Cedefop data analysis – About deduplication (draft version 20190802)

Claire de Maricourt, Alexis Eidelman (Dares - France)

Goal: better understanding of deduplication for provided data
Analyzed data:

- Period: data provided from 2019-01-01 to 2019-01-10 (included) [for now it is impossible for us to analyze more data since we have internal difficulties accessing the BDTI]
- Selected variables:
  "IDCOUNTRY" "DAY_GRAB_DATE" "IDESCO_LEVEL_4"
  "IDREGION" "IDCONTRACT" "IDEDUCATIONAL_LEVEL" "IDSECTOR"
  "IDSALARY" "IDWORKING_HOURS" "IDEXPERIENCE" "SOURCE"
  "SITE" "COMPANYNAME" "GENERAL_ID"

Code: https://framagit.org/jocas/jocas_cedefop/tree/master/program

1. Requested offers are not unique

First of all, only 95.6% of requested offers for selected variables are unique. It may be caused by internal deduplication within the sources (it could be worth asking Tabulex about that). We now consider unique requested offers.

2. General_id are reused

The following table shows the amount of unique ids among deduplicated offers.

Total number of offers over the period	730 548
Unique offers (deduplication on all variables)	705 772
Unique general_id	605 799
General_id used once	518 718
General_id repeated	87 081

We tried to identify causes of multiple general id.

3. Possible cases for reusing general_id: multiple sources or multiple regions

We identified two cases on our data when general_id is reused: similar offers with different regions or/and sources.

Repeated GENERAL_ID	
Total	87 081
Multiple regions (same source)	77 844
Mutiple sources (same region)	8 285
Multiples regions and multiple sources	952

It is consistent with what Tabulex said in Thessaloniki: if an offer is advertising for someone in multiple places, they duplicate the offer (eg "Statistician Milano/Roma").

4. Conclusion

Results can slightly change depending on which variables we perform deduplication (especially regional results).

5. Examples

- Similar offer with different sources

ID COUNTRY	DAY_GRAB_DATE	IDESCO_LEVEL_4	ID REGION	ID CONTRACT	ID EDUCATIONAL_LEVEL	ID SECTOR	ID SALARY	ID WORKING_HOURS	ID EXPERIENCE	SOURCE	SITE	COMPANY NAME	GENERAL_ID
FR	1	9313	FR22	2	5	80	*NA*	2	3	FR_JOBIJOBA	jobijoba	synergie	159205063
FR	1	9313	FR22	2	5	80	*NA*	2	3	ADZUNA	jobijoba	synergie	159205063

- Similar offer with different regions

ID COUNTRY	DAY_GRAB_DATE	IDESCO_LEVEL_4	IDREGION	IDCONTRACT	ID EDUCATIONAL_LEVEL	ID SECTOR	ID SALARY	IDWORKING_HOURS	IDEXPERIENCE	SOURCE	SITE	COMPANY NAME	GENERAL_ID
FR	1	2433	FR24	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR53	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR61	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR10	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR22	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR82	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR43	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR71	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR62	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362
FR	1	2433	FR26	4	5	69	*NA*	*NA*	*NA*	EL_SKYWALKER	wizbii	wizbii	159323362

- Similar offer with different regions and sources

ID COUNTRY	DAY_GRAB_DATE	IDESCO_LEVEL_4	IDREGION	ID CONTRACT	ID EDUCATIONAL_LEVEL	ID SECTOR	ID SALARY	ID WORKING_HOURS	ID EXPERIENCE	SOURCE	SITE	COMPANY NAME	GENERAL_ID
FR	3	9312	FR53	2	5	S	*NA*	*NA*	*NA*	NEUVOO	Multi posting	groupe	160061661
FR	3	9312	FR23	2	5	S	*NA*	*NA*	*NA*	NEUVOO	Multi posting	groupe	160061661
FR	3	9312	*NA*	2	5	S	*NA*	*NA*	*NA*	NEUVOO	Multi posting	groupe	160061661
FR	3	9312	FR23	2	5	S	*NA*	*NA*	*NA*	ADZUNA	Multi posting	groupe	160061661
FR	3	9312	*NA*	2	5	S	*NA*	*NA*	*NA*	ADZUNA	Multi posting	groupe	160061661
FR	3	9312	FR53	2	5	S	*NA*	*NA*	*NA*	ADZUNA	Multi posting	groupe	160061661