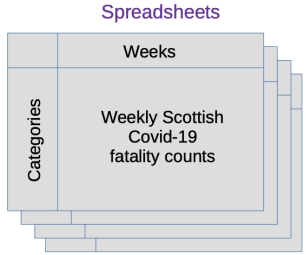
Temporal databases:

- → Introduction of time period information to individual rows by adding two columns: the start of the time period and the end of the time period
- → The time period describes when the data is valid and temporal queries use this time information
- → Interpretation of time period is application dependent: transaction time or another metric of validity such as validity in the real world.



Case study: Scottish Covid-19 figures

- → Weekly CSV files with weekly Covid-19 fatality counts for sex, age, health board, local authority, and location
- → Each CSV file contains new data for (at least) one week and may contain updates to counts from previous weeks



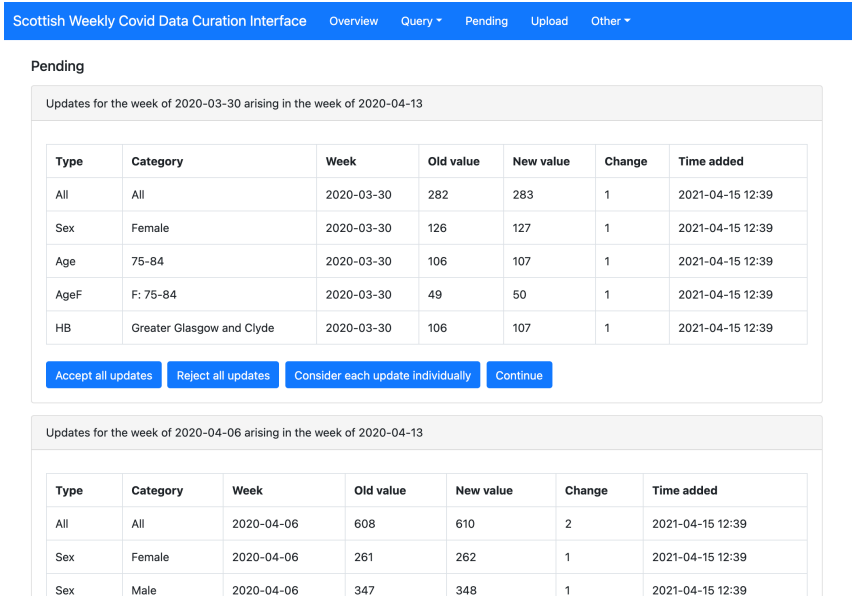
National Records of Scotland

www.nrscotland.gov.uk/statistics-and-data

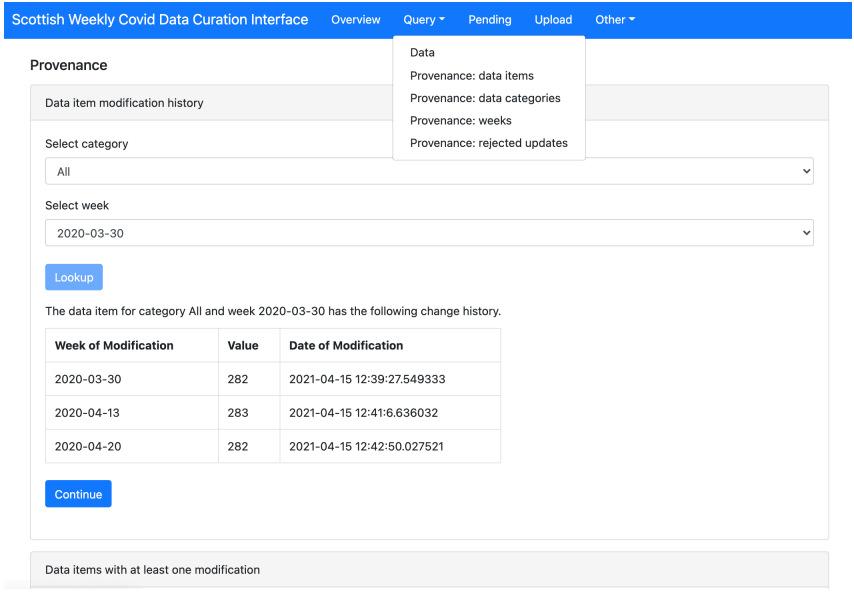
Temporal database table

Week date	Category	Fatality count	File_id	Date_valid_from	Date_valid_to

Curation interface prototype: update decisions



Curation interface prototype: provenance queries



Conclusion: Links for curation interfaces

- **Curated databases**:
  - Databases maintained using much human effort
  - Consist of multiple applications: database, web frontend, ...
  - How can we automate support for curation?

- **Curation interface**:
  - Links is a cross-tier programming language with client, server, database code in a single language
  - Temporal database features support update provenance
  - A bit like Github but for data

See our poster **Links between Temporal Databases and Curation** and complete our curation interface survey at **RDA Plenary 17**

Vashti.Galpin@ed.ac.uk

http://www.links-lang.org