Temporal data management for data stewardship: advancing provenance in reuseability

THE UNIVERSITY of EDINBURGH informatics

Vashti.Galpin@ed.ac.uk
James.Cheney@ed.ac.uk

Vashti Galpin
orcid: 0000-0001-8914-1122
James Cheney
orcid: 0000-0002-1307-9286

European Research Council
Established by the European Commission

ERC Consolidator Grant Skye: 682315

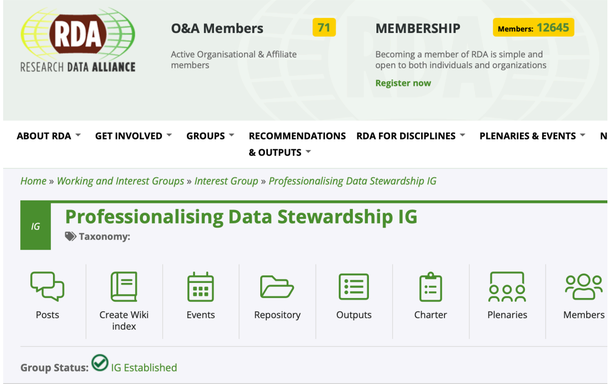
FAIR Principle R2.1

"(Meta)data are associated with detailed provenance."

Provenance provides the "where", "what", "when", "who" and "how" of data, giving information about the origin, creation and modification of a piece of data (versioning, by contrast, allows for the identification of the specific data that was used).

Provenance information can increase trust which in turn leads to data becoming more reusable.

We focus on automatically tracking the "when" of data using temporal DBs. This is an example of support for provenance that can be added to the toolbox of data stewardship.



Data stewards are "support staff from research communities and research libraries, and those managing data repositories" [1]. They provide professional support for FAIR data [2], including reusability.

Temporal Databases

Temporal DB techniques are well established in the database community but need to be better known by the broader scientific and research data community, especially in the case of stewardship and curation that occurs over long periods of time with data that varies.

Support for temporal data management can be provided by extending tables with information about the time when data changes. Updates preserve data rather than overwriting it.

Additional time columns

Key columns	Other columns	Time_from	Time_to
$K_1, \ldots K_n$	$A_1, \ldots A_m$	1 January 2021	end-of-time

UPDATE $(K_1,\ldots,K_n,B_1,\ldots,B_m)$

Key columns	Other columns	Time_from	Time_to
$K_1, \ldots K_n$	$A_1, \ldots A_m$	1 January 2021	now
$K_1, \ldots K_n$	$B_1, \ldots B_m$	now	end-of-time

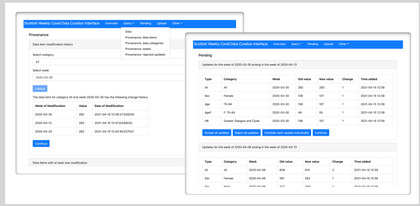
Temporal techniques can be expressed using standard SQL [3] but this requires substantial database programming skill.

Links provides a simpler approach [4]. It is a strongly typed functional language with temporal database support and language-integrated query.

Links supports the development of correct code for both web-based front-ends and database programming in a single language. This approach automatically generates correct SQL thereby avoiding the mistakes that can occur when explicit SQL statements are used for database interaction.

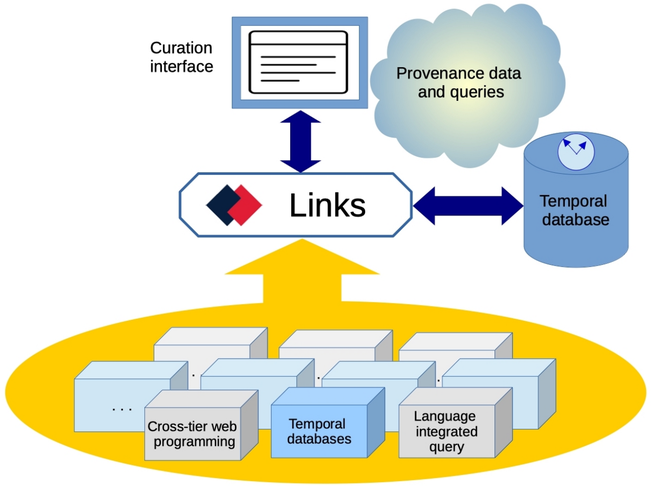
Update Provenance in Links

Case study 1: Scottish Covid-19 data released weekly with corrections, where the need is to track these corrections [5].



Case Study 2: Digital calibration certificates in metrology where measurement data will change over time due to repeated calibration [6].

Case study 3: Retrofitting update provenance support to an existing curated scientific database (GtoPdb) with a substantial user base and which has a couple of releases a year.



Retrofitting Provenance

The IUPHAR/BPS Guide to PHARMACOLOGY (GtoPdb) describes interactions between proteins and ligands from the medical research literature. It is curated by experts. Previously, GtoPdb has been implemented in Links to demonstrate the use of cross-tier programming in this context [7].

The curators of GtoPDB are also interested in understanding and quantifying the changes that occur between releases.

We provided SQL triggers for their development DB. Our triggers made no changes to the existing tables and only added information to new tables.

After several months, they provided us with a full database dump from which we instantiated a temporal database. Using Links, we developed a prototype which allows curators to see the number of changes for a table and the specifics of any change.

Further research: Allowing users of the DB to see changes between releases, and much more ambitiously, to support curation by users.

References

1. D3.4 Recommendations on practice to support FAIR data principle (Draft), FAIRsFAIR, https://doi.org/10.5281/zenodo.3924132)
2. https://www.rd-alliance.org/groups/professionalising-data-stewardship-ig
3. R. Snodgrass, Developing Time-Oriented Database Applications in SQL. Morgan Kaufmann, 1999
4. The Links Programming Language, https://links-lang.org
5. V. Galpin and J. Cheney, Curating Covid-19 Data in Links, IPAW 2020/21, LNCS 12839, 2021, https://doi.org/10.1007/978-3-030-80960-7_19
6. V. Galpin, I. Smith, J.-L. Hippolyte, Supporting Provenance of Digital Calibration Certificates With Temporal Databases, IMEKO T6 International Conference on Metrology and Digital Transformation, 2022, to appear
7. S. Fowler. S. Harding, J. Sharman, J. Cheney, Cross-tier Web Programming for Curated Databases: a Case Study, International Journal of Digital Curation 16(1), 2021, https://doi.org/10.2218/ijdc.v16i1.735