Data round-tripping with authority files

An overview and some examples

Camillo Carlo Pellizzari di San Girolamo (user:Epìdosis)
Scuola Normale Superiore

This session is recorded: Please mute your microphone and camera when you're not speaking.

DATA REUSE DAYS February 18-27_2025

Summary

1. Data reuse and data round-tripping
2. Types of mistakes and what to report
3. Twelve examples of data round-tripping with national authority files
4. Conclusions

1) Data reuse and data round-tripping

DATA REUSE DAYS February 18-27_2025

Data round-tripping

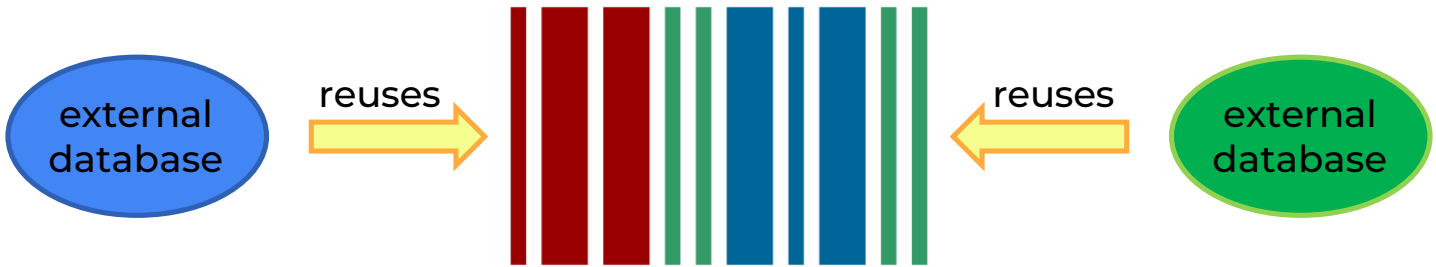
Cf. https://www.wikidata.org/wiki/Wikidata:Data_round-tripping

(created after Data Quality Days 2022)

Data round-tripping consists in the reciprocal synchronisation of data between Wikidata and the external databases to which it links, with main focus on the **correction of errors found by Wikidata users in external databases**.

Why are we talking of data round-tripping here?

Yes, it is more of an issue of data quality, but ...



Why are we talking of data round-tripping here?

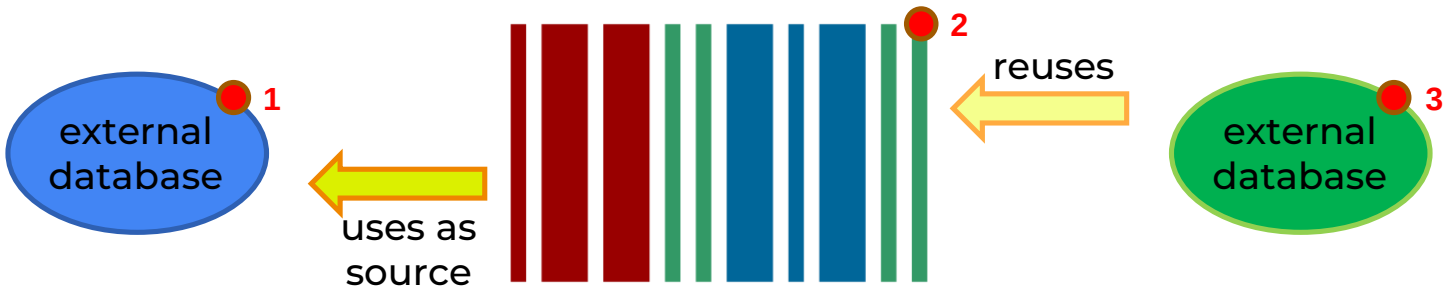
Yes, it is more of an issue of data quality, but ...

some databases only reuse Wikidata, some databases are only used as source by Wikidata, for some both things are true.

Why are we talking of data round-tripping here?

Yes, it is more of an issue of data quality, but ...

● = mistake

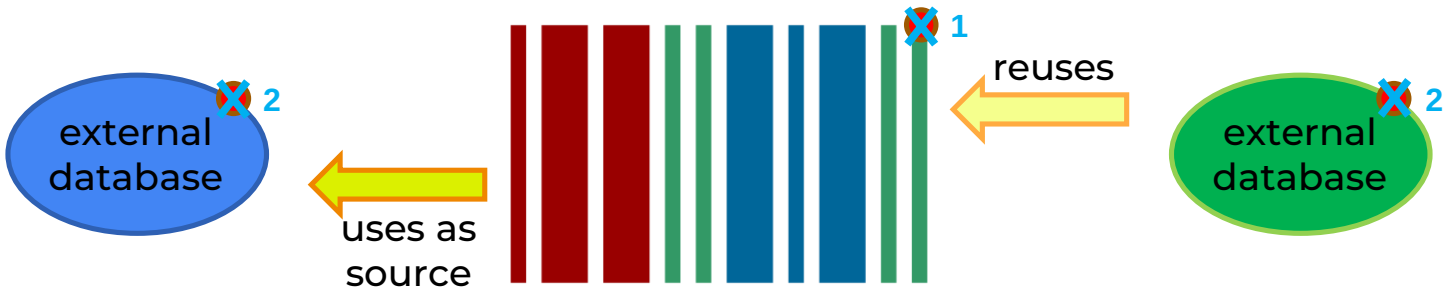


some databases only reuse Wikidata, some databases are only used as source by Wikidata, for some both things are true.

Why are we talking of data round-tripping here?

Yes, it is more of an issue of data quality, but ...

● = mistake



some databases only reuse Wikidata, some databases are only used as source by Wikidata, for some both things are true.

2) Types of mistakes and what to report

DATA REUSE DAYS February 18-27_2025

Types of mistakes

- **Duplications** (same entity, 2 or more IDs) – by far the most frequent
- **Conflations** (same ID, 2 or more entities)
  - **in the entry** itself
  - **in the bibliographic records** connected to the entry
- **Wrong data** (mistakes in the name, in the data etc.)
- [**Obsolete data** (e.g. missing death dates)]

What to report

- **Duplications**: since they are many, and can be extracted through single-value constraint violations, usually they are ***not*** reported
- **Conflations**: very dangerous, should ***always*** be reported
- **Wrong data**: very dangerous, should ***always*** be reported
- [**Obsolete data**: very frequent, and not mistakes *stricto sensu*, so usually they are ***not*** reported]

Methods of mistake reporting

external (in the database)

- **e-mail**: one e-mail address dedicated to mistake reports
- **web form**: one webform for all reports, or one webform for each entry

internal (in Wikidata)

- **page on Wikidata** (or Wikipedia): one page for all reports
- **user(s)**: one or more users to which you can send reports

Cf. https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control/Error_reporting_procedures and https://www.wikidata.org/wiki/Property:P10923

Degrees of cooperation with databases

1. **passive correction**: receiving mistake reports
2. **active correction**: actively using Wikidata to find mistakes
3. **editing Wikidata**

This presentation is mainly concerned with case 1)

Case 2) is still very rare, especially for national authority file.

3) Twelve examples of data round-tripping with national authority files

DATA REUSE DAYS February 18-27_2025

Premise

- in the following slides I report:
  - the name of the authority file, as abbreviated in VIAF
  - the ID of the property
  - the number of values presently used in Wikidata on humans (P31=Q5), rounded to the thousands
  - some comments about existing procedures of data round-tripping
- the slides are ordered by decreasing number of values in Wikidata

VIAF (P214) – 3.029 M values

- reporting method: e-mail bibchange@oclc.org
  - *in my experience*, never effective, maybe because the clusterization is difficult to correct manually if the algorithm used makes the wrong choices
  - also, wrong clusterization depends often on wrong data sent by VIAF members, so mistakes should be corrected first in the members to improve clusterization, and VIAF *seems* to have no workflow to forward mistake reports to members
- good news: adding duplicate VIAF clusters in the same Wikidata item, or marking a VIAF cluster as conflated in Wikidata, *seems* to have *some* impact on VIAF clusterization after a few weeks months

ISNI (P213) – 1.747 M values

- reporting method: a webform “Please help us improve this record” for each entry
  - very effective: *in my experience*, the great majority of reports is solved in maximum two/three workdays; reports rarely get lost and usually, if you send them again, they are solved as quickly as the others
  - filling in your e-mail is not compulsory; in easy cases it is unnecessary; it is useful when reporting complex cases (e.g. big conflations), sometimes they write to ask clarifications on how to solve the case

DNB [GND] (P227) – 1.718 M values

- in Wikipedia: since 2005, the Wikipedia page https://de.wikipedia.org/wiki/Wikipedia:GND/Fehlermeldung
  - effective, but with a delay of more than one year – given the number of reports, more resources should be allocated to take care of these reports more quickly
- in Wikidata: user Kolja21 has editing access to some parts of the authority records and is available for receiving mistake reports

LC (P244) – 1.354 M values

- reporting method: one webform https://www.loc.gov/contact/catalog-record-error-report/
  - not effective *in my experience*, although I have sent very few reports; at least, I have never been contacted through the e-mail (that is compulsory in the webform)
- since in Wikidata there is an ongoing collaboration with LC (see user LccnBot), probably the situation can be improved

SUDOC [IDREF] (P269) – 0.932 M values

- reporting method: a webform “Améliorer la notice” for each entry
  - effective: *in my experience*, the vast majority of reports are solved in a few workdays; however, reports rarely get lost and it may be difficult, also sending them again, to obtain their solution
- in Wikidata: user Thomas Kerboul (BGE) has wide editing access and is available for receiving mistake reports

NKC (P691) – 0.692 M values

- reporting method: one webform “Feedback”
    - effective: *in my experience* reports are solved, although the necessary time could be sometimes longer than some workdays

NUKAT (P1207) – 0.648 M values

- no known reporting method
- in Wikidata: users NukatCatalog and NukatCatalogBot exist since 2018 but have been used very rarely and it is not fully clear who operated (and maybe still operates) them

J9U (P8189) – 0.620 M values

- in Wikidata: user Geagea is available for receiving mistake reports

BNF (P268) – 0.597 M values

- reporting method: e-mail corrections.autorites@bnf.fr (or a webform on each bibliographic record) – but only corrections based on RGPD are accepted, although in my experience they made about 10 exceptions before declaring this policy
- until January 2020 all types of reports were accepted, but the policy has been restricted because “Ce service a été victime de son succès et a dû être arrêté pour permettre à la BNF de travailler sur ces propres chantiers de correction” (e-mail, 2025-01-15)

ICCU [SBN] (P396) – 0.147 M values

- reporting method: ic-cu.AFnomi@cultura.gov.it
  - effective: *in my experience*, mistakes are usually solved in a few workdays; reports rarely get lost, but, if you send them again, they are solved as quickly as the others
- in Wikidata: user Epìdosis has wide editing access and is available for receiving mistake reports

SIMACOB [COBISS.SI] (P1280) – 0.146 M values

- reporting method: e-mail obdelava@nuk.uni-lj.si
  - very effective: *in my experience*, mistakes are usually solved in a few workdays

BAV [VcBA] (P8034) – 0.123 M values

- no known reporting method

4) Conclusions

DATA REUSE DAYS February 18-27_2025

Conclusions

- Most of the analyzed authority files have at least one more or less effective method of reporting mistakes, although improvements are always possible (e.g. giving the possibility to the user to register an account, in order not to have to fill in their e-mail each time!)
- Having at least one user available on Wikidata for receiving mistake reports usually makes the workflow more efficient
- Having authority files actively using Wikidata to find and correct their mistakes would improve significantly the situation

Thanks for your attention!

Get in touch with us:

Camillo Carlo Pellizzari di San Girolamo
camillo.pellizzaridisangirolamo@sns.it

Credits
Information about the licence of your slides, the pictures used in your slides, links to additional resources, etc.

DATA REUSE DAYS February 18-27_2025