



VIAF and ISNI

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

Summary

1. Introduction
2. Opportunities
3. Issues
4. How to manage their issues on Wikidata



1) Introduction

to VIAF and ISNI in Wikidata

VIAF and ISNI: what they are

- **VIAF** (<https://viaf.org/>) collects authority data from 40+ national libraries (or similar institutions) and clusters them through algorithms, with very little manual curation
- **ISNI** (<https://isni.oclc.org/cbs/>) is curated by multiple ISNI registration agencies (<https://isni.org/page/isni-registration-agencies/>); at least in initial stages it has been widely curated with algorithms, now manual curation seems more relevant

VIAF and ISNI in Wikidata (as of 14/09/2024)

- **VIAF** (in Wikidata [P214](#)):
 - [3,4 M](#) values in Wikidata
 - [2,9 M](#) values on humans in Wikidata

- **ISNI** (in Wikidata [P213](#)):
 - [1,7 M](#) values in Wikidata
 - [1,5 M](#) values on humans in Wikidata

A comparison between VIAF and Wikidata

- My first scientific article (2021):
Bianchini-Bargioni-Pellizzari di San Girolamo, *Beyond VIAF. Wikidata as a Complementary Tool for Authority Control in Libraries* (DOI: [10.6017/ITAL.V40I2.12959](https://doi.org/10.6017/ITAL.V40I2.12959))
- Data from 2020 here: https://catalogo.pusc.it/beyond_viaf/
 - at that time only 2 M humans in Wikidata had VIAF



2) Opportunities

of using VIAF and ISNI for Wikidata

VIAF and ISNI opportunities for Wikidata

- **both** are useful as for **matching** Wikidata with other databases (i.e. if DB X links to VIAF/ISNI and Wikidata links to VIAF/ISNI, then it's easy to link DB X from Wikidata)
- **VIAF** is particularly useful for easily **importing** in Wikidata the **IDs of its members** (mainly through **moreidentifiers**, or *very careful* massive bot imports)
- ISNI has no other practical advantages (being an ISO standard is a more of a theoretical advantage)

VIAF and ISNI: coverage comparison

- **VIAF** is much more complete than ISNI on authors of written publications (thanks to library authority files: not all VIAF members are ISNI registration agencies)
- **ISNI** is more complete than VIAF on artists, musicians, actors etc. (thanks to non-library registration agencies) and on young researchers (because a few universities create ISNI IDs for their PhDs when they discuss their thesis)



3) Issues

of using VIAF and ISNI for Wikidata

Main issues of VIAF (1)

- **identification (= clusterization) issues:** VIAF clusters, especially non-personal ones (organizations, places etc.), are wrong in a significant minority of cases; the main issues are:
 - **duplications:** many clusters for the same entity (especially frequent for ancient and medieval authors and authors whose native name is not in Latin scripts)
 - **conflations:** one cluster mixing 2(+) entities (usually homonyms or near-homonyms)

Main issues of VIAF (2)

- **no round-tripping**: although in theory VIAF has an email for mistake reporting (bibchange@oclc.org), in my experience it has never been effective in the last years; this is also due to the nature of VIAF itself, which clusters data sent to it by national libraries but has no direct control on them:
 - VIAF could completely solve wrong clusters only if they are entirely caused by its algorithms (i.e. if the data by libraries don't contain mistakes); but in most cases, to be honest, the issue (especially for conflations) originates **at the level of specific national libraries**

Main issues of VIAF (3)

- **opacity in the use of Wikidata:** whilst we have empirical clues that VIAF uses Wikidata in order to reduce its duplications and conflations, it is not documented anywhere how exactly this happens; so, in fact, after 2013 there has not been any structured collaboration between VIAF (i.e. OCLC) and Wikidata in order to improve their reciprocal quality; such a collaboration could have great advantages for both

Main issues of ISNI

- **identification issues:** ISNI IDs are wrong in a significant minority of cases; the main issues are **duplications** and **conflations** (since ISNI IDs originated from VIAF clusters, usually a problem in VIAF corresponds to a problem in ISNI; and ISNI sometimes retains issues already solved in VIAF).
 - However, unlike VIAF, ISNI has an **effective round-tripping** system: sending a mistake report through the online form usually assures that the issue is solved in a few workdays



4) How to manage their issues on Wikidata

primarily, a lot of patience is needed

VIAF and ISNI duplications

- *Not so worrying*: just **add all IDs** to the item they describe; this has only one minor side-effect, i.e. it makes impossible to use single-value constraint violations to find mistakes in Wikidata, since most of the violations are due to external mistakes (for more on this, see my 2023 presentation: [slides](#) and [video](#))
 - for **ISNI** you can **send a mistake report** through the online form and the issue will be quickly solved
 - for **VIAF** you can just **wait**; usually, in a few months, some of the duplicates are merged by VIAF (probably using Wikidata)

VIAF and ISNI conflations

- *Worrying*: **add the IDs** to all the items they describe, ranking them as **deprecated** with qualifier **reason for deprecated rank** (P2241) = **conflation** (Q14946528)
 - for **ISNI** you can **send a mistake report** through the online form and the issue will be quickly solved
 - for **VIAF**, if the conflation is (at least partially) due to mistakes of **single national libraries**, you should try to **report** them the mistakes; hopefully, if they fix them and send correct data to VIAF, the clusterization will then be improved accordingly

Extra: duplications and conflations in Wikidata

Of course duplications and conflations also affect Wikidata, not only external databases (VIAF, ISNI, authority files etc.).

I produced some documentation about this theme in 2023:

- my presentation at NLG *Managing conflations and duplications of personal items in Wikidata*: [slides](#) + [video](#)
- my [paper](#) *Conflations and duplications in Wikidata items: causes, detection, solutions, and issues*
- my presentation at Data Modeling Days *Conflations and duplications*: [slides](#) + [video](#)



Thanks for your attention!

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