



# Conflations and duplications

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.

# Summary

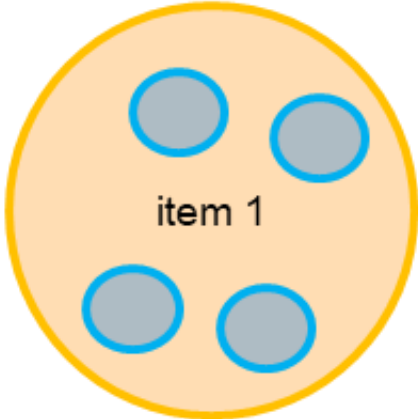
- Definitions
  1. Causes
  2. Detection
  3. Solutions
  4. Issues



# Definitions

# Conflation

conflation



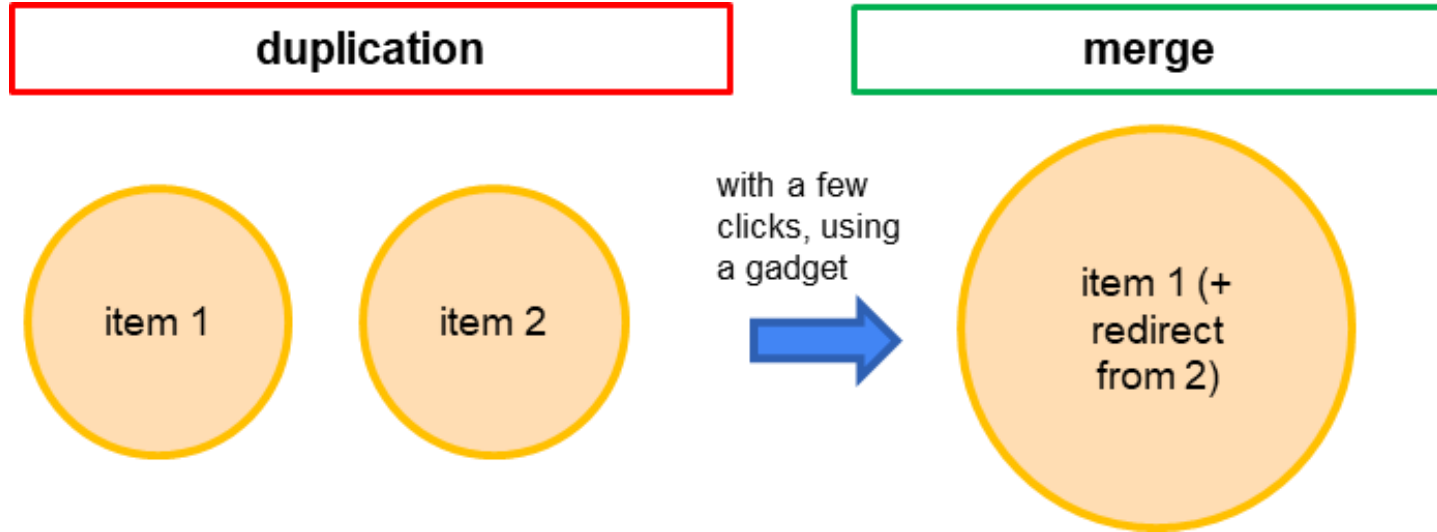
manually



split



# Duplication





# 1) Causes (and statistics)

# General causes

Conflations and duplications are issues which affect not only Wikidata, but also other databases (e.g. authority files managed by librarians)

- If many entities have the same name (or similar names), they risk to be conflated into one entry/item
  - e.g. John Smith, Hans Meyer, Mario Rossi
- If one entity has many names, it risks to be duplicated into many entries/items (one for each name)
  - e.g. Aristotle / Aristoteles / Aristotele / Ἀριστοτέλης

# How are conflations and duplications added to Wikidata?

- manual edits (by IPs and registered users)
- batches run through semi-automated tools (mainly QuickStatements and OpenRefine)
  - no approval process before running; discussion usually required to undo wrong batches, but no strict guideline presently exists (see [Wikidata:Edit groups](#))
- batches run through bot accounts
  - approval required before running (see [Wikidata:Bot requests](#))





# How many conflations and duplications are added to Wikidata?

- it is impossible to know exactly how many conflations and duplications are still waiting to be solved
- it is very difficult to know how many conflations have been already solved because splits cannot be easily traced through tools
- it is possible to know (approximately) how many duplications have been already solved counting the number of merges and the number of redirected items

# So, how many merges and how many redirects? (1)

Out of 107.6 M items:

- 3.6 M merges (see [NavelGazer](#)); about 30k merges each month
- 4.1 M redirects (see [Wikiscan](#))
  - 3.0 M redirects created by bot accounts
  - 1.1 M redirects created by non-bot accounts

# So, how many merges and how many redirects? (2)

Considering redirected items by year of creation:

- **bots** have created more redirected items\* than non-bots each year from 2012 to 2019 (in total 2.7 M vs 0.6 M)
- **non-bots** have created more redirected items\* than bots each year from 2020 to 2023 (in total 0.5 M vs 0.3 M)

\* obviously these items weren't redirects at the moment of their creation, but they were merged and became redirects afterwards



## 2) Detection

# Use of constraint violations

The two most important constraint violations used to discover conflations and duplications are:

- **the same ID value in two (or more) items: **unique-value**** constraint violation
- **two (or more) ID values in the same item: **single-value**** constraint violation







Constraint violations are *usually* caused by a conflation or duplication; however, the issue may be not in Wikidata, but in the ID.



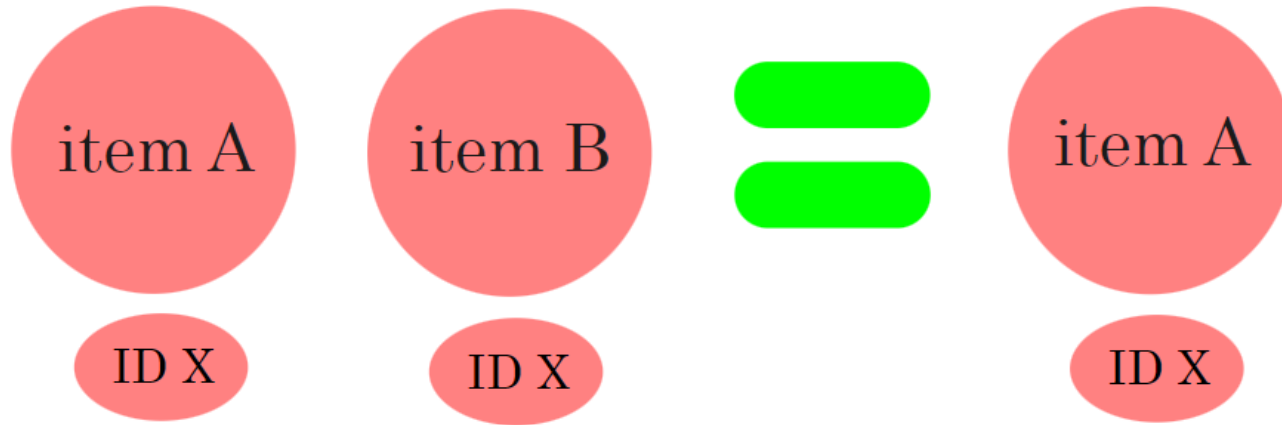
# Use of constraint violations

	<b>Wikidata</b>	<b>external DB</b>
<i>frequent</i> <b>duplication</b>	duplicate items same ID in two items	duplicate IDs two IDs in the same item
<i>usually rare</i> <b>conflation</b>	conflated items two IDs in the same item same ID in two items	conflated IDs two IDs in the same item same ID in two items

# Use of constraint violations: what you can solve

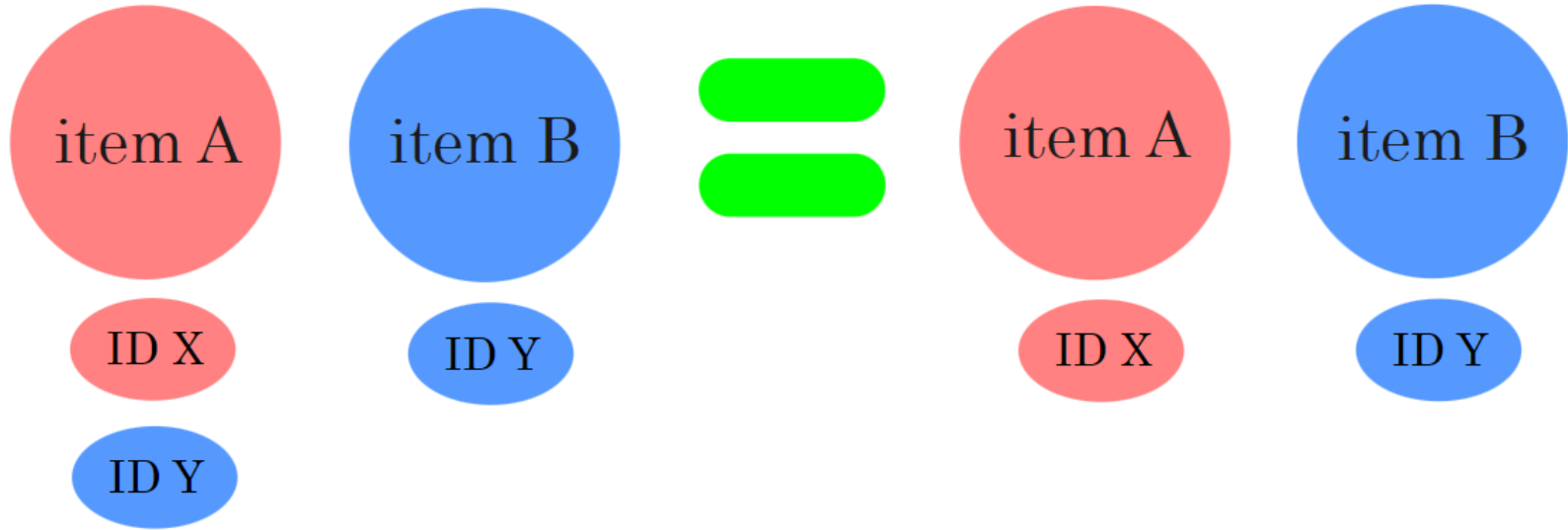
	<b>Wikidata</b> 	<b>external DB</b> 
<i>frequent</i> <b>duplication</b>	duplicate items  1) same ID in two items	duplicate IDs  4) two IDs in same item
<i>usually rare</i> <b>conflation</b>	conflated items  5) two IDs in same item 2) same ID in two items	conflated IDs  6) two IDs in same item 3) same ID in two items

# 1) same ID in two items: duplicate items

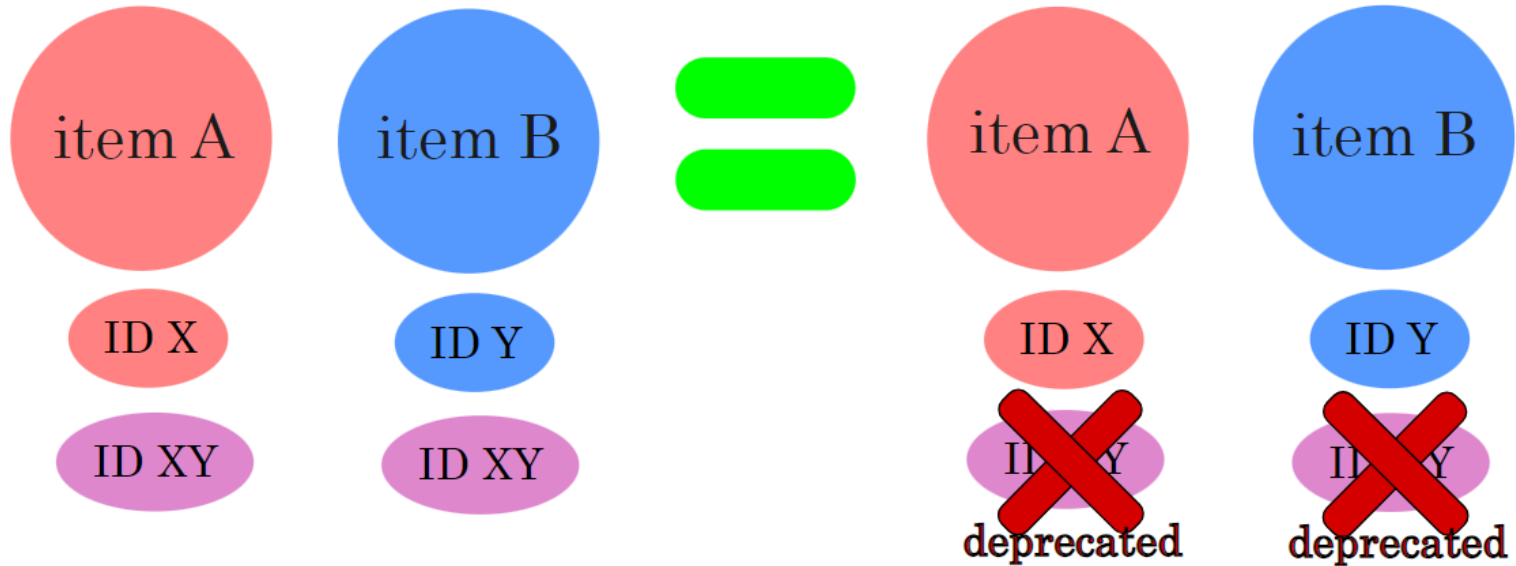




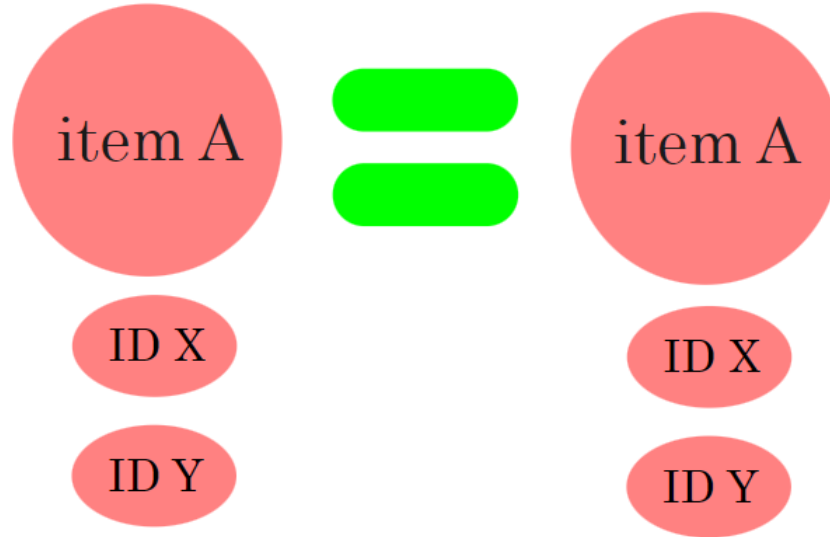
## 2) same ID in two items: conflated item



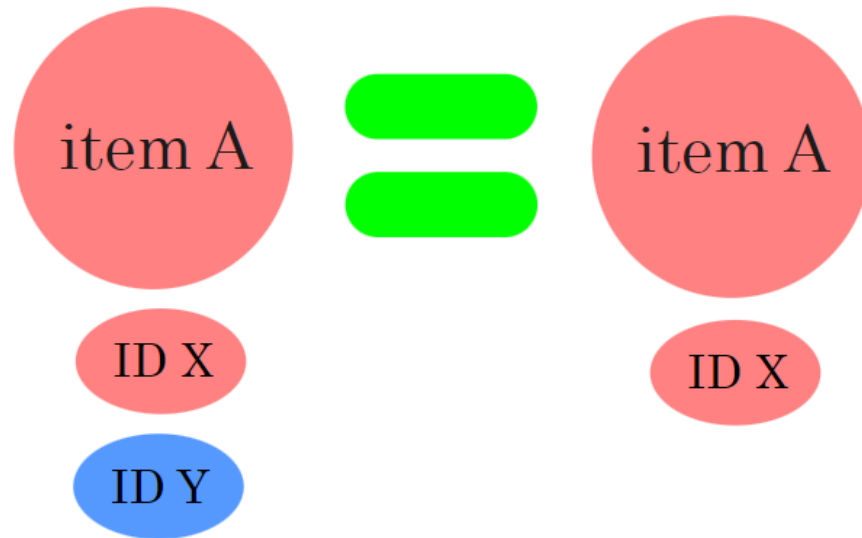
### 3) same ID in two items: conflated ID



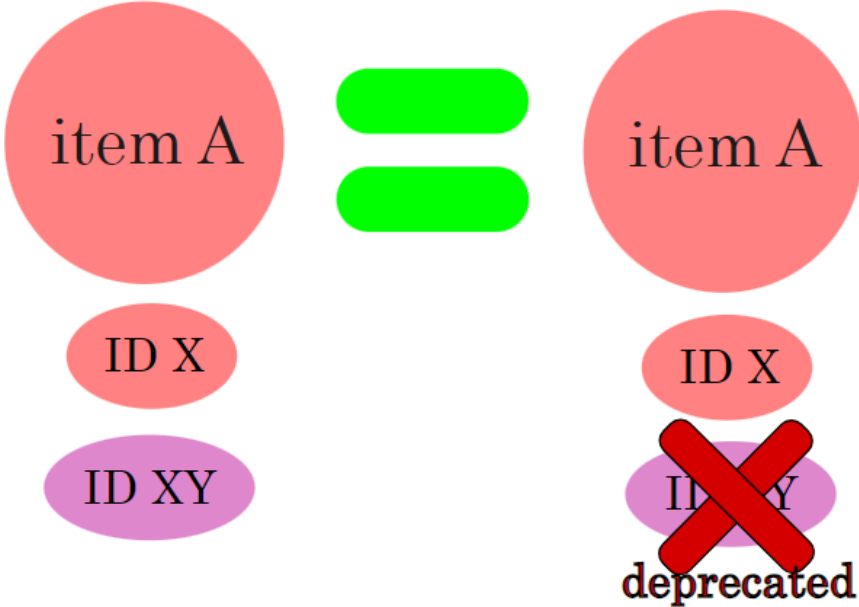
## 4) two IDs in same item: duplicate IDs



## 5) two IDs in same item: conflated item



# 6) two IDs in same item: conflated ID



# Normal workflow

If you are unable to obtain efficient corrections in the external database whose IDs you are checking the constraint violations:

- you can (and should) **solve** the violations “**same ID in two items**” because they are mostly caused by duplicate items, that you can personally merge (see also: [Wikidata:WikiProject Duplicates/VIAF members](https://www.wikidata.org/wiki/Wikidata:WikiProject_Duplicates/VIAF_members))
- you should **ignore** the violations “**two IDs in the same item**” because they are mostly caused by duplicate IDs, that you cannot personally merge



## 3) Solution

# Merges and splits

- Merges are performed through the **Merge gadget**: the procedure is started with a few clicks, is automatic and runs in a few seconds
- Splits are performed **manually** and consist in a long series of edits which remove from the conflated item all the non-pertinent data:
  - labels, descriptions, aliases
  - statements and identifiers
  - sitelinks
  - incoming links from other items





## 4) Issues

# 1) Mitigate **causes**: prevent low quality batches

- seemingly most duplications in recent years have been added by batches run through QuickStatements and OpenRefine
- no strict guideline presently exists about criteria to undo batches containing many mistakes (see [Wikidata:Edit groups](#))
- **proposal**: approve a new policy containing precise standards of quality for semi-automated batches

## 2) Improve **detection**: strengthen data round-tripping

- constraint violations are an effective way to find conflation and duplications; however, some of the issues they find are not in Wikidata, but in external databases, and thus cannot be solved directly
- the presence of abundant cases of duplicate IDs in external databases makes the lists of violations “two IDs in the same item” unusable, although they would be useful to discover some conflated items
- **proposal**: improve [data round-tripping](#), at least for the biggest databases (e.g. national authority files; cf. phab:[T312718](#))

### 3) Improve **solutions**: create a gadget for splits

- the split procedure involves a lot of manual edits and thus could be long and complex; this discourages some users from making splits
- moreover, when a split is performed, it is easy to forget checking some parts of the item, thus leaving the item partially conflated
- **proposal**: create a new gadget helping users in performing splits; it should guide the user step by step, so that no step is forgotten, and it should speed up the whole process



# Thanks for your attention!

Get in touch with me:

Camillo Carlo Pellizzari di San Girolamo  
[camillo.pellizzaridisangirolamo@sns.it](mailto:camillo.pellizzaridisangirolamo@sns.it)

For the 6 constraint schemes, see  
[https://commons.wikimedia.org/wiki/Category:Conflations\\_and\\_duplications\\_in\\_Wikidata\\_schemes](https://commons.wikimedia.org/wiki/Category:Conflations_and_duplications_in_Wikidata_schemes)

