Disambiguating and Geocoding Organizations

Matching WOS to Wikidata and GRID

Aliakbar Akbaritabar (Ali)

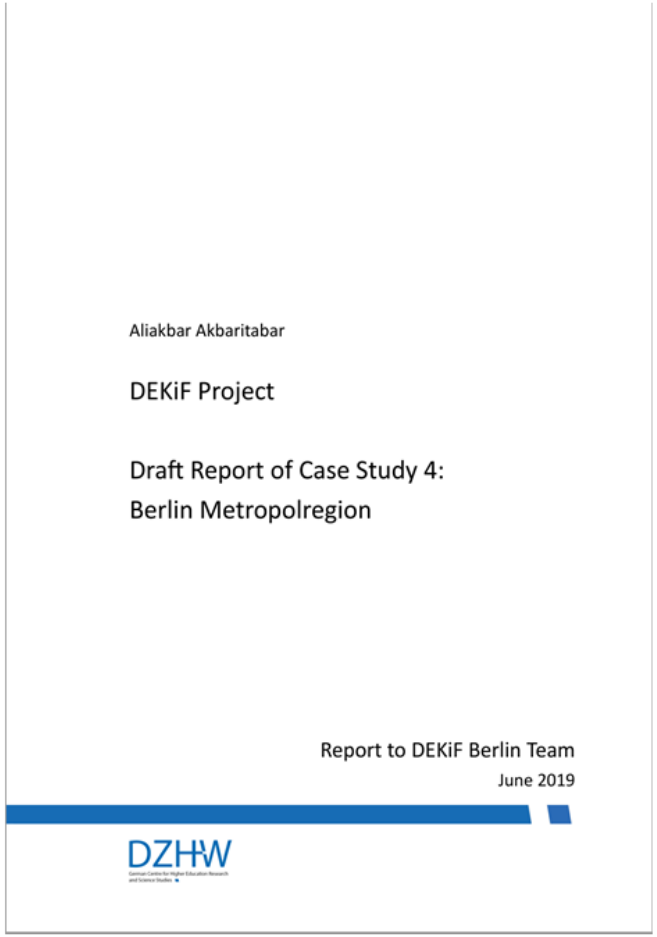
Akbaritabar@DZHW.eu

20 November, 2019

DZHW
German Centre for Higher Education Research and Science Studies

Two draft reports are prepared

- For DEKiF AP9-4 case study
- The report on matching has sample code to replicate

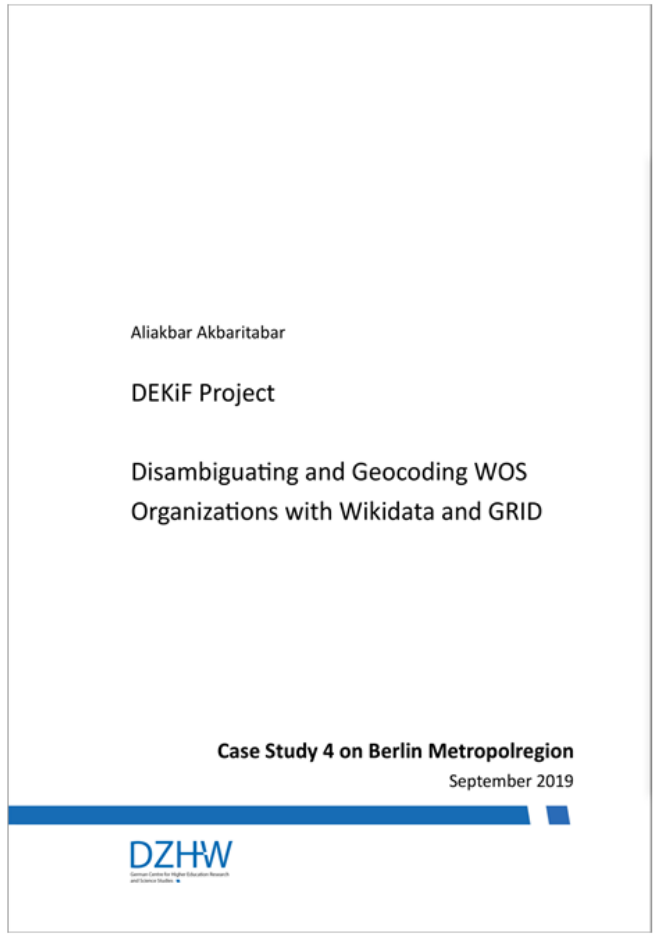

Aliakbar Akbaritabar

DEKiF Project

Draft Report of Case Study 4:
Berlin Metropolregion

Report to DEKiF Berlin Team
June 2019

DZHW


Aliakbar Akbaritabar

DEKiF Project

Disambiguating and Geocoding WOS
Organizations with Wikidata and GRID

Case Study 4 on Berlin Metropolregion
September 2019

DZHW

Brief on data and methods (Matching report)

DZHW

1. All WOS ~-**2017** (b_2018 KB) pubs
2. *Article*, *Review* as **document types**
3. With at least one author/institution from **Berlin**, **Germany**
4. Number of publications, fractional count, 3 years citations
5. **Wikidata** 27th March 2019, **GRID** 17th February 2019

Brief on data and methods (Matching report)

DZHW

1. All WOS ~-**2017** (b_2018 KB) pubs
2. *Article*, *Review* as **document types**
3. With at least one author/institution from **Berlin**, **Germany**
4. Number of publications, fractional count, 3 years citations
5. **Wikidata** 27th March 2019, **GRID** 17th February 2019

Brief on data and methods (Matching report)

DZHW

1. All WOS ~-**2017** (b_2018 KB) pubs
2. *Article*, *Review* as **document types**
3. With at least one author/institution from **Berlin**, **Germany**
4. Number of publications, fractional count, 3 years citations
5. **Wikidata** 27th March 2019, **GRID** 17th February 2019

Brief on data and methods (Matching report)

DZHW

1. All WOS ~-**2017** (b_2018 KB) pubs
2. *Article*, *Review* as **document types**
3. With at least one author/institution from **Berlin**, **Germany**
4. Number of publications, fractional count, 3 years citations
5. **Wikidata** 27th March 2019, **GRID** 17th February 2019

Brief on data and methods (Matching report)

DZHW

1. All WOS ~-**2017** (b_2018 KB) pubs
2. *Article*, *Review* as **document types**
3. With at least one author/institution from **Berlin**, **Germany**
4. Number of publications, fractional count, 3 years citations
5. **Wikidata** 27th March 2019, **GRID** 17th February 2019

In matching report, Wikidata was limited

DZHW

- To instances of:
  - '*Comprehensive university*' (Q1767829)
  - '*Public university*' (Q875538)
  - '*University*' (Q3918)
  - '*Academic institution*' (Q4671277)
  - '*Fraunhofer Institute*' (Q20168706)
  - '*Research institute*' (Q31855)
  - '*Scientific society*' (Q748019)
  - '*Scientific organisation*' (Q45103187)
  - '*Max Planck Society*' (Q158085)
  - '*Max Planck Institute*' (Q6019423).
- These limited our data from over 55 million cases to **106,794** entities.

Unique organizations (problematic?!)

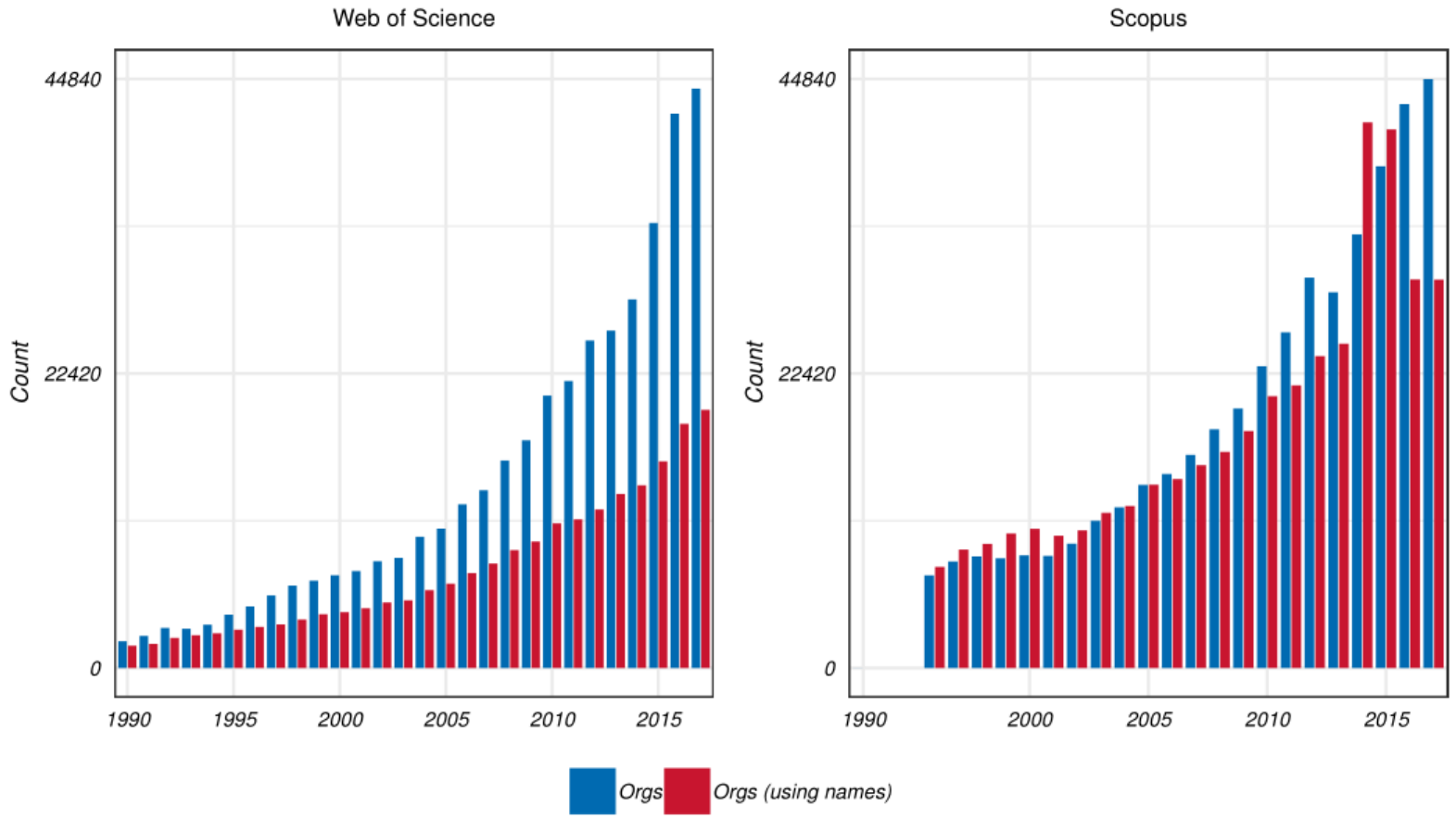
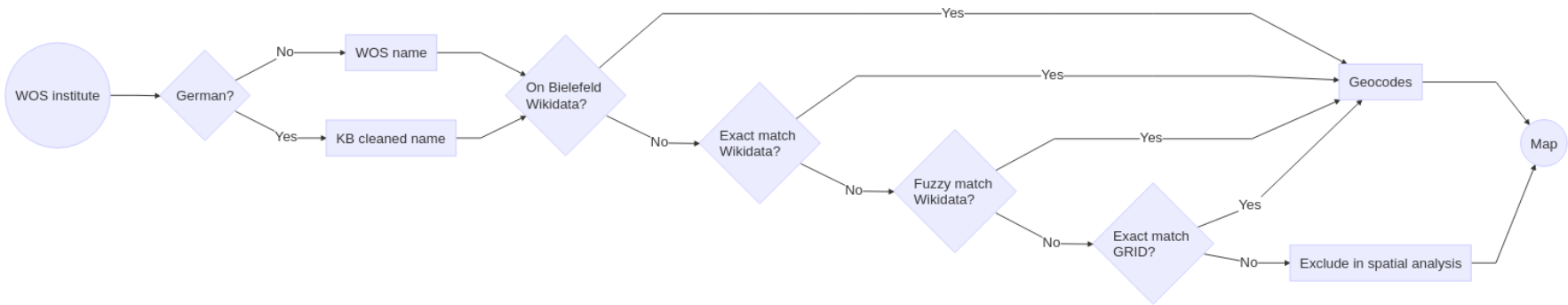


Figure 10: Unique organizations with which Berlin region institutes and universities have collaborated in Articles, Reviews and Conference proceedings in WOS and Scopus in 1990 - 2017

Disambiguation & geocoding logic



Exact and fuzzy matching with GRID and Wikidata

DZHW

Table 1: Descriptive metrics on Berlin metropolitan region organizations, countries and cities (WOS, only article and reviews and where organization was not previously disambiguated by Rimmert et al's project)

Metric	Value
# FK_INSTITUTIONS	287087
# with unique KB name	1028
# Without KB name	286060
# unique WOS country	201
# unique WOS city	11299
# unique German orgs	21726
# unique orgs located in Berlin	10554
# unique WOS ORGANIZATION1	72486
# Normalized name (baseline)	72609
# unique WOS orgs exact match with Wikidata (%)	42665 (58.76%)
# unique Wikidata organizations (exact match disambiguation result)	4444
# unique WOS orgs exact match with GRID (%)	45409 (62.54%)
# unique Wikidata organizations (GRID match disambiguation result)	4614
# unique WOS orgs FUZZY match with Wikidata (%)	60841 (83.79%)
# unique Wikidata organizations (FUZZY match disambiguation result)	7004

International organization example (1/2)

	FK_INSTITUTIONS	PK_KB_INST	KB_NAME	ORGANIZATION1
	All	All	All	aalborg univ
2	29009365	NA	NA	aalborg univ
3	1077601	NA	NA	aalborg univ
4	30422015	NA	NA	aalborg univ
6	8497964	NA	NA	aalborg univ
7	3823950	NA	NA	aalborg univ
8	5559997	NA	NA	aalborg univ
9	20140789	NA	NA	aalborg univ
10	2583771	NA	NA	aalborg univ
11	21284558	NA	NA	aalborg univ
12	27026789	NA	NA	aalborg univ
14	34366196	NA	NA	aalborg univ
15	23317559	NA	NA	aalborg univ
16	24242911	NA	NA	aalborg univ
17	28629165	NA	NA	aalborg univ
18	4170458	NA	NA	aalborg univ
19	6898547	NA	NA	aalborg univ
20	33059701	NA	NA	aalborg univ
22	24567558	NA	NA	aalborg univ
23	18351477	NA	NA	aalborg univ
24	17626795	NA	NA	aalborg univ
25	22866279	NA	NA	aalborg univ
26	7812723	NA	NA	aalborg univ
27	18269492	NA	NA	aalborg univ
28	14067026	NA	NA	aalborg univ
30	27769743	NA	NA	aalborg univ
31	18178514	NA	NA	aalborg univ
33	4883551	NA	NA	aalborg univ
34	16148970	NA	NA	aalborg univ
35	26011466	NA	NA	aalborg univ
36	23019925	NA	NA	aalborg univ
38	18861657	NA	NA	aalborg univ
39	20241696	NA	NA	aalborg univ

fuzzy_match_wiki_id	fuzzy_match_wiki_name	fuzzy_jw_level
All	All	All
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1
Q601956	aalborg university	0.85 < JW < 1

German organization example (2/2)

DZHW

	FK_INSTITUTIONS	PK_KB_INST	KB_NAME	ORGANIZATION1	CITY	COUNTRYCODE	POSTALCODE
1	24966247	NA	NA	alexander von humboldt inst internet & gesell	berlin	deu	D-10117
2	26263851	NA	NA	alexander von humboldt inst internet & gesell	berlin	deu	NA
3	25284785	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	NA
4	19041909	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	D-10117
5	23459193	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	NA
6	9814790	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	D-10117
7	5548014	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	NA
8	32465471	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	D-10117
9	6357212	NA	NA	alexander von humboldt inst internet & soc	berlin	deu	NA
10	2595255	NA	NA	alexander von humboldt inst internet & soc hiig	berlin	deu	NA

fuzzy_match_wiki_id	fuzzy_match_wiki_name	fuzzy_jw_level	fuzzy_city_status
NA	NA	NA	NA
NA	NA	NA	NA
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches
Q30261359	alexander von humboldt institute for internet and society	0.85 < JW < 1	Matches

Only matched with English Wikipedia, “gesell”

Comparative view to match procedure performances

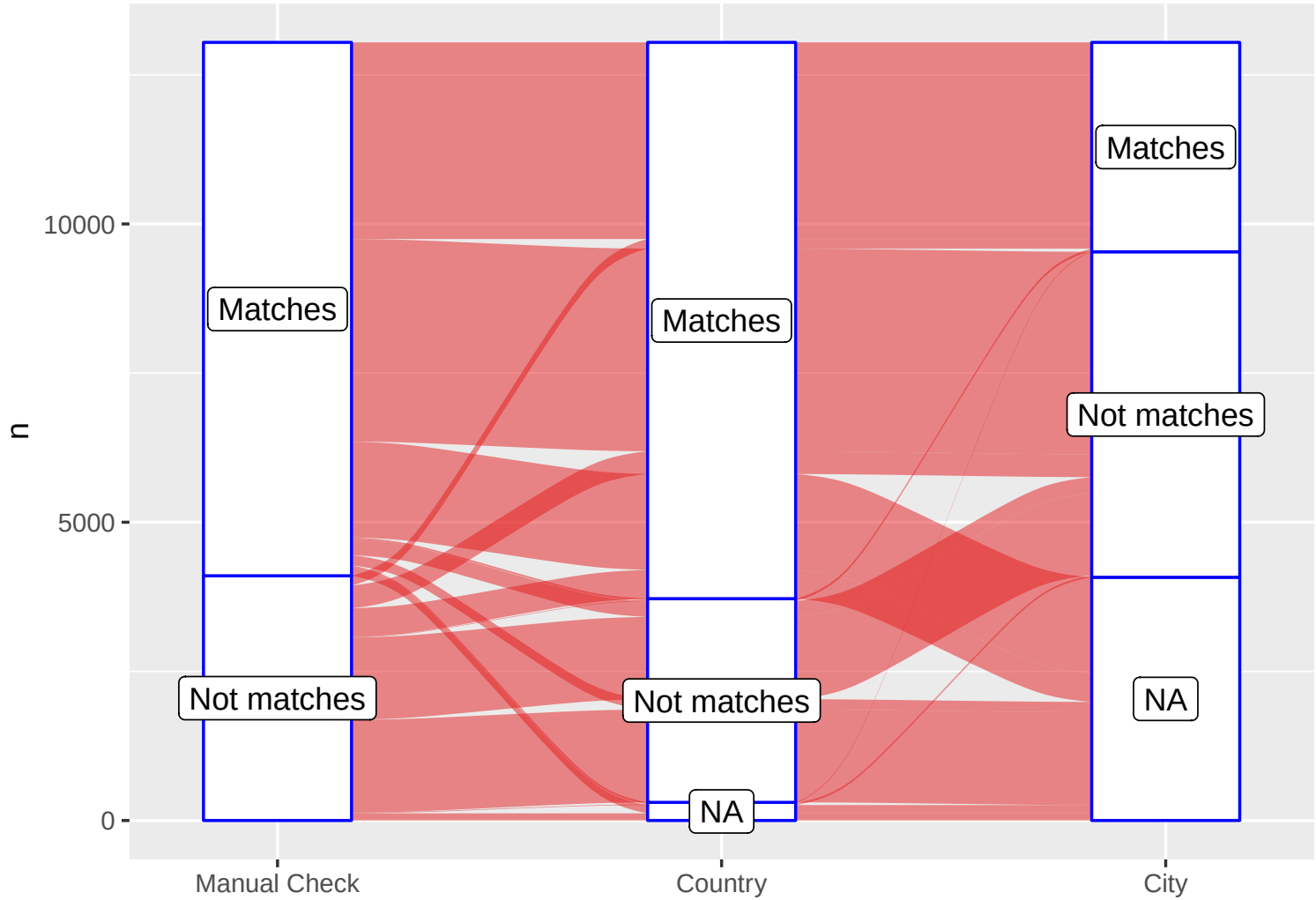
DZHW

Table 2: Comparative view to exact, Fuzzy and GRID match performances in case of each WOS institution

Exact Match	GRID Match	Fuzzy Match	Count
Not match found	Not match found	Best match	32,076
Not match found	Best match	Not match found	2,292
Not match found	Best match	Best match	3,688
Not match found	Best match with 0.85 < JW	Not match found	329
Not match found	Best match with 0.85 < JW	Best match	800
Best match	Not match found	Not match found	1,568
Best match	Not match found	Best match	1,554
Best match	Best match	Not match found	9,839
Best match	Best match	Best match	16,541
Best match	Best match with 0.85 < JW	Not match found	385
Best match	Best match with 0.85 < JW	Best match	130
Best match with 0.85 < JW	Not match found	Not match found	6
Best match with 0.85 < JW	Not match found	Best match	1,237
Best match with 0.85 < JW	Best match	Not match found	595
Best match with 0.85 < JW	Best match	Best match	1,370
Best match with 0.85 < JW	Best match with 0.85 < JW	Not match found	5,995
Best match with 0.85 < JW	Best match with 0.85 < JW	Best match	3,445

Manual check of match procedure performances

DZHW



Countries match performances

DZHW

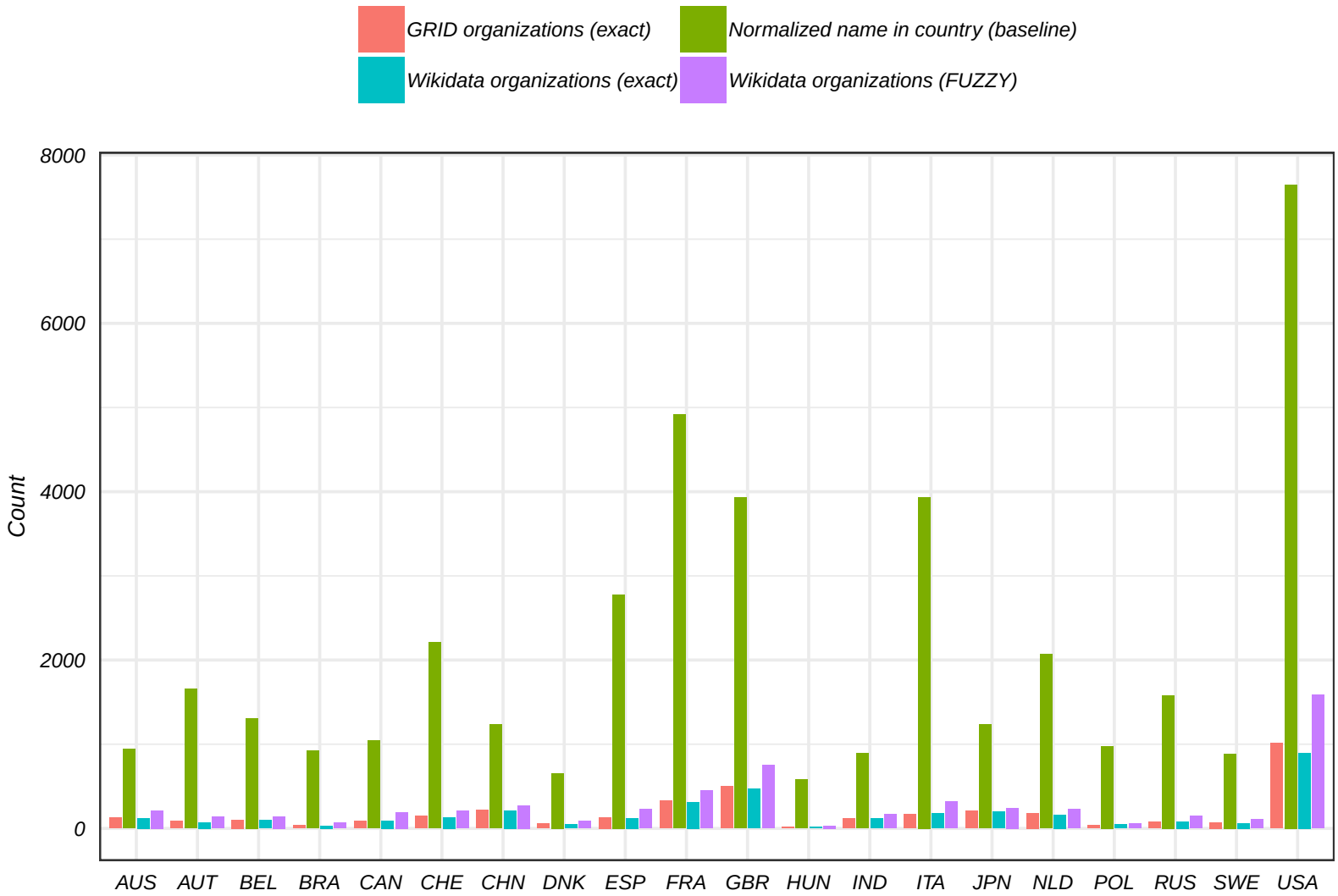
Table 3: Comparative view to exact, Fuzzy and GRID match performances in case of 20 countries with highest number of unique WOS institutions (limited to Berlin collaborations sample)

COUNTRYCODE	FK_INSTITUTIONS	ORGANIZATION1	Normalized name	WOS vs Wikidata (%)*	Wikidata orgs	WOS vs GRID (%)*	GRID orgs	WOS FUZZY vs Wikidata (%)*	Wikidata orgs (FUZZY)
DEU	31,788	21,631	21,726	1204 (5.54%)	706	1300 (5.98%)	765	2100 (9.67%)	1,052
USA	44,958	7,648	7,648	12011 (157.05%)	896	12570 (164.36%)	1,015	15662 (204.79%)	1,587
FRA	23,871	4,915	4,915	2600 (52.9%)	315	2304 (46.88%)	328	4364 (88.79%)	448
ITA	17,007	3,935	3,935	1410 (35.83%)	177	1349 (34.28%)	169	3045 (77.38%)	320
GBR	22,147	3,930	3,930	3990 (101.53%)	471	4104 (104.43%)	499	5891 (149.9%)	754
ESP	11,705	2,779	2,779	1164 (41.89%)	123	479 (17.24%)	134	1192 (42.89%)	230
CHE	10,675	2,212	2,212	992 (44.85%)	132	1109 (50.14%)	150	1369 (61.89%)	207
NLD	11,676	2,068	2,068	2420 (117.02%)	165	2172 (105.03%)	185	2473 (119.58%)	227
AUT	7,657	1,657	1,657	743 (44.84%)	74	792 (47.8%)	85	620 (37.42%)	136
RUS	5,747	1,582	1,582	571 (36.09%)	82	535 (33.82%)	78	787 (49.75%)	146
BEL	6,004	1,308	1,308	371 (28.36%)	99	347 (26.53%)	103	1157 (88.46%)	141
CHN	6,345	1,235	1,235	2309 (186.96%)	213	2332 (188.83%)	218	2210 (178.95%)	274
JPN	6,973	1,234	1,234	2433 (197.16%)	196	2836 (229.82%)	208	1805 (146.27%)	245
CAN	6,643	1,042	1,042	1167 (112%)	92	1493 (143.28%)	90	2016 (193.47%)	192
POL	4,390	978	978	223 (22.8%)	46	540 (55.21%)	40	468 (47.85%)	60
AUS	5,944	945	945	1353 (143.17%)	121	1418 (150.05%)	126	2180 (230.69%)	215
BRA	3,604	925	925	65 (7.03%)	33	87 (9.41%)	38	289 (31.24%)	69
IND	2,096	894	894	349 (39.04%)	120	350 (39.15%)	119	430 (48.1%)	167
SWE	5,896	886	886	1134 (127.99%)	64	2248 (253.72%)	73	2083 (235.1%)	106
DNK	4,454	655	655	358 (54.66%)	45	883 (134.81%)	57	810 (123.66%)	87

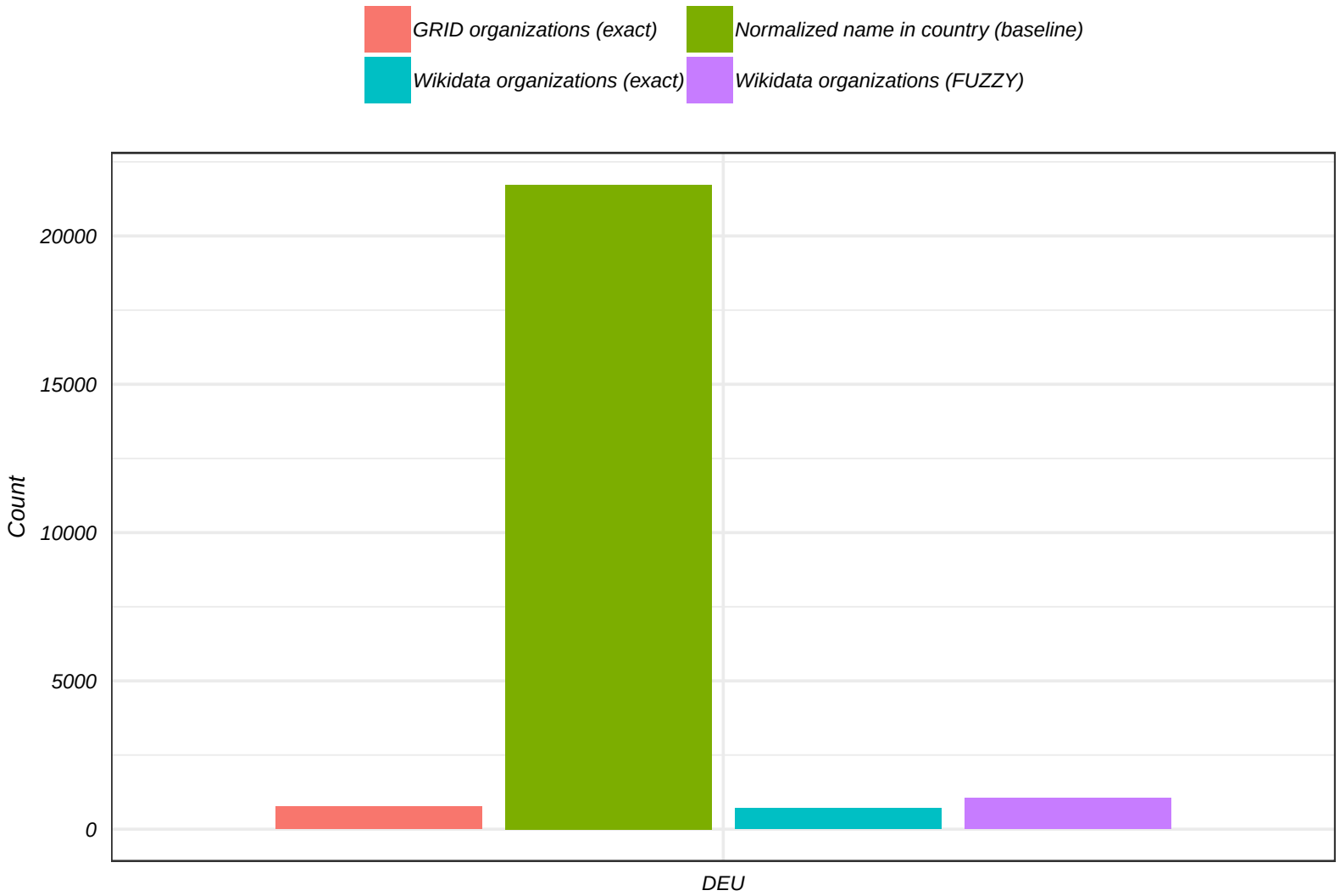
Note:

* These are the number of matches detected in Wikidata or GRID for the unique WOS organizations.
* There could be more than one match detected for any organization.
* Therefore counts and percentages can be higher than unique WOS organizations.

Countries match performances



How about match performance for Germany?



Revised approach, data and methods

DZHW

1. A preliminary view to Berlin region's scientific output
2. All WOS/Scopus **1990-2017** (b_2018 KB) pubs
3. *Article*, *Review* and *Conference proceeding* as **document types**
4. With at least one author/institution from **Berlin**, **Germany**
5. Number of publications, fractional count, 3 years citations

Revised approach, data and methods

DZHW

1. A preliminary view to Berlin region's scientific output
2. All WOS/Scopus **1990-2017** (b_2018 KB) pubs
3. *Article*, *Review* and *Conference proceeding* as **document types**
4. With at least one author/institution from **Berlin**, **Germany**
5. Number of publications, fractional count, 3 years citations

Revised approach, data and methods

DZHW

1. A preliminary view to Berlin region's scientific output
2. All WOS/Scopus **1990-2017** (b_2018 KB) pubs
3. *Article*, *Review* and *Conference proceeding* as **document types**
4. With at least one author/institution from **Berlin**, **Germany**
5. Number of publications, fractional count, 3 years citations

Revised approach, data and methods

DZHW

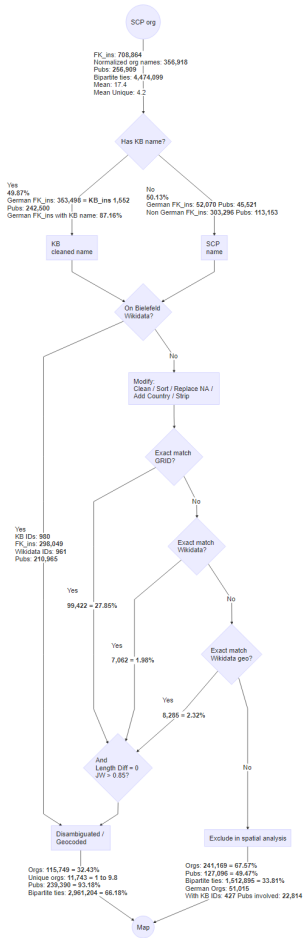
1. A preliminary view to Berlin region's scientific output
2. All WOS/Scopus **1990-2017** (b_2018 KB) pubs
3. *Article*, *Review* and *Conference proceeding* as **document types**
4. With at least one author/institution from **Berlin**, **Germany**
5. Number of publications, fractional count, 3 years citations

Revised approach, data and methods

DZHW

1. A preliminary view to Berlin region's scientific output
2. All WOS/Scopus **1990-2017** (b_2018 KB) pubs
3. *Article*, *Review* and *Conference proceeding* as **document types**
4. With at least one author/institution from **Berlin**, **Germany**
5. Number of publications, fractional count, 3 years citations

Matching logic updated & SCP results



Encoding (problematic?!)

DZHW

Worksheet | Query Builder

```
select PK_INSTITUTIONS, ORGANIZATION1, INSTITUTION_FULL from scopus_b_2018.institutions
where pk_institutions in ('10897353', '29324169', '14472460', '46335179', '31666249', '696390', '40120836','27118260', '28089912',
```

Query Result

SQL | All Rows Fetched: 10 in 0.004 seconds

	PK_INSTITUTIONS	ORGANIZATION1	INSTITUTION_FULL
1	696390	Thï¿½ringer Landessternwarte Tautenburg	Thï¿½ringer Landessternwarte Tautenburg
2	6942610	Universität Bern	Universität Bern, Institute of Psychology
3	10897353	Charitï¿½ - Universitï¿½tsmedizin	Charitï¿½ - Universitï¿½tsmedizin, Department of Neurology
4	14472460	UniversitätSiegen	UniversitätSiegen, Fachbereich Physik
5	27118260	Bundestierärztekammer e.V	Bundestierärztekammer e.V
6	28089912	RÖNTEC GmbH	RÖNTEC GmbH
7	29324169	Max-Delbrï¿½ck-Center for Molecular Medicine	Max-Delbrï¿½ck-Center for Molecular Medicine
8	31666249	University of Tï¿½bingen	University of Tï¿½bingen, Institute of Tropical Medicine
9	40120836	Else-Kröner-Fresenius-Zentrum	Else-Kröner-Fresenius-Zentrum
10	46335179	Institut für Ärztliche Begutachtung	Institut für Ärztliche Begutachtung

We did all this, so what? (1/2)

- Homogeneous and Hetrogeneous collaboration
- Spatial network analysis of coauthorships
- If pubs $>$ 20,000 name, if 1,000 $<$ pubs $<$ 20,000 number

We did all this, so what? (2/2)

KB Sectors

DZHW

Subsector	Color	Count
Akademien der Wissenschaften	grey	4
Behörden/öffentliche Einrichtungen	grey	13
Fachhochschulen	blue	91
Forschungsverbünde, Virtuelle Einrichtungen	grey	7
Fraunhofer-Gesellschaft	pink	57
Helmholtz-Gemeinschaft	grey	18
Internationale Organisationen	grey	2
Kliniken (ausgenommen Universitätskliniken)	red	128
Leibniz-Gemeinschaft	brown	69
Max-Planck-Gesellschaft	black	73
Praxen/Labore	grey	2
Ressortforschung-Bund	grey	12
Ressortforschung-Länder	grey	17
Sonstige	grey	1
Universitäten, Kunst- und Musikhochschulen	yellow	101
Vereine/Verbände	orange	56
Wirtschaft	green	10

One example: Humboldt-Universität zu Berlin

DZHW

- Number of articles: **32,578**
- Number of unique FK_ins: **93,251**
- Number of unique German FK_ins: **50,725**
- Number of unique FK_KB_INST: **700**
- Number of German orgs **without** FK_KB_INST: **2,762**
- Number of NON disambiguated orgs: **2,766**

Preliminary conclusions

DZHW

1. Need for lengthy & time consuming disambiguation
2. It is a must as 1 in 8 WOS (1/9.8 SCP) unique organization IDs proved reliable
3. Network analysis view to collaborations, composition & temporal evolution will be biased without disambigaution

Preliminary conclusions

DZHW

1. Need for lengthy & time consuming disambiguation
2. It is a must as 1 in 8 WOS (1/9.8 SCP) unique organization IDs proved reliable
3. Network analysis view to collaborations, composition & temporal evolution will be biased without disambigaution

Preliminary conclusions

DZHW

1. Need for lengthy & time consuming disambiguation
2. It is a must as 1 in 8 WOS (1/9.8 SCP) unique organization IDs proved reliable
3. Network analysis view to collaborations, composition & temporal evolution will be biased without disambigaution