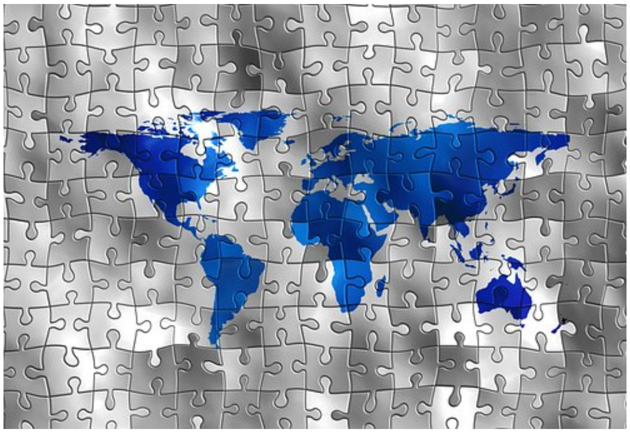
Institutional disambiguation for further countries - an exploration with extensive use of wikidata (project report).

August 7, 2018



Christine Rimmert

Contents

1 The goal of the project

An institutional disambiguation (defined as the assignment of author addresses in bibliometric databases to the corresponding research institutions) for German author addresses has been developed and processed by the Bibliometrics Group of Bielefeld University in the framework of the German Competence Center for Bibliometrics for years now[1]. The procedure is based on a manual effort in several steps whereby it is not scalable with regard to institutional disambiguation for other or even all countries. Due to needs in projects pursued by the Competence Center, this project is a study on the feasibility of an institutional disambiguation for further countries based on mainly automated steps in order to provide scalability, using six example countries.

The project is conducted in cooperation with Fraunhofer ISI in order to test two different approaches and evaluate the possibilities of combining them in the end. While ISI aims at creating a thesaurus with VantagePoint[2], Bielefeld pursues a wikidata[3]-based approach for the extraction of basic information concerning institutions as well as using wikidata for the institutional disambiguation of author addresses from the Web of Science (WoS).

The following steps have been performed:

- Creation of a gold standard by manual assignments of author addresses to wikidata entities
- Evaluation of wikidata as a source for basic information concerning institutional entities, development of an extraction procedure for basic information
- Evaluation of wikidata as a tool for institutional disambiguation, development and evaluation of a wikidata-based disambiguation procedure

[1] cf. Rimmert C, Schwechheimer H, Winterhager M. Disambiguation of author addresses in bibliometric databases - technical report. Bielefeld: 2017. https://pub.uni-bielefeld.de/record/2914944

[2] https://www.thevantagepoint.com/

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

2 Data & methods

For wikidata a full dump from May 2017 was used, supplemented by entities with a 'last modified' date between May 2017 and January 2018. Attributes concerning wikidata entities of interest were obtained by querying the wikidata API with python wikidata client library[4] in order to receive the current version at the time the basic data tables were created.

Author addresses of six example countries (Switzerland (CHE), Germany (DEU), Spain (ESP), Great Britain (GBR), Korea (KOR) und South Africa (ZAF)) were extracted from Web of Science raw data[5] with a restriction to the publication years 2012 to 2016.

Interaction with wikidata was performed with python while all steps within the disambiguation procedure were written in SQL or PL/SQL.

3 Creation of gold standard

In order to create a gold standard for the evaluation of the disambiguation procedures developed here, for each of the six example countries a random sample was drawn from the distinct author addresses in the Web of Science in the publication period 2012 to 2016 (2.5%, at least 500 addresses) with a minimum frequency threshold of 10. Table 1 shows the resulting number of addresses per example country. For CHE and ZAF the minimum limit applies.

These addresses were assigned manually to the corresponding wikidata entities (in case of existence, otherwise a corresponding url was recorded). This step was processed by Fraunhofer ISI for CHE, GBR and KOR, for DEU, ESP and ZAF by the Bibliometrics Group of Bielefeld University.

In order to simplify the manual assignments for the gold standard by creating suggestion lists, a first simple matching was performed: the first part of the address (separated by ',') – describing the main institution in most

[4] https://pypi.python.org/pypi/Wikidata

[5] 17th calendar week 2017

country	**# addresses**
CHE	500
DEU	1,471
ESP	1,158
GBR	1,195
KOR	767
ZAF	500

Table 1: Number of addresses (gold standard).

country	**# dist. addresses**	**thereof matched**	**in %**	**thereof uniquely matched**	**in %**
CHE	500	279	55.8	171	34.2
DEU	1,471	1,064	72.3	908	61.7
ESP	1,158	609	52.6	486	42.0
GBR	1,195	969	81.1	782	65.4
KOR	767	679	88.5	468	61.0
ZAF	500	447	89.4	430	86.0

Table 2: Results of the suggestion list matching (gold standard).

cases of WoS addresses – was matched with wikidata labels taking only exact matches into account. Addresses as well as wikidata labels were prepared by a transformation step (e.g. deletion of stop words and special characters, permutations, replacements and a special stemming – the transformation step of the institutional disambiguation for Germany was reused for this purpose[6]).

Table 2 shows the number of addresses with at least one assignment and the number of addresses with unique assignment – regardless of correctness. The number of addresses assigned differs among countries while in all countries a large proportion of addresses could be assigned.

Table 3 shows the share of correctly assigned addresses on the number of uniquely assigned addresses.

[6] For details of the transformation see chapter 3.3 in Rimmert, C. et al., Disambiguation of author addresses in bibliometric databases – technical report. Bielefeld: 2017, p.6f., https://pub.uni-bielefeld.de/record/2914944

country	**# addresses with unique assignment**	**thereof correct**	**in %**
CHE	171	167	97.7
DEU	908	902	99.3
ESP	486	483	99.4
GBR	782	768	98.2
KOR	468	447	95.5
ZAF	430	425	98.8

Table 3: Correct unique assignments (gold standard).

country	**# addresses with more than one assignment**	**thereof correct assignment contained**	**in %**
CHE	108	103	95.4
DEU	156	144	92.3
ESP	123	117	95.1
GBR	187	181	96.8
KOR	211	207	98.1
ZAF	17	13	76.5

Table 4: Correct assignment contained in suggestions (gold standard)

For addresses with at least two assignments to different wikidata entities table 4 shows the number (and share) of addresses, for which the correct assignment is contained in the suggestions.
Finally, table 5 shows the number of addresses per country where the corresponding institution is contained in wikidata, providing a hint for the maximum achievable recall for disambiguation procedures based on wikidata. As wikidata is changing every day, this can of course only be a snapshot.

In summary, even with a simple matching several addresses can be assigned to wikidata entities where in case of unique assignments the matching is most likely correct and in case of ambiguous assignments, the correct one is among the suggestions in most cases.
For 89.5% of the addresses from ESP the corresponding institution is recorded in wikidata – this is the case for more than 95% of the goldstandard addresses from all other countries. This may be a hint at different wikidata entity cov-

country	# dist. addresses	institution in wikidata	in %
CHE	500	477	95.4
DEU	1,471	1,417	96.3
ESP	1,158	1,036	89.5
GBR	1,195	1,156	96.7
KOR	767	760	99.1
ZAF	500	477	95.4

Table 5: Institutions contained in wikidata (gold standard).

erage for different countries.

The number of addresses matched also differs clearly among the example countries: while for CHE and ESP only 52% (respectively 55%) could be matched, ZAF is on top with already 89%.

In cases of missing allocations, for instance, the following reasons appeared:

1. the research institution mentioned is not covered by wikidata
2. a subdivision is mentioned instead of the main institution and the subdivision name is not contained in wikidata as name variant for the main institution
3. a wikidata name variant is not exactly like the first part of the address but a substring of it – or the other way around
4. the name variant mentioned in the address is not covered in wikidata
5. the first part of the address does not contain enough information for a unique allocation

Multiple assignments appeared, for instance, due to the following reasons:

1. the name of the institution is ambiguous
2. wikidata duplicates
3. the first part of the address does not contain enough information for a unique allocation
4. the first part of the address mentions more than one institution

While in the first case of reasons for missing allocations an allocation based on wikidata definitely has to fail, other reasons for missing allocations can be addressed by automated disambiguation procedures.

The gold standard described in this section was used for the evaluation of the wikidata-based institutional disambiguation procedure designed in the context of this project (section 5).

4 Wikidata as a source for basic information concerning institutional entities

The usefulness of an institutional disambiguation based on wikidata depends to a large extent on the wikidata coverage of the entities of interest (research institutions mentioned in author addresses in WoS) and their attributes.
Thereby, 'research institutions' mentioned in WoS addresses are not only research institutions in the common sense (such as universities or Max Planck Institutes) but also companies, societies, research networks, hospitals and so on. All these are referred to as 'research institutions' in the following, due to simplification of notation.

4.1 Entity coverage

A first hint concerning entity coverage is given by the insights emerging from the creation of the gold standard in table 5: although the percentage of addresses with institutions covered in wikidata differs among the countries (with a minimal value for ESP), the coverage seems to be a solid foundation.

Of course this is only a sample of addresses from the example countries and a check with complete lists of research institutions would be a better option for a wikidata coverage analysis – but in the absence of such lists of research institutions (in the sense described above) for all countries, entity coverage in wikidata cannot be checked in a complete and easy way.
However, for research institutions mentioned in German author addresses in WoS, such a list (approximately complete) is available due to the German institutional disambiguation. Therefore, Germany was used as an example (bearing in mind that there may be differences in entity coverage among countries) to get an impression of entity coverage in terms of research institutions mentioned in WoS author addresses.

Therefore, identifiers used in the institutional disambiguation for Germany were assigned manually to the corresponding wikidata id if existent, which allows the observation of the share of research institutions contained in wikidata among research institutions mentioned in WoS author addresses.
Existing as well as already closed institutions were included here. Institu-

type of institution (sector)	**share institutions recorded in wikidata in %**
Fraunhofer-Gesellschaft	89.3
Helmholtz-Association	91.7
Leibniz Association	97.1
Max-Planck Society	89.1
Universities	98.4
Universities of Applied Sciences	100
Companies	46.0
Federal and State Government R&D institutions	85.7
– State Government R&D inst.	94.0
– Federal Government R&D inst.	75.6
Others (e.g. hospitals without university hospitals, public inst.)	62.3

Table 6: Entity coverage for German institutions by sector.

tions may be contained more than once due to sector changes or sector hybrid status (where sector defines the type of the institution).

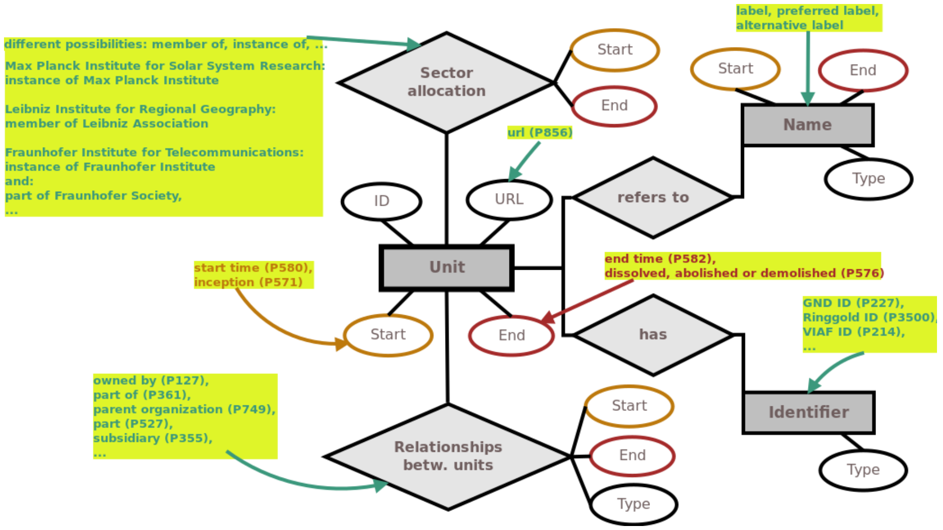
Table 6 shows the results per sector. It should be mentioned here that the upper half of table 6 (representing the core of the German academic system) is responsible for more than 85% of the total publication output of Germany as covered in WoS. Differences among sectors are clearly visible. It could also be observed that already closed institutions are more likely not recorded than existing ones (which is expectable as wikidata is rather new and does not aim at complete historicisation).

In addition, this check leads to the assumption of wikidata providing a solid basis of research institutions.

4.2 Property coverage

In addition to the question of entity coverage, it is also of relevance if attributes (called 'properties' in wikidata) and relations of interest are contained in wikidata. For attributes and relations used in the institutional disambiguation procedure for Germany, belonging properties were researched manually in order to examine the existence of suitable properties in wikidata[7].

Figure 1 shows an entity relationship model of attributes and relations of interest together with corresponding wikidata properties:



different possibilities: member of, instance of, ...
Max Planck Institute for Solar System Research:
instance of Max Planck Institute
Leibniz Institute for Regional Geography:
member of Leibniz Association
Fraunhofer Institute for Telecommunications:
instance of Fraunhofer Institute
and:
part of Fraunhofer Society,
...
Sector allocation
Start
End
label, preferred label,
alternative label
Start
End
Name
Type
url (P856)
ID
URL
refers to
Unit
start time (P580),
inception (P571)
end time (P582),
dissolved, abolished or demolished (P576)
Start
End
has
GND ID (P227),
Ringgold ID (P3500),
VIAF ID (P214),
...
owned by (P127),
part of (P361),
parent organization (P749),
part (P527),
subsidiary (P355),
...
Relationships betw. units
Start
End
Type
Identifier
Type

Figure 1: Attributes and relations with corresponding wikidata properties.

As can be seen, suitable properties exist in wikidata where sometimes more than one property contains information for an attribute or relation. The sector allocation in the institutional disambiguation for Germany is interpreted in different ways in wikidata: sometimes the sector is given as a parent organization (e.g. Fraunhofer Institute for Telecommunications), in other cases the single institute is linked to the corresponding sector via the 'member of' property or the sector is given as the type of the institution ('instance of' property).

[7] https://www.wikidata.org/wiki/Wikidata:List_of_properties

Data concerning research institutions (attributes and relationships, either used in the procedure or providing additional information for analyses/evaluations on the basis of research institutions) is referred to as 'basic data' in the following.

4.3 Table schema & wikidata properties for basic data

Of course, wikidata provides a large amount of basic data for research institutions, so the selection of information for the basic data tables is a restriction to information of interest in this context. We decided to include information concerning names, geographical attributes, urls, identifiers (to enable links to other sources), descriptions, types of research institutions (classification based on wikidata property 'instance of' (P31)), relations among research institutions and structural changes over time. To store this basic data, the table schema displayed in figure 2 was designed. A brief description of the tables is given in the following:

- **IW_UNIT:** main table to store research institution entities on any hierarchical level with their url and start and end date.

- **IW_NAME:** names for units – for every unit several names in different languages are given. We decided against including the language flag because this leads to storing the same name with many different language flags for many different units, therefore leading to duplicate information. Adding the language flag (accepting to use much more space) is of course possible without difficulties.

- **IW_IDENTIFIER:** different identifiers may be given for one unit. This is a list containing all wikidata properties included:
  - P227 (GND ID)
  - P3500 (Ringgold ID)
  - P268 (BnF ID)
  - P214 (VIAF ID)
  - P1662 (DOI Prefix)
  - P2427 (GRID ID)
  - P213 (ISNI)
  - P646 (Freebase ID)

  - P950 (BNE ID)
  - P1566 (GeoNames ID)
  - P4096 (RePEc institute ID)
  - P2002 (Twitter username)
  - P2740 (ResearchGate institute ID)
  - P2013 (Facebook profile ID)

  Again, of course there are more possibilities – we use this selection as an example. Deleting properties or adding further properties of interest is always possible without difficulty. Property ids (appearing in IW_IDENTIFIER as well as IW_GEO and IW_RELATION) can be looked up in table 'IW_PROPERTY'.

- **IW_DESCRIPTION:** description given in wikidata (in different languages). For the reason already mentioned for table 'IW_NAME' the language flag was excluded (and can be added without difficulty also in this case).

- **IW_TYPE:** wikidata classification provided by property P31 ('instance of'). One wikidata entity may have no, exactly one or more P31-values. For P31 values, wikidata ids as well as labels are given (English label if existent, other language otherwise).

- **IW_GEO:** several geographical attributes:
  - P281 (postal code)
  - P625 (coordinate location)
  - P740 (location of formation)
  - P131 (located in the administrative territorial entity)
  - P159 (headquarters location)
  - P495 (country of origin)
  - P17 (country)
  - P276 (location)
  - P969 (located at street address)
  - P669 (street)

  Property ids can be looked up in table 'IW_PROPERTY'.

- **IW_RELATION:** relations among units. The following properties were included so far:
  - P361 (part of)

  - P749 ((has) parent organization)
  - P527 ((has) part)
  - P355 (subsidiary)
  - P463 (member of)

  where the first four cover hierarchical relationships (the first two from a child's, the latter two from a parent's point of view) and the last is an example of a completely different relational structure. Again, property ids can be looked up in table 'IW_PROPERTY'. In a first version only direct children and direct parents of entities were included (not, e.g., parents of parents or other children of parents). In case of existence, start and end dates for relations are also provided. In case of the existence of a fact both from a child's and a parent's point of view, it is contained just once in order to avoid redundancy.

- **IW_DESCENT:** time sequences of units (in contrast to relations among units within a time period). In this case, a transition date is given (if existent) in addition to predecessor and successor units. In contrast to IW_RELATION belonging property ids are not given (because their distinction does not seem to be of valuable interest in this context). Underlying properties are
  - P156 (followed by)
  - P1366 (replaced by)
  - P155 (follows)
  - P1365 (replaces)

  where the first two show a predecessors, the last two a successors point of view. Like for IW_RELATION, in case of the existence of a fact both from a predecessors and a successors point of view, it is contained just once in order to avoid redundancy.

- **IW_PROPERTY:** look-up table for property ids,
  used in IW_IDENTIFIER, IW_GEO and IW_relation.

- **IW_WIKIDATA_DUPLICATES:** manually collected duplicates detected while procedure development. WIKIDATA_ID_PREF is treated as the preferred entity – meaning that other wikidata ids are replaced by this preferred one in case of occurrence.

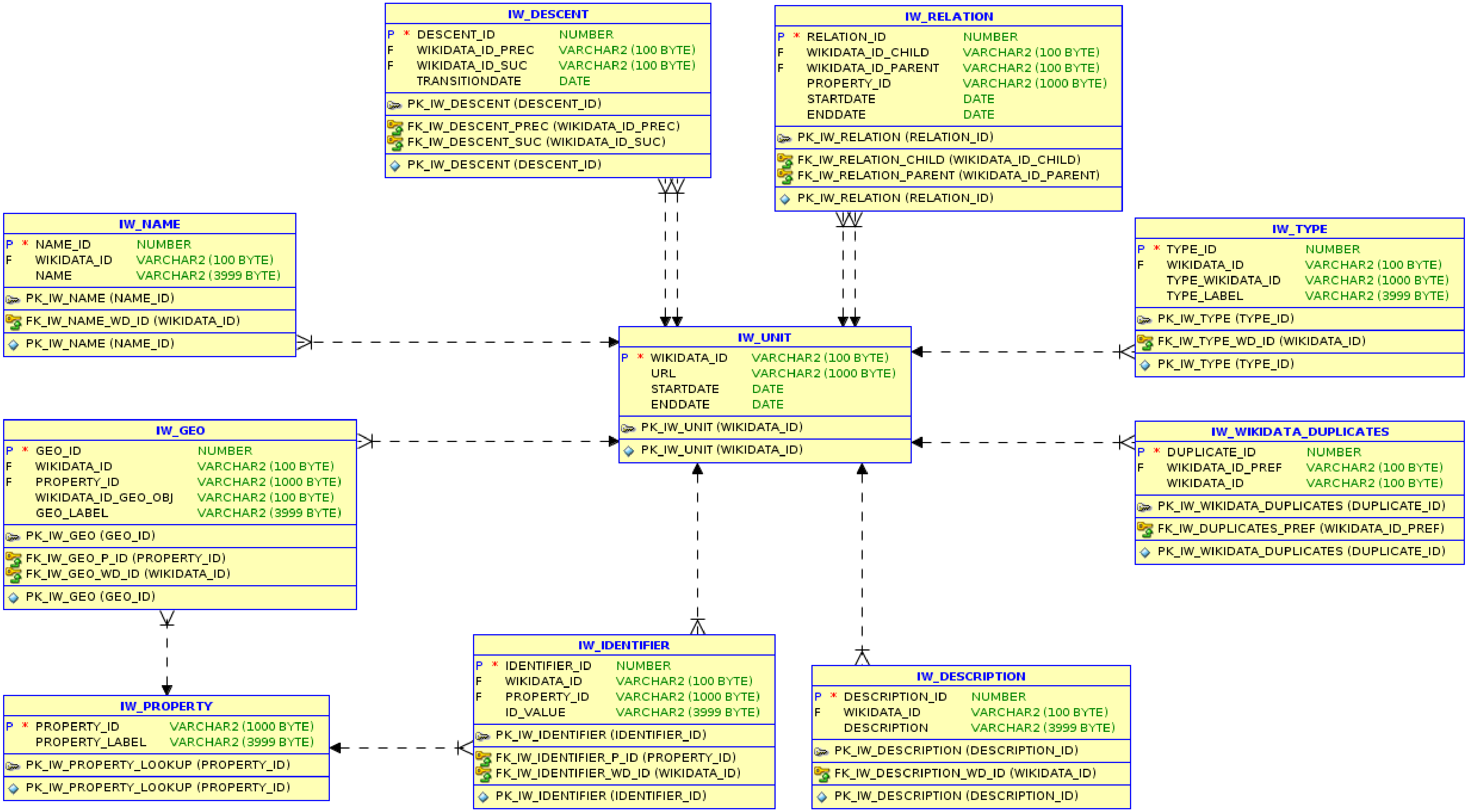

IW_DESCENT
P * DESCENT_ID NUMBER
F WIKIDATA_ID_PREC VARCHAR2 (100 BYTE)
F WIKIDATA_ID_SUC VARCHAR2 (100 BYTE)
TRANSITIONDATE DATE
PK_IW_DESCENT (DESCENT_ID)
FK_IW_DESCENT_PREC (WIKIDATA_ID_PREC)
FK_IW_DESCENT_SUC (WIKIDATA_ID_SUC)
PK_IW_DESCENT (DESCENT_ID)
IW_RELATION
P * RELATION_ID NUMBER
F WIKIDATA_ID_CHILD VARCHAR2 (100 BYTE)
F WIKIDATA_ID_PARENT VARCHAR2 (100 BYTE)
PROPERTY_ID VARCHAR2 (1000 BYTE)
STARTDATE DATE
ENDDATE DATE
PK_IW_RELATION (RELATION_ID)
FK_IW_RELATION_CHILD (WIKIDATA_ID_CHILD)
FK_IW_RELATION_PARENT (WIKIDATA_ID_PARENT)
PK_IW_RELATION (RELATION_ID)
IW_NAME
P * NAME_ID NUMBER
F WIKIDATA_ID VARCHAR2 (100 BYTE)
NAME VARCHAR2 (3999 BYTE)
PK_IW_NAME (NAME_ID)
FK_IW_NAME_WD_ID (WIKIDATA_ID)
PK_IW_NAME (NAME_ID)
IW_TYPE
P * TYPE_ID NUMBER
F WIKIDATA_ID VARCHAR2 (100 BYTE)
TYPE_WIKIDATA_ID VARCHAR2 (1000 BYTE)
TYPE_LABEL VARCHAR2 (3999 BYTE)
PK_IW_TYPE (TYPE_ID)
FK_IW_TYPE_WD_ID (WIKIDATA_ID)
PK_IW_TYPE (TYPE_ID)
IW_UNIT
P * WIKIDATA_ID VARCHAR2 (100 BYTE)
URL VARCHAR2 (1000 BYTE)
STARTDATE DATE
ENDDATE DATE
PK_IW_UNIT (WIKIDATA_ID)
PK_IW_UNIT (WIKIDATA_ID)
IW_GEO
P * GEO_ID NUMBER
F WIKIDATA_ID VARCHAR2 (100 BYTE)
F PROPERTY_ID VARCHAR2 (1000 BYTE)
WIKIDATA_ID_GEO_OBJ VARCHAR2 (100 BYTE)
GEO_LABEL VARCHAR2 (3999 BYTE)
PK_IW_GEO (GEO_ID)
FK_IW_GEO_P_ID (PROPERTY_ID)
FK_IW_GEO_WD_ID (WIKIDATA_ID)
PK_IW_GEO (GEO_ID)
IW_WIKIDATA_DUPLICATES
P * DUPLICATE_ID NUMBER
F WIKIDATA_ID_PREF VARCHAR2 (100 BYTE)
WIKIDATA_ID VARCHAR2 (100 BYTE)
PK_IW_WIKIDATA_DUPLICATES (DUPLICATE_ID)
FK_IW_DUPLICATES_PREF (WIKIDATA_ID_PREF)
PK_IW_WIKIDATA_DUPLICATES (DUPLICATE_ID)
IW_IDENTIFIER
P * IDENTIFIER_ID NUMBER
F WIKIDATA_ID VARCHAR2 (100 BYTE)
F PROPERTY_ID VARCHAR2 (1000 BYTE)
ID_VALUE VARCHAR2 (3999 BYTE)
PK_IW_IDENTIFIER (IDENTIFIER_ID)
FK_IW_IDENTIFIER_P_ID (PROPERTY_ID)
FK_IW_IDENTIFIER_WD_ID (WIKIDATA_ID)
PK_IW_IDENTIFIER (IDENTIFIER_ID)
IW_PROPERTY
P * PROPERTY_ID VARCHAR2 (1000 BYTE)
PROPERTY_LABEL VARCHAR2 (3999 BYTE)
PK_IW_PROPERTY_LOOKUP (PROPERTY_ID)
PK_IW_PROPERTY_LOOKUP (PROPERTY_ID)
IW_DESCRIPTION
P * DESCRIPTION_ID NUMBER
F WIKIDATA_ID VARCHAR2 (100 BYTE)
DESCRIPTION VARCHAR2 (3999 BYTE)
PK_IW_DESCRIPTION (DESCRIPTION_ID)
FK_IW_DESCRIPTION_WD_ID (WIKIDATA_ID)
PK_IW_DESCRIPTION (DESCRIPTION_ID)

Figure 2: Table schema basic data.

This is the selection of attributes chosen for the project – nevertheless, more information is available and can be added if required.

4.4 Check of basic data values

The basic data tables were filled with all entities (and their attributes and relations of interest) showing up in the first matching step of the procedure (which is described in section 5.1).

In iw_unit 17,755 wikidata entities are recorded, for 11,624 of them a url is given (65.47 %), 3,989 have a start date and 1,290 an end date.
Urls were checked concerning HTTP status code with a python script. Results are presented in table 7. By far most of the urls are valid and produce a status code of the form 2xx [8].

# of urls	HTTP statuscode
10,443	2xx (success)
629	4xx (client errors)
204	max retries exceeded
140	other errors
112	timeout error
28	5xx (server errors)
9	3xx (redirection)
2	9xx (proprietary codes)

Table 7: Url check: status codes.

Tables 8 and 9 show the extent of completeness of geographical attributes and identifiers. For all attributes and also identifiers coverage is far from complete. In case of identifiers it has to be taken into account that not every identifier is applicable for every type of institution.
For a random sample of 100 wikidata entities country and city values were checked concerning correctness: both city and country were correct in all cases of existence (79 for country, 81 for city).

[8] This check was performed on March 09, 2018

property ID	property label	# of wikidata entities	in%
P17	country	12,341	69.51
P131	located in the administrative territorial entity	7,816	44.02
P625	coordinate location	7,268	40.93
P159	headquarters location	4,966	27.97
P969	located at street address	631	3.55
P740	location of formation	408	2.3
P281	postal code	399	2.25
P276	location	289	1.63
P495	country of origin	84	0.47
P669	street	77	0.43

Table 8: Completeness of geographical attribute values: number of wikidata entities in iw_unit with the corresponding attribute in absolute numbers as well as in % of the total number of wikidata entities in iw_unit.

property ID	property label	# of wikidata entities	in%
P2427	GRID ID	7,248	40.82
P213	ISNI	6,212	34.99
P646	Freebase ID	4,712	26.54
P214	VIAF ID	3,744	21.09
P3500	Ringgold ID	2,917	16.43
P227	GND ID	2,829	15.93
P268	BnF ID	1,351	7.61
P2002	Twitter username	963	5.42
P1566	GeoNames ID	945	5.32
P2013	Facebook profile	400	2.25
P950	BNE ID	385	2.17
P1662	DOI Prefix	201	1.13
P2740	ResearchGate institute ID	72	0.41
P4096	RePEc institute ID	8	0.05

Table 9: Completeness of identifier values: number of wikidata entities in iw_unit with the corresponding identifier in absolute numbers as well as in % of the total number of wikidata entities in iw_unit.

In iw_relation 12,652 relations are recorded. Table 10 shows the number of relations by property id. 2,101 wikidata entities appear as parents in iw_relation (which means they have at least one child assigned) while 3,458 wikidata entities are recorded as children (meaning they have at least one parent entity).

property ID	**property label**	**# of wikidata entities**
P355	subsidiary	5,210
P361	part of	1,283
P749	(has) parent organization	1,981
P463	member of	2,055
P527	(has) part	2,123

Table 10: Number of relations recorded in iw_relation by property.

A random sample of 100 wikidata entities appearing either as parent or child in iw_relation was checked concerning all relations recorded for this specific wikidata entity. Four entities turned out to be no institution, in two cases no information could be found to verify the relation information. For 84 entities all relations recorded have been correct and of the desired type (relation in the sense of subdivision). In the remaining cases, there is (among other correct ones) at least one incorrect/undesired relation recorded (2 cases) or there are relations that are rather undesired in this context (relations in terms of geographical ties, stock index relations or humans recorded as 'child of' institutions). A more detailed check of the different properties included may be of value.

Furthermore, for three types (university hospitals, academic hospitals and affiliated institutes), handling of relations to universities in wikidata was checked for a sample set. In case of university hospitals and medical faculties, a relation (e.g. in terms of 'part of') to the university is desired in most contexts while affiliated institutes and hospitals are independent from the corresponding university – here a relation may be helpful but not in the sense of a 'part of' relation.
For only six of 26 university hospitals an 'is part of' relation is recorded in wikidata, in one case the relation to the university is given by property 'applies to jurisdiction' (P1001). Thus if university hospitals should be assigned

to the corresponding hospitals these relations have to be added in a further step (not yet included in the procedure/extraction of basic data here).
In case of academic hospitals none of the 26 hospitals is related to the corresponding university via 'is part of' which is the desired result. Two of them are classified as 'teaching hospital' via 'instance of' (P31).
The case of affiliated institutes is similar: here also none of the 16 affiliated institutes checked has an 'is part of' relation to the corresponding university. In one case the affiliated institute is declared as 'An-Institut (Q482329)' and the corresponding university is given. This is the optimal way of handling the situation in wikidata.

The table iw_descent covers 216 succession cases where a transition date is given for only 14 of them. 83 wikidata entities occur as successors (meaning they have at least one predecessor) while 121 predecessors are recorded (having at least on successor).
30 entries of iw_descent were checked manually. One of the entries is an error (one wikidata entity is just a subdivision of the other, no structural change), another one is unclear. For the remaining 28 (correct) entities the following underlying cases of structural changes could be identified (with frequency in brackets):

- replacements (9)
- incorporations (8)
- fusions (7)
- spin offs (1)
- name changes (2)

The classification given by wikidata is contained in iw_type. For 14,914 (84% of the total number of wikidata entities in iw_unit) at least one assignment to an entity type is given.

In summary it can be stated that a large amount of basic data can be gained from wikidata whereby most of the attributes are far from complete (in the sense of existence for all wikidata entities). In case of existence, basic data from wikidata seems to be of good quality due to random sample checks.

5 Wikidata-based institutional disambiguation

Roughly described, the procedure is structured in the following way: Starting from rather clear allocations gained from simple allocation types (referred to as 'seed' here), more and more allocations are derived based on this seed. The underlying assumption of this approach is that every institution is mentioned in the WoS addresses at least on one occasion with a name variant appearing exactly this way (with respect to the transformation step) in wikidata.

The following attributes were used for addresses from the WoS:

1. **full address string** (referred to as *fulladdress* in the following)
2. **country** (assigned to an iso_3 code – referred to as *iso_3/country*)
3. **city** (given separately in the WoS – referred to as *city*)
4. **postalcode** (given separately in the WoS – referred to as *postalcode*)
5. **email addresses** from authors (author addresses were linked to mailaddresses via authors enabling the assignment of maildomains to author addresses – referred to as *maildomain*)

Obviously, maildomains cannot be used standing alone: for a large amount of authors no email address is given, sometimes authors use private email addresses (maildomains like 'web.de' or 'gmail.com') or addresses belonging to research institutions they no longer work for (which therefore do not match the institution mentioned in the address). But – in case of their existence and used in combination with other attributes or in a more statistical way – maildomains provide valuable information (e.g., email addresses of authors employed at Bielefeld University very often have the maildomain 'uni-bielefeld.de' – indicating the corresponding research institution).

From wikidata, several properties are used. Examples are:

1. **all kinds of labels**
2. **P17** (country)
3. several properties containing further **geographical information**
4. **P856** (url)
5. **P31** (instance of)

5.1 Creating the seed

For the creation of the seed, three simple matchings were performed with all WoS addresses (as described in section 2) where only the first part of the address (separated by ',') – referred to as 'orga1' in the following – and the city were used. In all matchings the transformation from the disambiguation procedure for German addresses was used as preparation step for addresses as well as wikidata labels.

1. **orga1-city: the concatenation of orga1 and city matches a wikidata label.**
   Example: Orga1: 'ALBERT LUDWIGS UNIV', City: 'FREIBURG', wikidata label: 'ALBERT LUDWIGS UNIV FREIBURG' (all after transformation).
   Further matches were derived in the following way: orga1 values of remaining addresses that are concatenations of orga1 and city (in any order) of an address with orga1-city-match were also matched.
   Example: Orga1: 'ALBERT LUDWIGS UNIV FREIBURG' and 'FREIBURG ALBERT LUDWIGS UNIV'.
   Taking into account orga1 and city information this is expected to be the most reliable match type.
2. **combination of name variants: the concatenation of two wikidata labels belonging to the same wikidata entity matches orga1.**
   Example: Orga1: 'RWTH TECH UNIV AACHEN' where both 'RWTH' as well as 'TECH UNIV AACHEN' are labels for the wikidata entity Q273263.
3. **orga1: orga1 matches a wikidata label.**
   Example: Orga1: 'CHARITE'.

Results were recorded with the priority given above: first orga1-city matches were recorded, for the remaining addresses combination matches were recorded, again for the remaining addresses orga1 matches were recorded.

5.1.1 Reduction of multiple assignments.

Although in most cases the orga1 values mention just one research institution, these matchings obviously do not necessarily lead to unique results due to several reasons. On the one hand there are structural circumstances in wikidata leading to undesired multiple assignments. This is the case, e.g., with redirections[9] or wikidata duplicates. Redirections were removed using the python wikidata client library, a duplicate check was performed only later (after other steps for reducing multiple assignments in order to minimize manual effort).

A further reason for multiple (undesired) assignments is an orga1 value not mentioning an institution. There are several addresses containing street information in orga1 (with mostly no institution given in other parts of the string). Addresses of this kind were excluded from these matchings. The identification of such cases is simple in case of equal values for street and orga1 attribute. In addition, some simple statistics were created: for all orga1 values appearing as street value at least once, frequencies as street value and frequencies as orga1 value were recorded. This information was used in some steps of the procedure.

Sometimes name variants are recorded for main institutions as well as affiliated subdivisions or predecessors and successors. To reduce multiple assignments in these cases, assignments to subdivisions were removed in case of the existence of an allocation to the belonging parent institution for the same fulladdress, analogous allocations of predecessors were removed in case of another allocation to a belonging successor.

Table 11 displays assignment rates at this point of the procedure. As already mentioned in section 3, assignment rates differ among the example countries where again CHE and ESP fall far behind the others. The number/percentage of matched and uniquely matched addresses vary widely – many multiple assignments remained.

Further reasons for multiple assignments are of course ambiguous labels for

[9] Redirections are wikidata ids without an own page but just redirecting to another wikidata entity, e.g. searching for Q35326428 leads to a redirection to the entity Q55044, with comment 'Redirected from Q35326428'.

country	**# addresses**	**matched**	**matched** (in %)	**uniquely matched** (in %)
CHE	123,141	60,091	48.8	38.26
DEU	606,775	403,866	66.56	57.50
ESP	477,325	244,991	51.33	42.86
GBR	496,464	350,082	70.52	56.31
KOR	318,207	259,017	81.40	75.74
ZAF	56,285	43,980	78.14	71.96

Table 11: Assignment rates (seed), distinct addresses, before feature step (description in the following).

wikidata entities. Examples of wikidata entities with the label 'Charité' are Q1077064 ('Charity: voluntary giving help to those who need it', 'charité' in French), Q3658725 ('a sculpture by Tino di Camaino'), Q184353 ('Greek goddesses of charm, beauty, nature, human creativity, and fertility') , Q29374135 ('statue') and – luckily – also a hospital in Berlin (Q162684).

In order to distinguish further wikidata entities with the same label from the desired allocation, attributes of addresses and wikidata entities were compared in case of their existence. Four different features were applied here:

1. entity type (instance of check)
2. country
3. city
4. maildomain/url

The four features and their application are described in more detail in the following sections.

Entity type. For the entity type feature, the value for property P31 (instance of) was recorded for all wikidata entities matched so far. This list was checked manually concerning types that may be research institutions (e.g. type university, hospital, company), types that are purely geographical entities (e.g. type city, mountain) and types that are definitely no research institutions (e.g. type band, sculpture, article). The list has not been edited completely but with priority on types that can be easily identified by helpful key words and types with high frequency. The list contains 5,044 'instance of' values, where 338 could be identified as institution types, 169 as pure geographical types and 281 as non-institution (and non-geographical) types.

Values: 0 (no decision), 1 (has institution type), allocations with other types were deleted directly.

Country. In order to perform a country matching between the country attribute of addresses and the country attribute of wikidata entities (property P17) both are assigned to iso3 country codes. For addresses, the country name disambiguation of the Competence Center was used, for wikidata entities the matching can be performed quite easily by using the property P298 which is existent for almost all countries.

Values: -1 (wikidata country differs from address country), 0 (no country given in wikidata), 1 (accordance in country values).

City. For a comparison of city values of wikidata entities and addresses, different name variants have to be taken into account. Of course a complete disambiguation of city names would be an oversized effort at this point, but existing name variants in wikidata can be used. Therefore, city attribute values of all addresses were matched against all wikidata labels (again after transformation), restricted to matches where also the address country value matches the country value of the wikidata entity allocated to the city (matching via iso3 codes as described above). Table 12 shows the number of distinct city values per country together with the number and share of matched values.
Thus, this matching provides an allocation of city values from addresses to city entities in wikidata. Subsequently the city entities were matched to the institution entities allocated to the address whenever the city entity appears as a value for an attribute of the institution entity (regardless of the property as we assume the entities to be cities and there are several geographical properties worth considering). As this may be confusing, figure 3 illustrates this step using an example.

country	# dist. city	thereof matched	in %
CHE	1,827	1,315	71.98
DEU	6,556	4,166	63.54
ESP	3,802	1,812	47.66
GBR	5,780	4,003	69.26
KOR	3,737	717	19.19
ZAF	1,236	690	55.83

Table 12: City matching.

The address in the example was assigned to the wikidata entity Q54096 (which is the University of Cologne) by the orga1-city matching described above. The city of the address ('COLOGNE') can be assigned to the wikidata entity Q365 by the city matching described in this section. As there is a connection between Q54096 and Q365 via the property P131 ('located in the administrative territorial entity'), the city value of the wikidata entity matches the city value of the address.

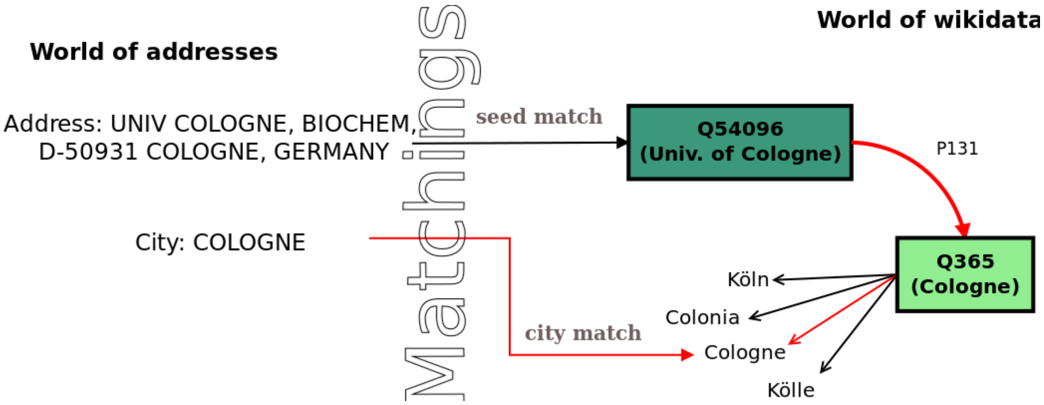


Figure 3: City feature.

Values: 0 (no city match), 1 (city match). In difference to handling the country feature, no negative values are assigned. For country values this can be done because country values are standardized using iso3 codes. In case of cities there is a danger of pruning unjustly just because of different name variants.

Maildomain/url. Email addresses of authors provide useful hints to corresponding institutions as maildomains often comply with or contain the url of the institution, e.g. many authors affiliated to Bielefeld University use email addresses with maildomain uni-bielefeld.de or a maildomain including uni-bielefeld.de (such as e.g. math.uni-bielefeld.de) – according to the url http://www.uni-bielefeld.de (figure 4).

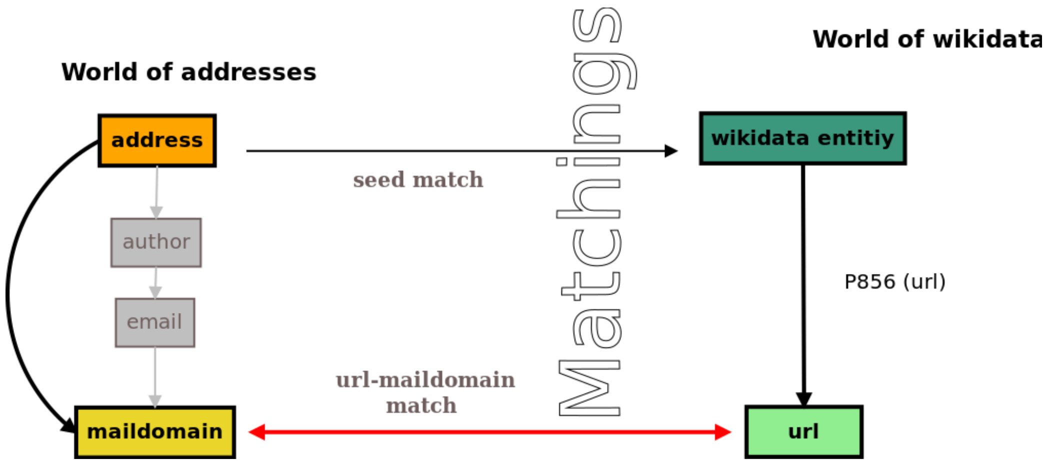


Figure 4: Maildomain/url feature.

Of course, email addresses are missing in many cases and sometimes private email addresses are used or email addresses are misleading in case, for example, authors are affiliated to more than one institution. Nevertheless this feature can provide useful hints in addition to the other features.

First, a connection between addresses and maildomains in Web of Science data is needed – this is not given directly, but can be gained by linking addresses to email addresses via authors, keeping only the maildomain (substring of the emailaddress after '@'). In wikidata, urls for entities can be extracted via property P856.

For every fulladdress there may be no, exactly one or more than one different maildomains with different frequencies (counting item-author-combinations) and every maildomain can be connected to one or more fulladdress values. To illustrate the first issue, table 13 contains some examples of maildomains for the fulladdress 'UNIV BIELEFELD, D-33615 BIELEFELD, GERMANY', table 14 gives some examples of fulladdress values for the maildomain 'kit.edu'. The examples in table 13 also demonstrate that in several cases the consideration of maildomains leads to conclusions concerning subdivisions of institutions. Another advantage of dealing with maildomains is the matter of fact that hints to institutions are given regardless of string similarities to institution name variants in the address (last example in table 14). This linking between addresses and maildomains was also used in further steps of the procedure.

At this point, maildomains are not yet used to group fulladdress values of one institution respectively to find more fulladdresses for an institution but just to check a given allocation (address → wikidata entity) via comparing maildomain and url (prefixes 'http://www.' removed).
For each fulladdress the most frequent maildomain was determined (in case of the existence of more than one maildomain per fulladdress), so different match levels could be distinguished: at least one maildomain contains the url, at least one maildomain equals the url, the most frequent maildomain (per address) contains the url, the most frequent maildomain equals the url.

Values: 0 (no match), 0.5 (at least one maildomain contains the url), 1 (at least one maildomain equals the url), 1.5 (the most frequent maildomain contains the url),2 (the most frequent maildomain equals the url).

count	**maildomain**
80	uni-bielefeld.de
19	physik.uni-bielefeld.de
13	techfak.uni-bielefeld.de
5	math.uni-bielefeld.de
5	cit-ec.uni-bielefeld.de
4	wiwi.uni-bielefeld.de

Table 13: Maildomain examples for fulladdress value 'UNIV BIELEFELD, D-33615 BIELEFELD, GERMANY'.

count	fulladdress
660	KARLSRUHE INST TECHNOL, D-76021 KARLSRUHE, GERMANY
57	KIT, KARLSRUHE, GERMANY
4	KARLSRUHE INSITUTE TECHNOL, INST CONTROL SYST, D-76131 KARLSRUHE, GERMANY
3	ENGLER BUNTE RING 1, D-76131 KARLSRUHE, GERMANY

Table 14: Fulladdress values with maildomain 'kit.edu'.

Feature sum. Proceeding this way, every allocation a of the form

$$\text{(iso\_3, city, fulladdress)} \rightarrow \text{wikidata entity}$$

in the seed set receives four scores – one for each feature – where

- $score_{type}(a) \in \{0, 1\}$
- $score_{country}(a) \in \{-1, 0, 1\}$
- $score_{city}(a) \in \{0, 1\}$
- $score_{domain}(a) \in \{0, 0.5, 1, 1.5, 2\}$.

For every allocation a the sum of all scores was calculated (referred to as feature sum of a in the following):

$$feature\_sum(a) = score_{type}(a) + score_{country}(a) + score_{city}(a) + score_{domain}(a),$$

so for every (iso_3, city, fulladdress) the allocation(s) with maximal feature sum could be determined. Allocations with lower feature sums for the same address were removed. As a last step, all (iso_3, city, fulladdress) with allocations to more than two different wikidata entities were removed from the seed.

For the identification of wikidata duplicates, a list of all fulladdresses with exactly two wikidata entities assigned was created from the seed set. From this, a list of wikidata pairs (as candidates for duplicates) was generated (wikidata entities connected by the allocation to one common fulladdress) with their frequency of occurrence (in the sense of the number of fulladdresses allocated to both wikidata entities of the pair). From this list, the

top 50 entries (concerning frequency) were checked manually for duplicate detection (a preferred wikidata entity was chosen in case of duplicates). Of the 50 wikidata entities checked, 13 could be identified as duplicates.
For all duplicates identified in this step, allocations of the seed were switched to the preferred wikidata entity.

Table 15 presents the number of fulladdresses handled per match type for the allocations remaining in the seed set, figures 5 and 6 show the frequencies of feature sum values as well as separate frequencies for the single features in terms of allocations.

match type	**# dist. fulladdresses**
orga1 city	32,003
orga1 city (further additions)	273,811
combination	16,168
orga1	1,030,013

Table 15: Number of fulladdresses handled per match type.

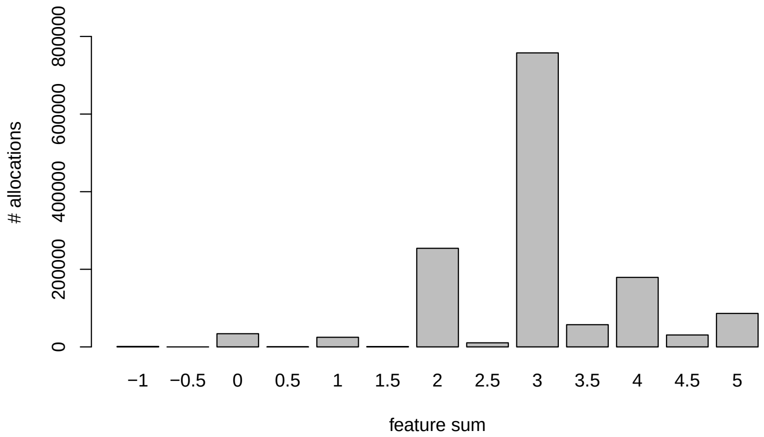


Figure 5: Frequencies of feature sum values in the seed set.

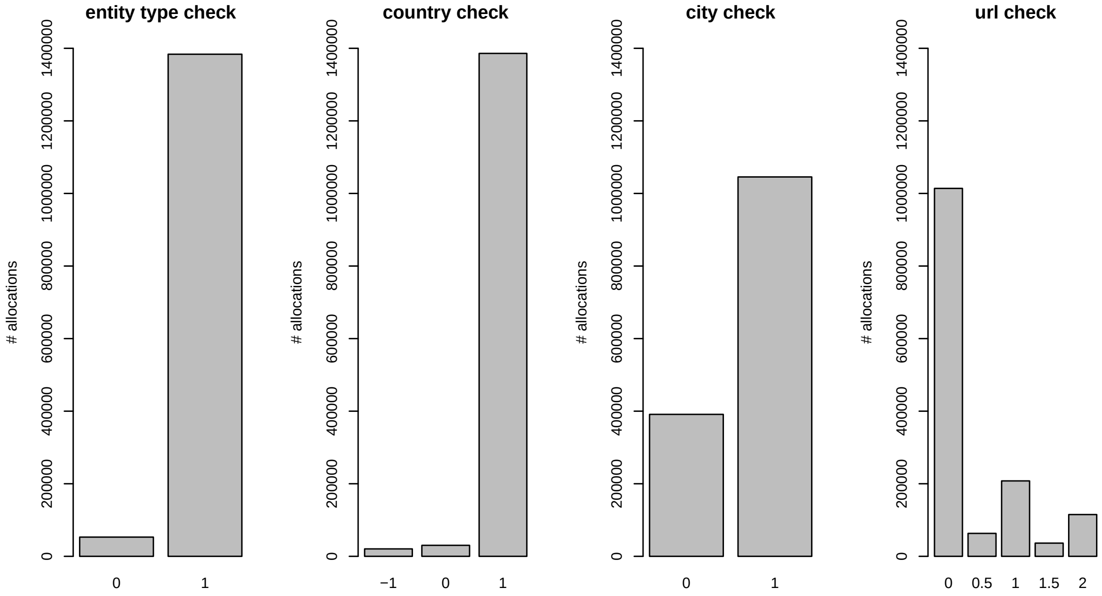


Figure 6: Frequency of feature values in the seed set.

Results. The seed serves as a reference point for all further allocations. Tables 16 and 17 display the assignment rates achieved by the seed set – once counting distinct addresses, once taking frequencies into account (which leads to counting item-address-combinations).

country	**# addresses**	**matched**	**matched** (in %)	**uniquely matched** (in %)
CHE	123,141	59,475	48.3	47.17
DEU	606,775	400,400	65.99	65.45
ESP	477,325	244,177	51.16	50.81
GBR	496,464	347,905	70.08	69.75
KOR	318,207	256,251	80.53	79.86
ZAF	56,285	43,787	77.8	76.69

Table 16: Assignment rates (seed), distinct addresses.

country	**# addresses**	**matched**	**matched** (in %)	**uniquely matched** (in %)
CHE	368,697	214,091	58.07	57.43
DEU	1,688,849	1,286,649	76.18	75.82
ESP	955,455	538,306	56.34	56.06
GBR	1,705,950	1,384,765	81.17	81.02
KOR	880,626	778,873	88.45	87.92
ZAF	150,189	128,461	85.53	84.83

Table 17: Assignment rates (seed), addresses with frequency.

Compared to the assignment rates before the feature step (table 11) there is a slight loss of matched addresses combined with a significant increase of uniquely assigned addresses – reduction of multiple assignment succeeded without paying a high price.

Compared to the assignment rates in the gold standard set, assignment rates for the overall address set are lower – although the matching for the gold standard suggestion lists was an even simpler one. This may be due to the long tail of addresses with low frequencies – including addresses using seldom or incorrect spellings/name variants (more likely not covered in wikidata than more frequent ones) and addresses mentioning institutions with low frequency (e.g. small institutions or institutions that are no 'typical' research institutions with a low number of publications, for example companies or hospitals). This assumption is confirmed by the consideration of addresses with frequency providing higher assignment rates (table 17).

Again, differences among countries can be stated with the tendencies already observed. The wikidata entities received in this step form the content of the basic data tables.

5.2 Further allocations based on the seed

Using the seed as basis, further allocations were created using statistical considerations, string similarities/permutations and domain strings of email addresses. Three laps for the generation of new allocations were performed – with derivations of further allocations via the newly gained orga1 values in between. This section offers a brief description.

5.2.1 Lap 1:

Fulladdresses not yet allocated (that means not in the seed set) to one or more wikidata entities are matched to fulladdresses in the seed (and therefore the belonging wikidata entities) with the following methods:

1. **Country and city match, fulladdress match after deletion of whitespaces and commas.**
   Example:

   fulladdress seed:
   RHEIN WESTFAL TH AACHEN, UNIV HOSP, DEPT HEMATOL & ONCOL, AACHEN, GERMANY,
   fulladdress:
   RHEIN WESTFAL TH AACHEN UNIV HOSP, DEPT HEMATOL & ONCOL, AACHEN, GERMANY'

2. **Country and city match, orga1 match after replacement of terms defining university hospitals by 'UNIV'** (e.g. UNIV HOSP, UNIV MED CTR).
   Example:

   fulladdress seed:
   UNIV MAINZ,[...], MAINZ, GERMANY
   fulladdress:
   UNIV HOSP MAINZ,[...], MAINZ, GERMANY

3. **Country and city match, orga1 is a word permutation of the seed fulladdress.**
   Example:

fulladdress seed:
UNIV HOSP ZURICH, [...] ZURICH, SWITZERLAND
ZURICH UNIV HOSP, [...] ZURICH, SWITZERLAND
fulladdress:
UNIV ZURICH HOSP, [...] ZURICH, SWITZERLAND

4. **(Country, city, street) is unambiguously allocated in the seed set.**

5. **Country and city match, orga1 place change** (the orga1 value of the seed fulladdress appears somewhere 'in the middle' of the address, not as first part of the address). This method provides not only allocations for addresses not yet assigned to any wikidata entity but furthermore the identification of addresses mentioning more than one institution for fulladdresses where the orga1 values were already allocated in the seed set.
   Example:

   fulladdress seed:
   HERTIE INST CLIN BRAIN RES, LAB FUNCT NEUROGENET, TUBINGEN, GERMANY
   (assigned to Q30287260='Hertie Institute for Clinical Brain Research')
   fulladdress:
   UNIV TUBINGEN, HERTIE INST CLIN BRAIN RES, MED CTR, D-72076 TUBINGEN, GERMANY
   (was already assigned to the University of Tübingen in the seed but receives a new allocation to Q30287260 in this step)

6. **Country and city match, jaro winkler similarity[10] (range: 0-100) of fulladdresses $\geq$ 90** (separated into two levels: $<$ 95 and $\geq$ 95).
   Example:

   fulladdress seed:
   ALFRIED KRUPP WISSENSCHAFTSKOLLEG, GREIFSWALD, GERMANY

[10] https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

fulladdress:
ALFRIED KRUPP WISSENSCH KOLLEG GREIFSWALD,
GREIFSWALD, GERMANY

7. **Maildomain match**. For this method, the linking of addresses to maildomains is used again – this time for grouping fulladdresses. Maildomain-address-combinations appearing only once were excluded. For every maildomain, the most often assigned wikidata entity was determined from the seed set (taking into account frequencies of maildomain-address-combinations). These maildomains were matched with the most often assigned maildomain per (not yet allocated) address (two levels: exact match of maildomains, seed maildomain is a substring of the maildomain of the address not yet allocated).

After processing these methods, allocations to child and predecessor entities were removed as already described in 5.1.1. The new allocations were used to derive more fulladdresses/futher allocations in the following way:

1. **New orga1 values:** allocation methods based on orga1 values and considered as most likely safe (which are methods 1, 2 and 3) provide new orga1 values (respectively (country, city, orga1) values) for wikidata entities. A matching on country, city and these new orga1 values provides new allocations.

2. **No-street-address:** for all addresses, a version with the street value removed from the fulladdress was recorded. A matching on these 'no-street-addresses' gives new allocations for fulladdresses not yet allocated but with a no-street-address-value that also appears as a no-street-address-value for allocated addresses.

3. **Uniquely allocated orga1-city-combinations** lead to new allocations for addresses with the same orga1-city-combination.

4. **Uniquely allocated orga1 values:** analogous.

5. **New transformed orga1-city-combinations:** analogous to the no-street-address-approach, for each fulladdress a version with just a concatenation of transformed orga1 and city values was used to receive

new allocations (again only for methods 1, 2 and 3 of lap 1).

At the end of lap 1, new allocations had been gained with statistical methods, string similarities and permutations as well as maildomain matches followed by derivations of further allocations based on the new ones.

5.2.2 Lap 2:

As new allocations were received in lap 1, methods for gaining new allocations based on already existing ones can be applied one more time. Lap 2 was restricted to methods 1, 2 and 3, which are not applied to the seed set as in lap 1 but also on the new allocations received in lap 1.

Thereafter, the derivation of further allocations was performed exactly as after lap 1.

5.2.3 Lap 3:

Lap 3 aims at the collection of based on string similarity measures (jaro winkler similarity[11] and edit distance[12]) performed on the fulladdress string (in order to use not only information from orga1 values but including all information given for measuring similarity/distance – such as also street names, postal codes and subdivision information). Three methods were applied where the first and the second are hedged by a match of country, city and street values in addition while method 3 requires only country and city match (leading to much more allocations at the expense of causing more errors):

1. **Country, city and street match, edit distance of fulladdresses $\leq$ 2.**
   Example:

   fulladdress with allocation:
   UNIV DUSSELDORF, MATH INST, UNIV STR 1, D-40225
   DUSSELDORF, GERMANY
   fulladdress:
   UNIV DSSELDORF, MATH INST, UNIV STR 1, D-40225
   DSSELDORF, GERMANY

2. **Country, city and street match, jaro winkler similarity of fulladdress values $\geq$ 85 and matched with maximal similarity (may be more than one) provide an unambiguous wikidata entity.**
   Example:

   fulladdress:
   BAM FACHBEREICH 9 1, UNTER EICHEN 87, D-12205 BERLIN,
   GERMANY

   fulladdresses with allocations
   (similarity to fulladdress in brackets):

[11] https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance
[12] https://en.wikipedia.org/wiki/Edit_distance

BAM, FACHBEREICH 8 4, UNTER EICHEN 87, D-12205 BERLIN, GERMANY (90)
BAM BERLIN, FACHBEREICH INGN BAU 7 2, UNTER EICHEN 87, D-12205 BERLIN, GERMANY (85)
BAM FED INST MAT RES TESTING, UNTER EICHEN 87, D-12205 BERLIN, GERMANY (85)

Country and city values of the fulladdresses with allocations match the corresponding values for the fulladdress, the maximal similarity is 90 where there is only one address with this similarity – the match provides an unambiguous wikidata entity and the fulladdress can be assigned to the same wikidata entity.

3. **Country and city match jaro winkler similarity of fulladdresses $\geq$ 85 and the top 5 (concerning similarity values) matches provide an unambiguous wikidata entity.**
Example:

fulladdress:
RW1H AACHEN UNIV, INST STEEL STRUCT, D-52074 AACHEN, GERMANY
top 5 fulladdresses with allocations
(similarity to fulladdress in brackets):
RWTH AACHEN UNIV IKV, INST PLAST PROC, D-52064 AACHEN, GERMANY (85)
RWTH UNIV HOSP, DEPT INTENS CARE, D-52074 AACHEN, GERMANY (85)
RWTH AACHEN UNIV, MED CLIN 1, PAUWELSSTR 30, D-52074 AACHEN, GERMANY (85)
RWTH AACHEN UNIV HOSP, PAUWELSSTR 30, D-52074 AACHEN, GERMANY (85)
RWTH UNIV HOSP AACHEN, DEPT NEUROL, D-52074 AACHEN, GERMANY (85)

All top 5 fulladdresses with allocations (in this case all with the at least required similarity of 85 to the fulladdress of interest) have allocations to one single wikidata entity – so the matches provide an unambiguous entity – therefore the example fulladdress can be assigned to this wikidata entity. As the fulladdress in this example does not contain a street value the two previous methods failed in this case.

In a last step all fulladdresses were removed from the results where more than five wikidata entities were assigned to one single fulladdress.
Table 18 shows all methods applied together with the number of allocations gained.

Of course further extensions and potential improvements of this procedure are conceivable. In particular individual evaluations for all methods applied and an in-depth analysis of addresses not yet allocated by the procedure would be sensible steps in order to optimize priorities and add missing match types. For the purpose of this feasibility study this has been restricted to the methods applied as an initial test of a disambiguation procedure based on wikidata due to limited resources of time and manpower.

The methods were chosen with respect to practical implementation issues: although the jaro winkler similarity is a good measure in this context and would also be of interest without demanding city equality, this was not incorporated due to processing time observations, taking into account the aim of developing a scalable procedure.

5.3 Evaluation

5.3.1 Assignment rates

Tables 19 and 20 show allocation rates after processing all disambiguation steps. Multiple assignments occur – in difference to the first allocations (seed set) dealing only with orga1 values instead of fulladdresses, they are not necessarily unwanted but may be desired at this point in case of more than one institution mentioned in an address (like in the example of the Hertie Institute in the description of the place change matching method in lap 1).

Furthermore it has to be taken into account that the procedure does not contain an aggregation on the main institutional level so far – therefore e.g. allocations to a university hospital and a university may both appear for one

lap	method	# allocations
0	seed	1,433,876
1	Country and city match, fulladdress match after deletion of whitespaces and commas	9,343
1	Country and city match, orga1 match after replacement of terms defining university hospitals by 'UNIV'	13,695
1	Country and city match, orga1 is word permutation	29,992
1	(Country, city, street) is unambiguously allocated	34,691
1	Country and city match, orga1 place change	179,233
1	City match, jw-similarity<95	95,967
1	City match, jw-similarity≥95	8,420
1	Maildomain match (url substring)	7,175
1	Maildomain match (url match)	33,295
add 1	New orga1 values	8,892
add 1	No-street-address	28,428
add 1	(Orga1, city) unique	55,730
add 1	Orga1 unique	14,878
add 1	New transformed orga1-city-combinations	2,121
2	Country and city match, fulladdress match after deletion of whitespaces and commas	1,522
2	Country and city match, orga1 match after replacement of terms defining university hospitals by 'UNIV'	2,246
2	Country and city match, orga1 is word permutation	155,007
add 2	New orga1 values	156,465
add 2	No-street-address	265
add 2	(Orga1, city) unique	64
add 2	Orga1 unique	533
add 2	New transformed orga1-city-combinations	1612
3	Country, city and street match, jw-similarity $\geq$ 85. max. jw-sim unique	4,062
3	Country, city and street match, edit distance $\leq$ 2	581
3	Country and city match, jw-similarity≥85, top 5 unique	237,715

Table 18: Assignment rates (all allocations), distinct addresses.

fulladdress, especially if the university hospital wikidata entity is not linked to the wikidata entity of the university. In this case none of the allocations is incorrect, it is just not necessary to keep both of them recorded.

country	**# addresses**	**matched**	**matched** (in %)	**uniquely matched** (in %)
CHE	123,141	104,944	85.22	47.17
DEU	606,775	535,038	88.18	65.45
ESP	477,325	410,760	86.05	50.81
GBR	496,464	446,410	89.92	69.75
KOR	318,207	294,387	92.51	79.86
ZAF	56,285	50,834	90.32	76.69

Table 19: Assignment rates (all allocations), distinct addresses.

country	**# addresses**	**matched**	**matched** (in %)	**uniquely matched** (in %)
CHE	368,697	342,138	92.80	85.08
DEU	1,688,849	1,582,087	93.68	88.26
ESP	955,455	862,570	90.28	80.37
GBR	1,705,950	1,631,877	95.66	82.87
KOR	880,626	847,795	96.27	87.83
ZAF	150,189	142,935	95.17	87.16

Table 20: Assignment rates (all allocations), addresses with frequency.

Compared to the seed allocations, allocation rates could be increased by adding further allocations by additional steps of the disambiguation procedure so that now for each country the share of matched addresses exceeds the share of matched addresses in the gold standard.

Keeping in mind the share of addresses with institutions recorded in wikidata stated in the gold standard set (table 5) which provides a hint for the upper bound of assignment rates, the allocation rates achieved seem to be already acceptable at this point.

5.3.2 Check against gold standard

The creation of the gold standard enables not only considering assignment rates but also receiving information on the quality of results.

	CHE	DEU	ESP	GBR	KOR	ZAF
(1) # addresses in gold standard	500	1,471	1,158	1,195	767	500
(2) thereof assigned to at least one wikidata entity	477	1,399	1,032	1,156	760	472
(3) thereof contained in results of disambiguation procedure	471	1,388	1,001	1,141	756	469
(3) in % of (2)	98.74	99.21	97.00	98.70	99.47	99.36
(4) thereof with at least one assignment also appearing in gold standard	442	1,271	870	1,064	733	456
(4) in % of (2)	92.66	90.85	84.30	92.04	96.44	96.61

Table 21: Check against gold standard: numbers of addresses and allocations contained in the assignment results of the disambiguation procedure.

Table 21 shows that for each country almost all addresses assigned to wikidata entities manually for the gold standard have also been handled by the disambiguation procedure.

Concerning the entities assigned, in most cases at least one assignment (there may be more than one wikidata entity assigned) is also contained in the gold standard set.

Table 22 provides a closer look at the comparison of gold standard assignments and disambiguation results at the basis of allocations (not: addresses, which means that if an address is assigned to, e.g., two different wikidata entities this is counted as two allocations).

The set of allocations in the gold standard is referred to as G (without addresses assigned to only a url due to the lack of a corresponding wikidata entity) in the following while the set of allocations in the disambiguation results will be D.

The following figures are given, both in absolute numbers as well as shares of the number of allocations in the gold standard:

1. $G \cap D$**:**
   allocations appearing in both the gold standard and in the disambiguation results.
2. $G \setminus D$**:**
   allocations in the gold standard missing in the disambiguation results.
3. $D \setminus G$**:**
   allocations in the disambiguation results not given in the gold standard.
4. The number of addresses without any assignment to wikidata entities in the gold standard assigned to at least one wikidata entity in the disambiguation results.

In the gold standard as well as the disambiguation results multiple assignments are possible wherefore shares do not sum up to 100%.

	CHE	**DEU**	**ESP**	**GBR**	**KOR**	**ZAF**
(1) #allocations in gold standard	478	1,407	1,086	1,156	760	478
$G \cap D$ – in % of (1)	442 92.47	1,271 90.33	894 82.32	1,064 92.04	733 96.45	456 95.40
$G \setminus D$ – in % of (1)	36 7.53	136 9.67	192 17.68	92 7.96	27 3.55	22 4.60
$D \setminus G$ – in % of (1)	96 20.08	198 14.07	231 21.27	272 23.53	98 12.89	44 9.21
disambiguation results of addresses not in wikidata due to gold standard	15	64	139	26	5	30

Table 22: Check against gold standard: comparison on the basis of allocations.

It can be stated that the largest share by far of the disambiguation results can be found in the gold standard as well as for all countries – where the highest share applies for KOR with 96% followed by ZAF with 95% while the share for ESP is with 82% much lower.

For all countries, there are allocations missing in the disambiguation results that can be found in the gold standard as well as additional allocations in the disambiguation results that are not contained in the gold standard and therefore seem to be undesired allocations. Allocations of the last type seem to be most problematic as they appear to be allocations to wikidata entities in case of addresses with corresponding institutions not covered in wikidata.

To gain deeper insights into the last two groups of allocations (which seem to display errors), all allocations in the disambiguation results were sighted in detail for allocations for DEU addresses.

For the last group (addresses of institutions not covered in wikidata due to gold standard) it turned out that 32 of the 64 allocations are correct (while 4 stay unclear and only 28 are indeed incorrect). Thus, institutions mentioned in these addresses could not be found while creating the gold standard but are now definitely covered in wikidata. In some cases only a subdivision is mentioned in the address (where the subdivision is not recorded in wikidata) while the allocation directs to the parent institution. Furthermore, wikidata is continuously changed, updated and supplemented by new entities – so entities may be newly recorded after creation of the gold standard or new name variants are added which enable the previously unsuccessful identification of the corresponding wikidata entity.

The group $D \setminus G$ contains 141 correct assignments as well as in addition 6 allocations to sectors (allocations to, e.g., Max Planck Society instead of the single Max Planck Institute) and 8 allocations to libraries of the corresponding institution (it turned out that sometimes the label for the institution is also used as label for the library of the institution leading to allocations to libraries) – these are not exactly the desired results but they are no errors. Only 43 allocations are indeed incorrect.
Among the correct assignments there are more detailed assignments as the ones recorded in the gold standard (e.g., a faculty or institute of a university is in the disambiguation results while in the gold standard the university is assigned – here both assignments are correct, they just address different hierarchical levels). Similarly in case of university hospitals, in one set the university may be addressed while in the other the university hospital is used as allocation target. Also wikidata duplicates show up in this group.

Thus, both in the last and the second last group, by far not all allocations are errors – in case of $D \setminus G$ only 22% are indeed incorrect.

A closer look at the set $G \setminus D$ will show if these allocations are really 'missing' ones: for DEU, 11 allocations in $G \setminus D$ belong to addresses that are not at all handled in the disambiguation results. Here allocations are definitely missing.
In 8 cases, for the same address there are also allocations in $G \cap D$. All 8 addresses belong to one single institution (namely Ernst-Moritz-Arndt-Universität Greifswald) which is recorded twice in wikidata (wikidata duplicates, Q20426978 and Q165528) where both entities are assigned in the gold standard, only one (after removing duplicates identified) in the disambiguation results.
For 98 allocations, other allocations given in the disambiguation procedure have already been flagged as correct while inspection of $D \setminus G$ – these may just be alternative allocations (e.g., duplicates, allocations on other hierarchical levels).
For 23 allocations, other allocations given in the disambiguation procedure had been already flagged as incorrect (in $D \setminus G$), here allocations are also really missing. Overall, here again it can be stated that the first impression of really 'missing' allocations only holds for a small subset.

	# allocations DEU
(1) total # of allocations in D	1,533
# of allocations in $D \cap G$	1,271
# of allocations in $D \setminus G$, identified as correct	155
# of allocations for addresses of institutions not in wikidata due to gold standard, identified as correct	32
sum of allocations identified as correct	1,458
– in % of (1)	95.11

Table 23: Disambiguation results for DEU.

In summary, a gold standard created in the way as done here, is not sufficient for a complete quality evaluation of disambiguation procedures as there may be additional or alternative allocations to other wikidata entities than the ones looked up manually (due to several reasons). The intersection of disambiguation results and gold standard only provides a lower bound of the number of correct allocations. With the closer look at the different sets of allocations, the number of correct allocations can be updated for German addresses in the way displayed in table 23.

With this update, 95.11% of the allocations in the disambiguation results for DEU are correct (while the share of $G \cap D$ among the number of allocations in the gold standard is significantly lower with 90.33%).

6 Conclusion and further steps

6.1 Suitability of wikidata for the extraction of basic data and disambiguation procedure

In summary, wikidata appears to be valuable for the task of the disambiguation of author addresses. It turned out that a large amount of institutions is covered in wikidata which therefore is able to serve as source.
In addition, wikidata entities are equipped with attributes of interest (name variants, geographical information, urls, identifiers of different kinds enabling linkings to other sources). Although this information is not complete, it may be very useful in many contexts – and, of course, wikidata is changing and growing, so information may become more and more complete.
The results of the disambiguation procedure created here – containing only a minimum of manual effort at two points (identification of instance of values, duplicate check) – lead to the assumption that wikidata-based disambiguation is a promising approach to solve or at least contribute to the solution of the task.

Advantages of wikidata as a source for basic information and disambiguation are obvious: wikidata is a free source, is continuously updated, results can be easily exchanged as identifiers (wikidata ids) are available for anyone and independent of country and language.

Of course, disambiguation results could be improved by adding manual steps, especially using text patterns for allocations or exclusions in addition – this has deliberately not been done as the aim was to explore the potential of wikidata for a scalable procedure rather than tuning results for single countries.

Aggregation up to main institutional level has not been addressed in the procedure yet – proposing this may be left to users providing as much flexibility as possible. The basic data tables enable them to perform aggregations individually aligned to their specific project context.

Although the procedure was developed for WoS addresses in this context, it is not restricted to WoS addresses but may be applied to any other set of author addresses as well (provided country, city and a fulladdress string are available, email addresses should be existent – but not necessarily for all addresses – to take advantage of the maildomain match).

6.2 Next steps

Due to limited resources, an in-depth analysis of allocation results compared to the goldstandard (as done for DEU as an example, table 23) could not be provided in this project. This would be of value to get further insights into the actual precision values.

Furthermore, a detailed error analysis would be of high interest – this would provide hints for improvements of the disambiguation procedure as well as insights in the reasons for differences (of coverage and disambiguation results) among countries.

For missing institutions or attributes in wikidata, a systematic supplement could be considered – the resulting improvement of the basic data contained in wikidata would be freely available. It would be desirable to work jointly on the improvement of an open access data basis instead of hosting closed-shop data behind the walls of several institutions engaged in the business of bibliometrics. It could be considered to do this in the context of the WikiCite initiative[13].

[13] https://meta.wikimedia.org/wiki/WikiCite