String matching algorithms in OpenRefine clustering and reconciliation functions

A case study of person name matching

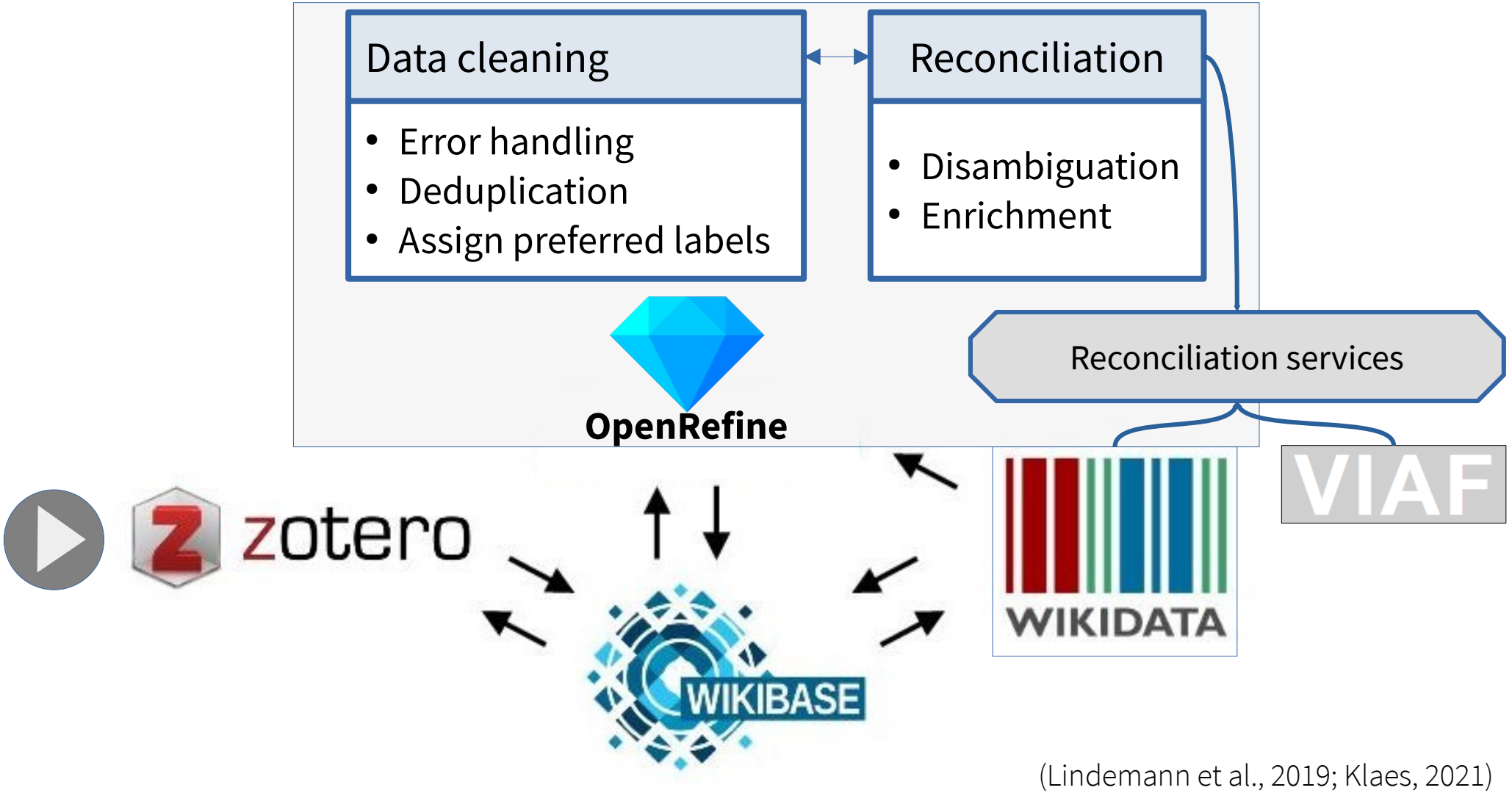
Christiane Klaes
Hildesheim University / University Library Braunschweig
c.klaes@tu-braunschweig.de

Agenda

1. Use case: domain knowledge base „LexBib“
2. String matching measures for person names
3. Clustering algorithms in OpenRefine
4. Matching algorithms in reconciliation services

1 Domain knowledge base „LexBib“

Data flow



(Lindemann et al., 2019; Klaes, 2021)

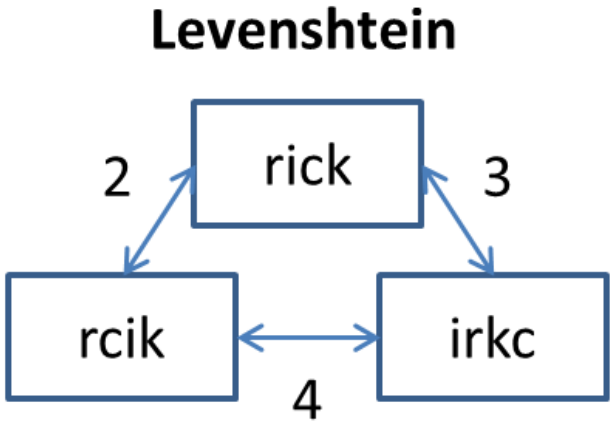
1 Domain knowledge base „LexBib“
Harmonizing name literals

Merging	Clustering	creatorName
Sue Atkins	Sue Atkins	Sue Atkins
	Sue Atkins	Sue B. T. Atkins
	Sue Atkins	B. T. Sue Atkins
	Sue Atkins	Sue Atkins
	Sue Atkins	B. T. Sue Atkins
	Sue Atkins	B. T. Sue Atkins
	Sue Atkins	B. T. S. Atkins

- Initials
- Double names
- Nicknames
- Order of name components
- Spelling errors

2 String matching measures for person names

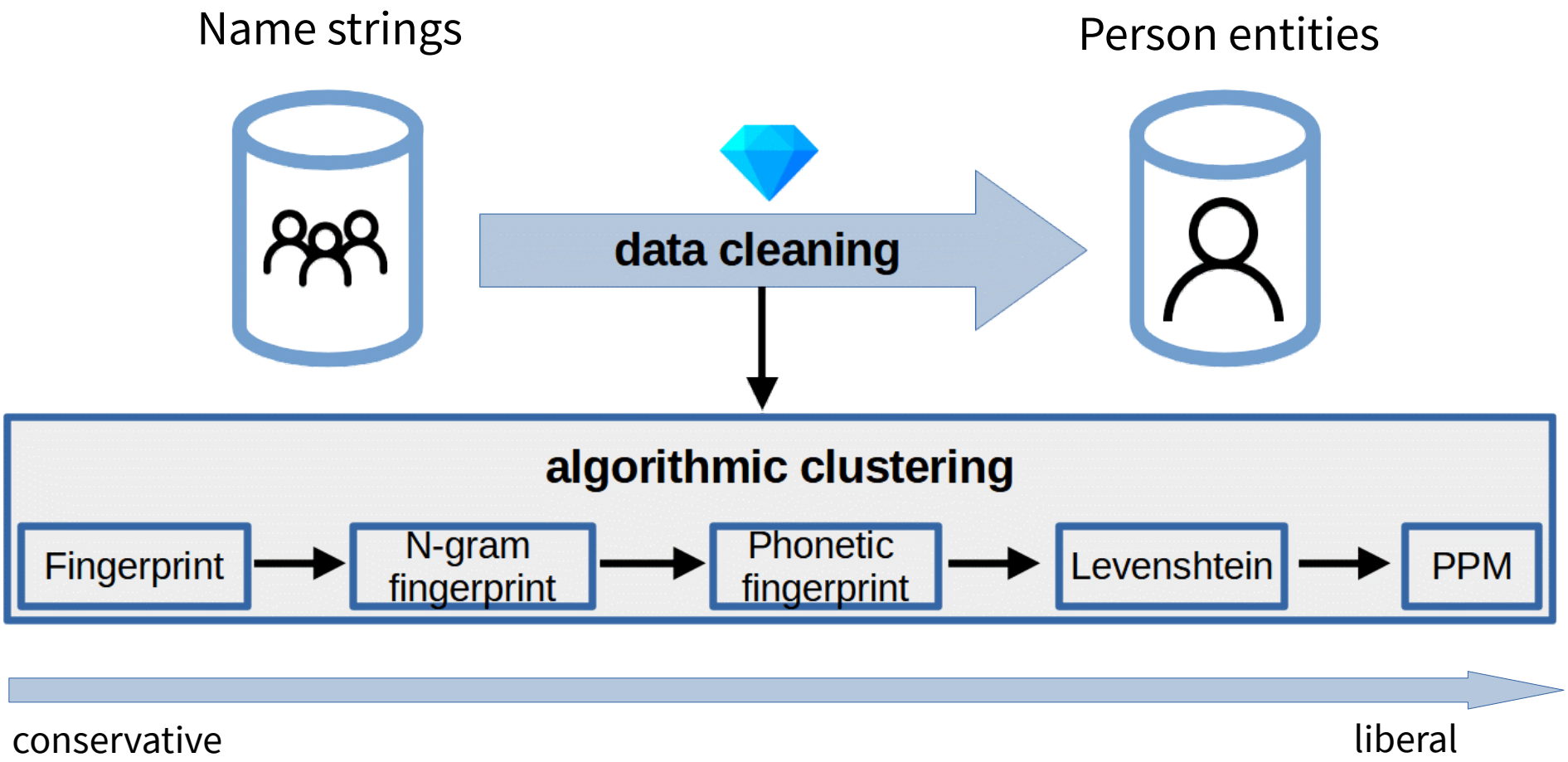
- Levenshtein
- N-grams, Skip-grams
- Phonetic measures:
  - Soundex, Metaphone *(for English)*
  - Cologne *(for German)*
- Jaro, JaroWinkler



Minerich, Richard. 2012. "Levenshtein Distance and the Triangle Inequality." Inviting Epiphany, September 04. https://devopedia.org/levenshtein-distance#Minerich-2012

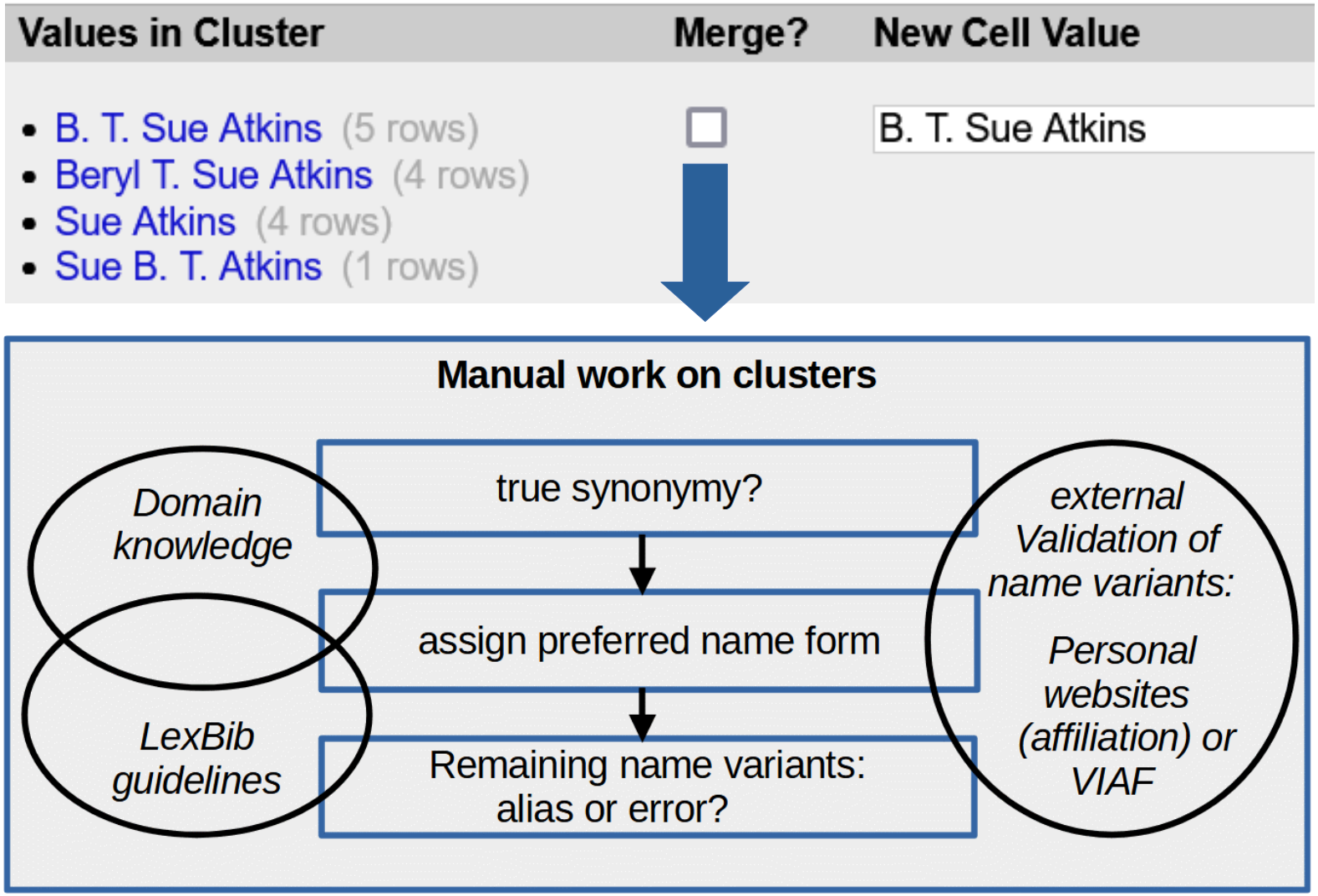
(Christen 2006; Recchia/Louwerse 2013; Pilania/Kumaran 2019)

3 Clustering algorithms in OpenRefine
Mixed methods approach



https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

3 Clustering algorithms in OpenRefine
Manual validation and post-processing

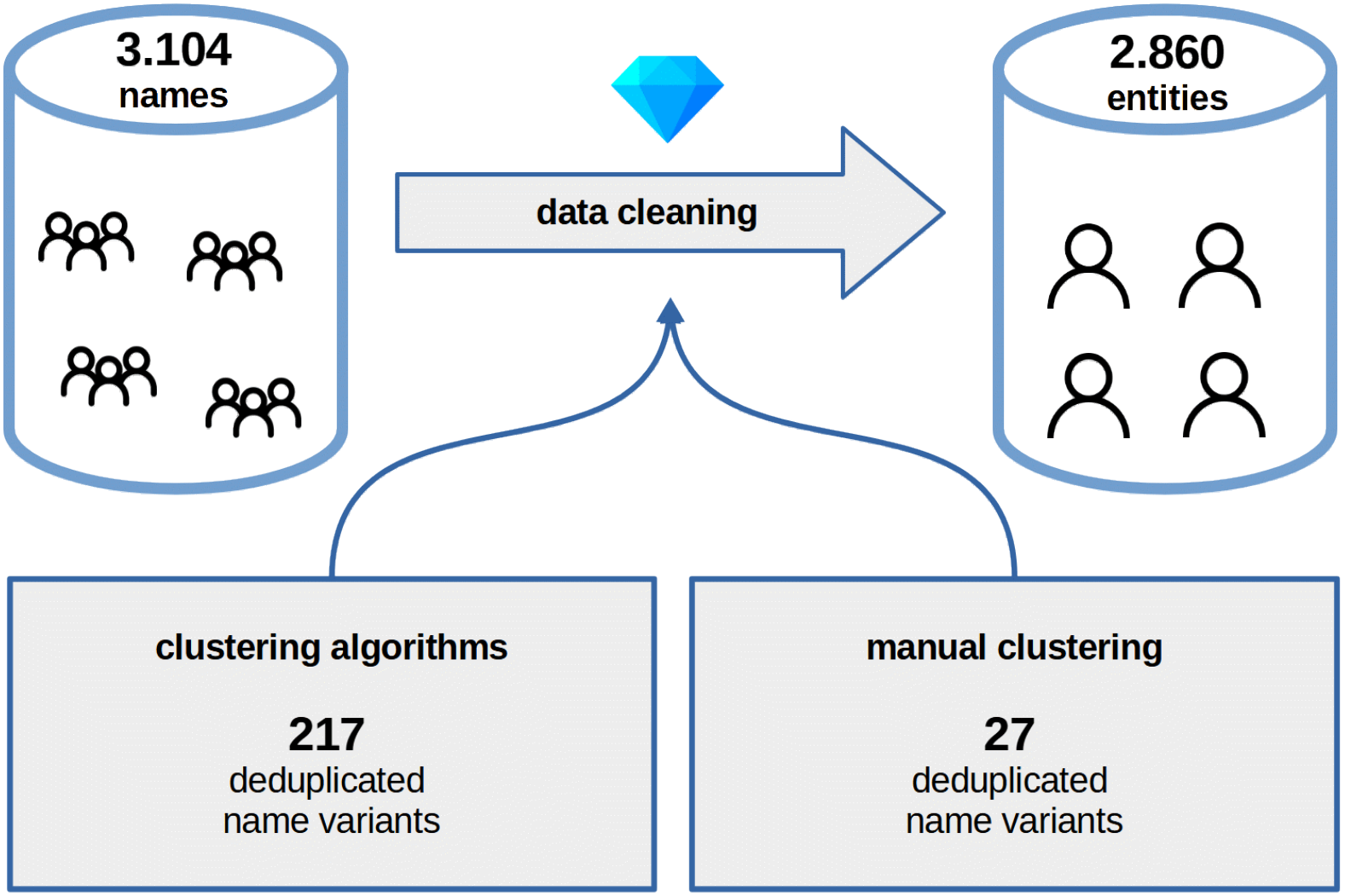


3 Clustering algorithms in OpenRefine
Results for clustering algorithms

Sample: 3.104 person names from LexBib

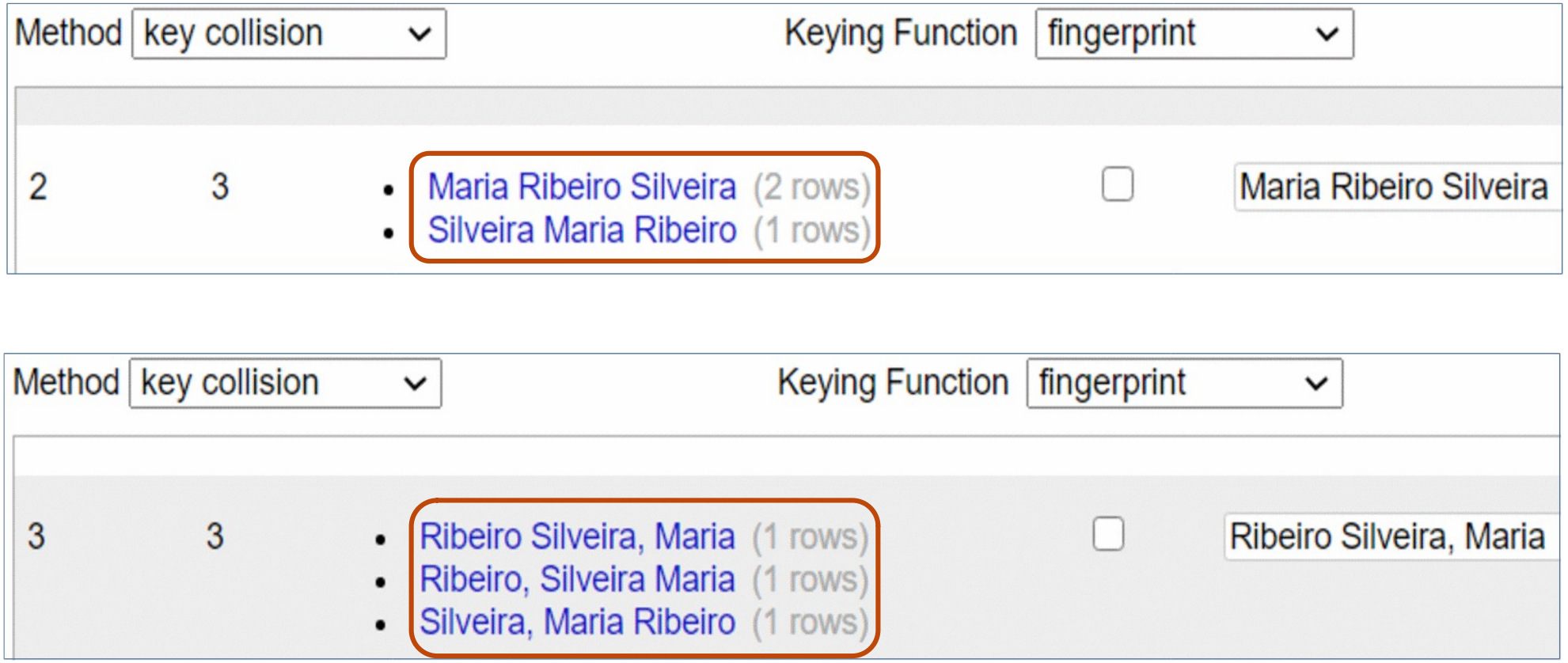
Clustering algorithm	Number of clusters	Precision of clusters	Typical deviations
Fingerprint	60	**1,0**	Agnès Tutin / Agnes Tutin
Bi-gram fingerprint	9	**1,0**	Sene-Mongaba / Sene Mongaba
Metaphone 3	66	**0,65**	Hannu Tommola / Hannu Tammala
Cologne	42	0,19	Hannu Tommola / Hannu Tuomola
Levenshtein	5	**1,0**	Franck Sajous / Frank Sajous
PPM (setting: 1.0 / 6)	15	**1,0**	Bolette S. Pedersen / Bolette Sandford Pedersen
PPM (setting: 2.0 / 4)	103	**0,43**	B. T. Sue Atkins / Sue Atkins / Beryl T. Sue Atkins

3 Clustering algorithms in OpenRefine
Automatic vs. manual clustering

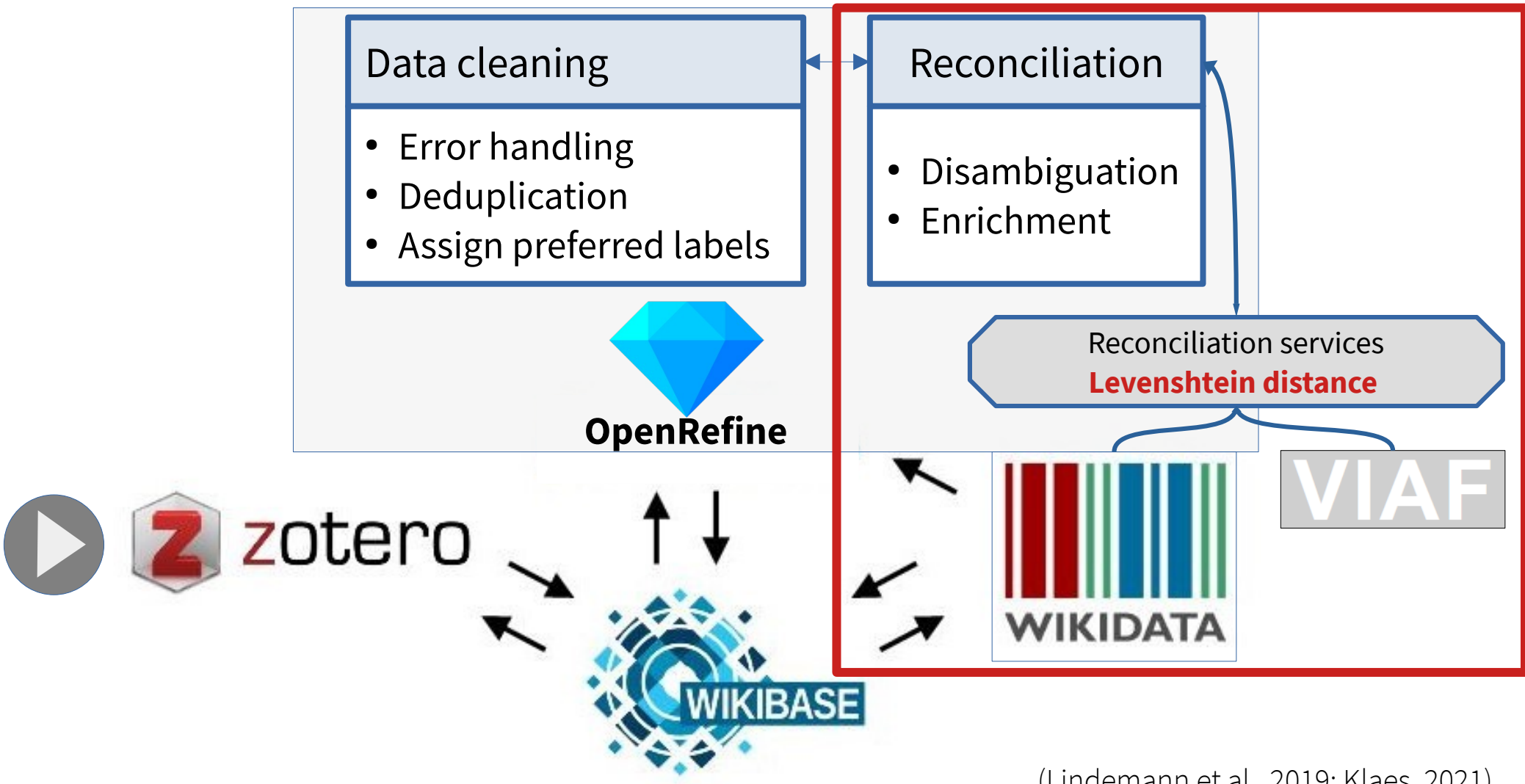


3 Clustering algorithms in OpenRefine
Lessons learned

Different name forms as input for clustering algorithms



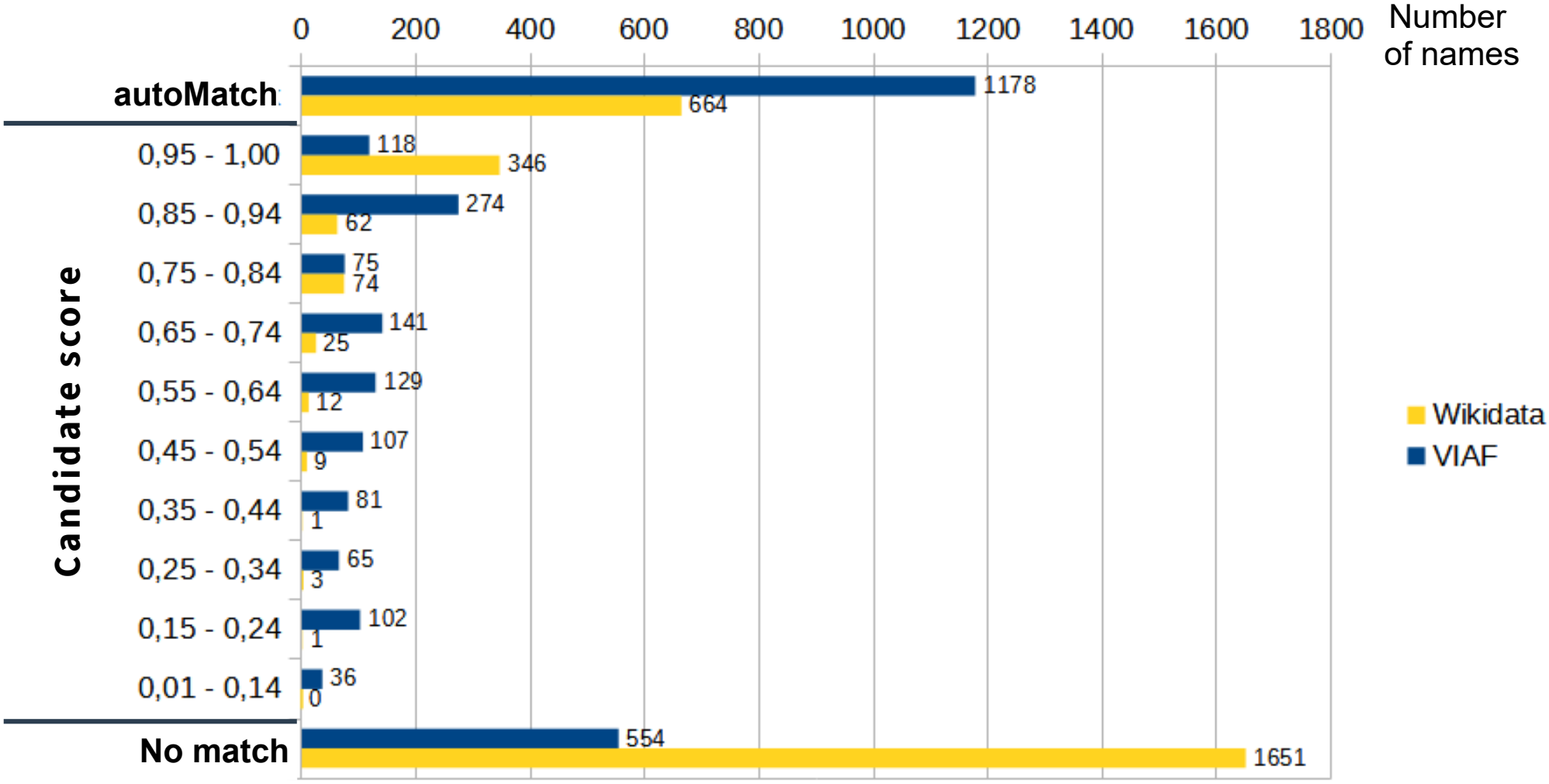
4 Matching algorithms in reconciliation services



(Lindemann et al., 2019; Klaes, 2021)

4 Matching algorithms in reconciliation services

Reconciliation results for Wikidata and VIAF



4 Matching algorithms in reconciliation services *Validation*

Judgement	Validation: precision of matches	
	Wikidata	VIAF
autoMatches	**0,9**	**0,91**
Linking candidates, Score 100 - 95	**0,29**	**1,00**
All linking candidates	**0,18**	**0,67**

4 Matching algorithms in reconciliation services
Misleading deviations in VIAF name strings

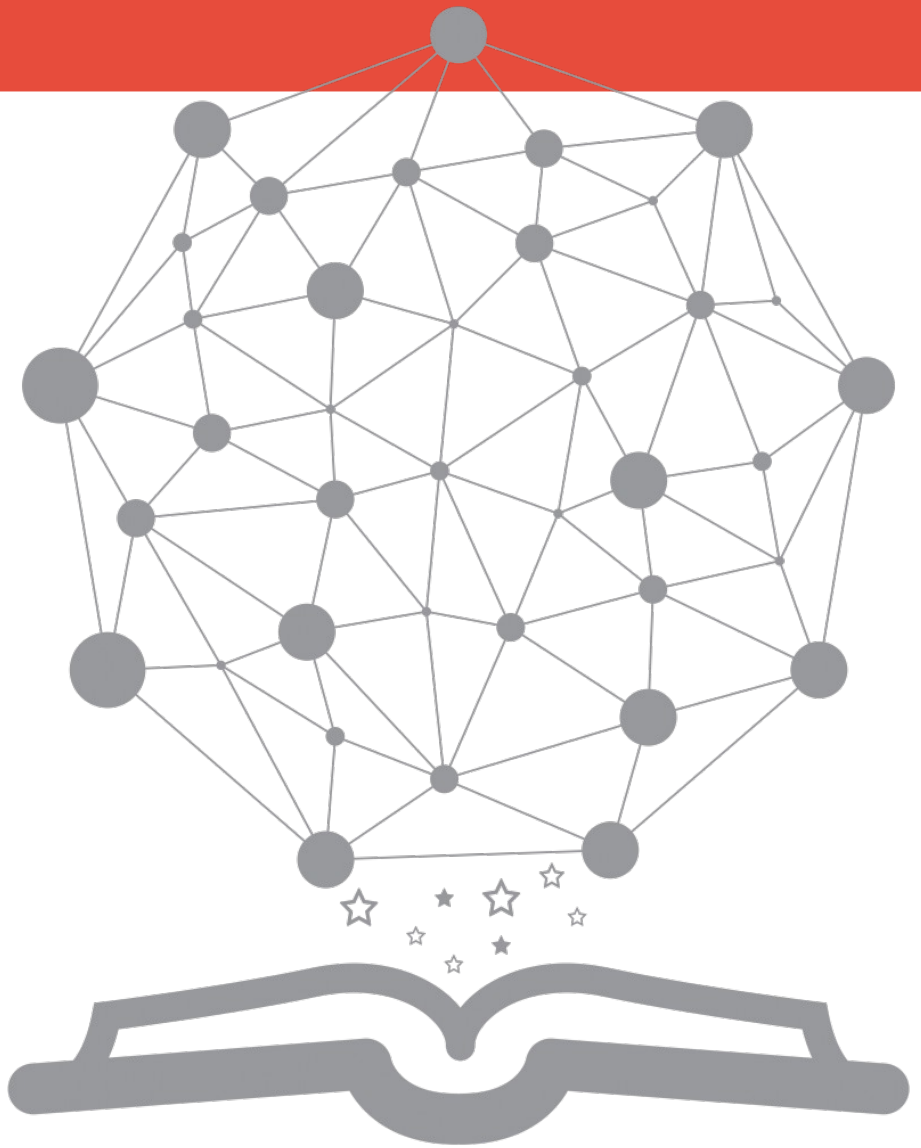
Birth year as part of name strings

Similarity score (Levenshtein distance)

Punctuation as part of name strings

Thank you!

Questions & comments welcome



Acknowledgements
This presentation is based on the author's Master thesis titled „Linked Open Data-Strategien zum Identity Management in einer Fachontologie – prototypische Entwicklung eines Workflows zur Aufbereitung und zum Interlinking von Personennamen", University of Hildesheim, August 2021.
Many thanks to PD Dr. Laura Giacomini and Prof. Dr. Ulrich Heid, and to Dr. David Lindemann.

References

Christen, Peter (2006): A Comparison of Personal Name Matching: Techniques and Practical Issues. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006, Hong Kong, China: IEEE, 290–294. http://doi.org/10.1109/ICDMW.2006.2

Christen, Peter (2012): Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin ; New York: Springer.

Delpeuch, Antonin (2019): A Survey of OpenRefine Reconciliation Services. http://arxiv.org/abs/1906.08092

Färber, Michael/Bartscherer, Frederic/Menne, Carsten/Rettinger, Achim (2017): Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. In: Zaveri, Amrapali et al. (Hrsg.), Semantic Web, 9 (1), 77–129.

Heath, Tom/Bizer, Christian (2011): Linked Data: Evolving the Web into a Global Data Space, Bd. 1. 1. Aufl. Morgan & Claypool. http://linkeddatabook.com/book

References

Klaes, Christiane (2021): Linked OpenData-Strategien zum Identity Management in einer Fachontologie - Prototypische Entwicklung eines Workflows zur Aufbereitung und zum Interlinking von Personennamen. Hildesheim: Universität Hildesheim.

Lindemann, David/Klaes, Christiane/Zumstein, Philipp (2019): Metalexicography as Knowledge Graph. In: Eskevich, Maria/De Melo, Gerard/Fäth, Christian/McCrae, John P./Buitelaar, Paul/Chiarcos, Christian/Klimek, Bettina/Dojchinovski, Milan (Hrsg.), OASICS, 70. https://doi.org/10.4230/OASIcs.LDK.2019.19

Pilania, Ankita/Kumaran, Gnanamani Mayyil Muthuil Muthu (2019): Comparative Study of Name Matching Algorithms. Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), Bharati Vidyapeeth, New Delhi: IEEE Computer Society, 1174–1178. https://ieeexplore.ieee.org/document/8991380

Pratter, Yves (2020): Clustering in Depth: Methods and Theory Behind the Clustering Functionality in OpenRefine. GitHub. https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Recchia, Gabriel/Louwerse, Max (2013): A Comparison of String Similarity Measures for Toponym Matching. COMP 2013 - ACM SIGSPATIAL International Workshop on Computational Models of Place, 5. November 2013, Orlando, Florida, USA, 54–61. https://doi.org/10.1145/2534848.2534850