ASSIGNMENT 3: SILK

Lorenzo Lorgna, Marzorati Stefano, Università degli Studi di Milano Bicocca 08/02/2022

Richiesta

Create and execute a pipeline in SILK to generate sameAs links between Google Adwords and a country of your choice (from Geonames).

Create and execute to the best of your knowledge a pipeline using the SILK framework to generate sameAs links between Google Adwords (https://developers.google.com/adwords/api/docs/appendix/geo/geotargets-2020-03-03.csv) and the country of your choice from Geonames.

Before starting, please refer to the list of the countries in this link (https://docs.google.com/spreadsheets/d/1UKQbgpwOpBbV7Nuc_mKf6a_fUpkHElKnOCrc MulraZg/edit?usp=sharing) to choose the country for which you want to generate links. Add your name near to the country that you prefer to generate sameAs links. You can choose one of the first 10 countries.

ATTENTION

- The max group size refers to the number of students that can choose that state for the generation of links. You can also work in groups for this task.
- Finding the links for a country which has few places does not mean the task is easier :)

1. Links should be generated for each level of the administrative organisation of the country (one pipeline for each level). Tips: Before constructing the pipeline refer to this link https://en.wikipedia.org/wiki/List_of_administrative_divisions_by_country to get familiar with the administrative organisation level for each country.
2. Evaluate the correctness of the links that you have generated.

Remarks

- At the exam, you may be asked questions about how you constructed the pipelines.
- You can export and upload the pipeline (it should have inside the file with the generated links and the ones that you use as reference) in this folder (https://drive.google.com/drive/folders/1VzmWBhWd_ERrDSaAfRzLwxGGI0cQHL Xf?usp=sharing) renamed in the following way: YourSurnameName_CountryAdministrativeLevel.

ASSIGNMENT 3: SILK

Risoluzione

Per la risoluzione è stato utilizzato il framework Silk.

I dati di Google AdWords, informazioni riguardanti posti nel mondo, come i luoghi in cui sono state visualizzate le pubblicità, devono essere linkati ai dati di GeoNames, database di dati geografici.

Come paese è stato considerato l'Italia che presenta i seguenti livelli amministrativi:

- City: 1176
- Province: 110
- Region: 20
- Neighborhood: 7
- Municipality: 5
- District: 1
- Country: 1

Per ognuno di questi è stato creato un linking task e un relativo workflow per generare un file

di testo contenente le istanze matchate in formato N-Triples.

City

Per quanto riguarda City sono stati trovati 1032 match su un totale di 1176.

Province

Per quanto riguarda Province sono stati trovati 110 match su un totale di 110.

Region

Per quanto riguarda Region sono stati trovati 20 match su un totale di 20.

Neighborhood

Per quanto riguarda Neighborhood sono stati trovati 5 match su un totale di 7.

Risultano infatti mancanti su GeoNames i neighborhoods di Pontecchio Marconi e San Nicolò a Trebbia. I 5 neighborhoods matchati risultano essere:

- Lancenigo
- Monsagrati
- Torregaia
- Spezzano
- Silvi Marina

ASSIGNMENT 3: SILK

Municipality

Per quanto riguarda Municipality sono stati trovati 4 match su un totale di 5.

La municipality di Ratschings non viene considerata tra i match perchè possiede un feature code differente (P.PPL) dalle restanti 4 municipality. Quest'ultime 4 risultano essere:

- Napoli
- Fiumicino
- Buccinasco
- Bari

District

Per quanto riguarda District sono stati trovati 1 match su un totale di 1. Risulta essere Trastevere.

Country

Per quanto riguarda Country sono stati trovati 1 match su un totale di 1.

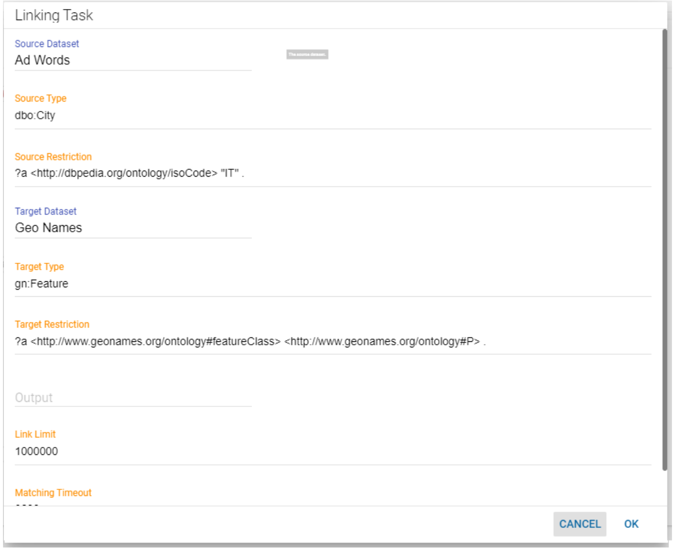
Di seguito la **tabella riassuntiva**:

Livello amministrativo	Tot in Adwords	Matched
City	1176	1032
Province	110	110
Region	20	20
Neighborhood	7	5
Municipality	5	4
District	1	1
Country	1	1

ASSIGNMENT 3: SILK

CITY LINKING

PROPERTIES:



Linking Task

Source Dataset
Ad Words

Source Type
dbo:City

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

Target Type
gn:Feature

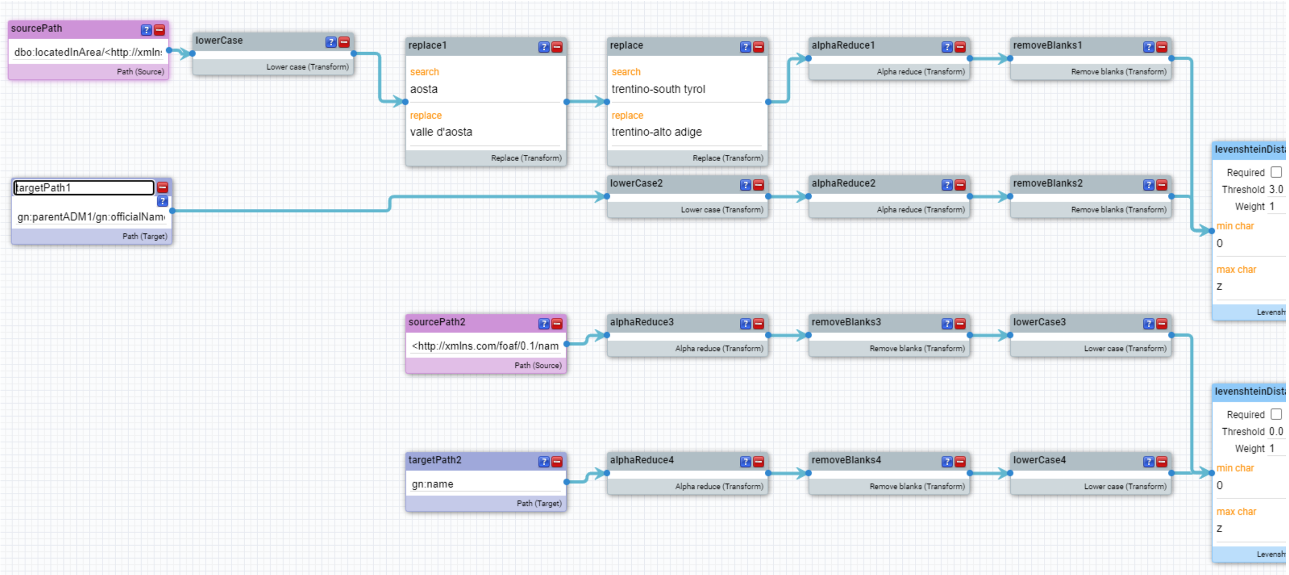
Target Restriction
?a <http://www.geonames.org/ontology#featureClass> <http://www.geonames.org/ontology#P> .

Output

Link Limit
1000000

Matching Timeout

PIPELINE:



sourcePath: dbo:locatedInArea/<http://xmln:

replace1: search aosta; replace valle d'aosta

replace: search trentino-south tyrol; replace trentino-alto adige

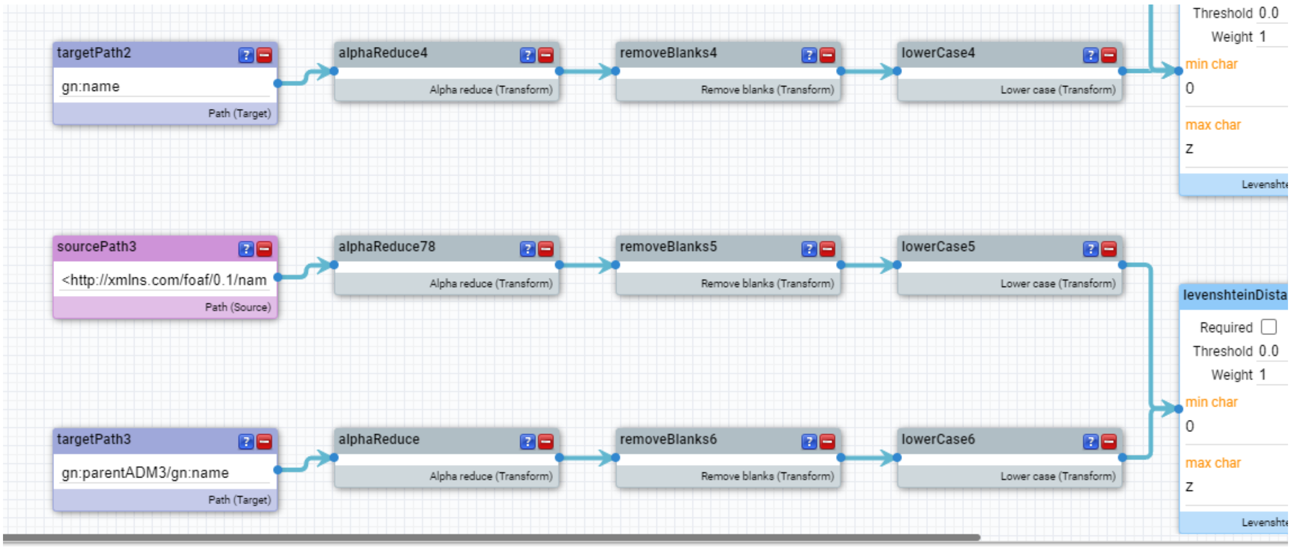
targetPath1: gn:parentADM1/gn:officialNam

levenshteinDista: Required ☐ Threshold 3.0 Weight 1; min char 0; max char z

sourcePath2: <http://xmlns.com/foaf/0.1/nam

targetPath2: gn:name

levenshteinDista: Required ☐ Threshold 0.0 Weight 1; min char 0; max char z



targetPath2
gn:name
Path (Target)

alphaReduce4
Alpha reduce (Transform)

removeBlanks4
Remove blanks (Transform)

lowerCase4
Lower case (Transform)

Threshold 0.0
Weight 1
min char
0
max char
z
Levensht

sourcePath3
<http://xmlns.com/foaf/0.1/nam
Path (Source)

alphaReduce78
Alpha reduce (Transform)

removeBlanks5
Remove blanks (Transform)

lowerCase5
Lower case (Transform)

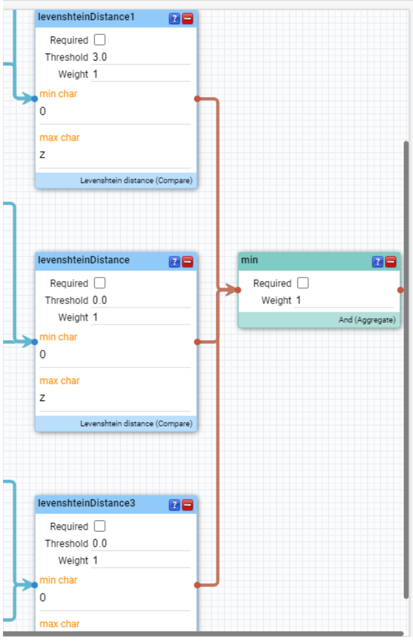
levenshteinDista
Required □
Threshold 0.0
Weight 1
min char
0
max char
z
Levensht

targetPath3
gn:parentADM3/gn:name
Path (Target)

alphaReduce
Alpha reduce (Transform)

removeBlanks6
Remove blanks (Transform)

lowerCase6
Lower case (Transform)



levenshteinDistance1
Required □
Threshold 3.0
Weight 1
min char
0
max char
z
Levenshtein distance (Compare)

levenshteinDistance
Required □
Threshold 0.0
Weight 1
min char
0
max char
z
Levenshtein distance (Compare)

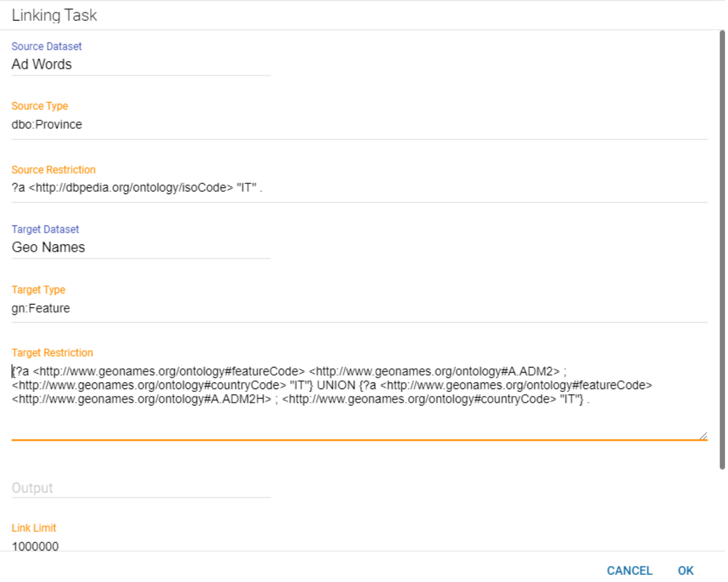
levenshteinDistance3
Required □
Threshold 0.0
Weight 1
min char
0
max char

min
Required □
Weight 1
And (Aggregate)

PROVINCE LINKING

PROPERTIES:

ASSIGNMENT 3: SILK



Linking Task

Source Dataset
Ad Words

Source Type
dbo:Province

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

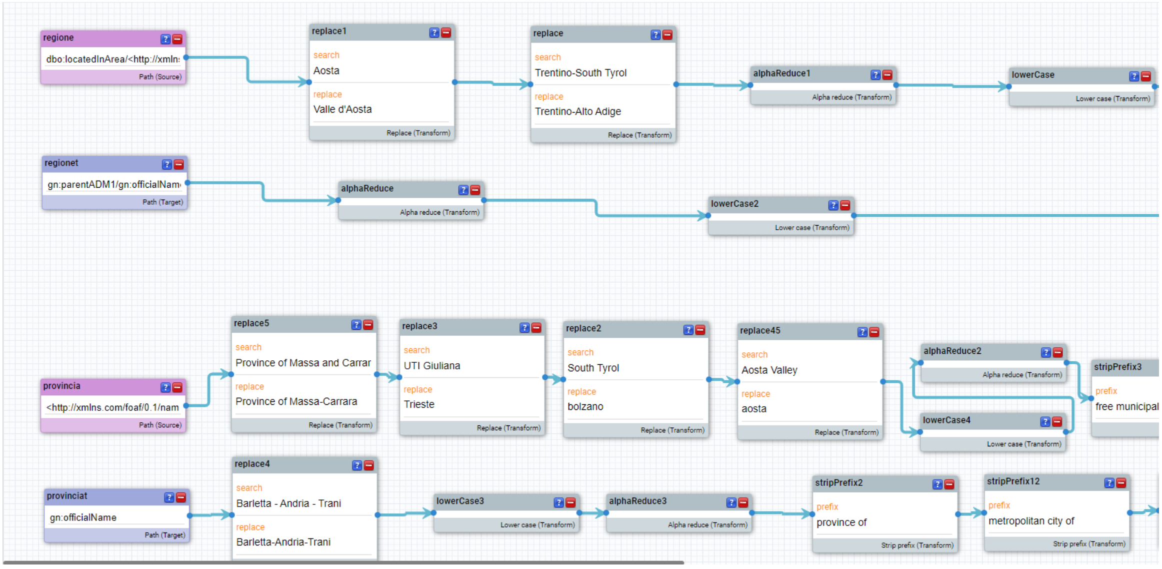
Target Type
gn:Feature

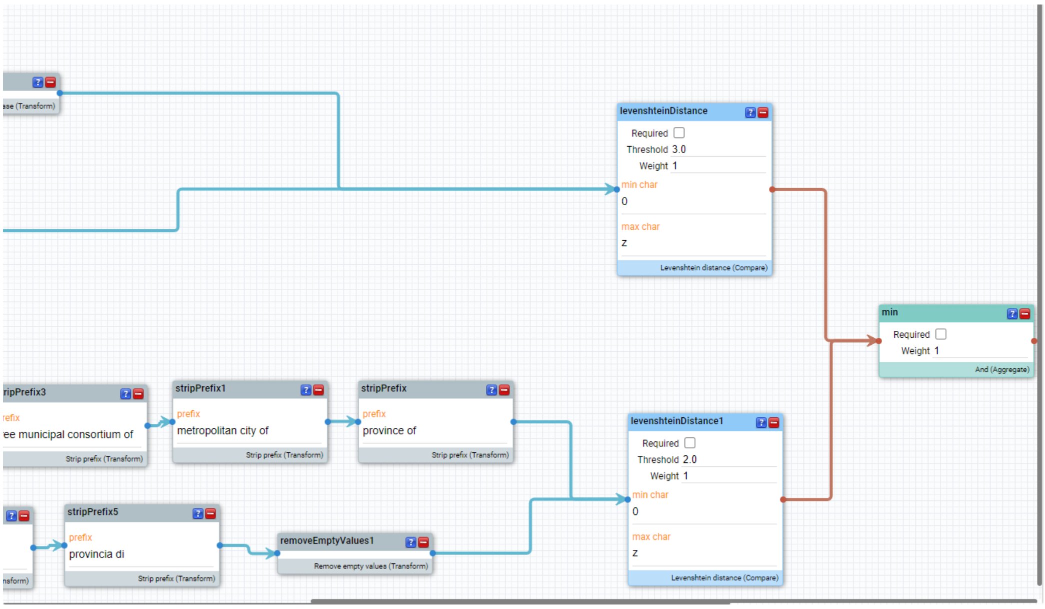
Target Restriction
{?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.ADM2> ; <http://www.geonames.org/ontology#countryCode> "IT"} UNION {?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.ADM2H> ; <http://www.geonames.org/ontology#countryCode> "IT"} .

Output

Link Limit
1000000

PIPELINE:

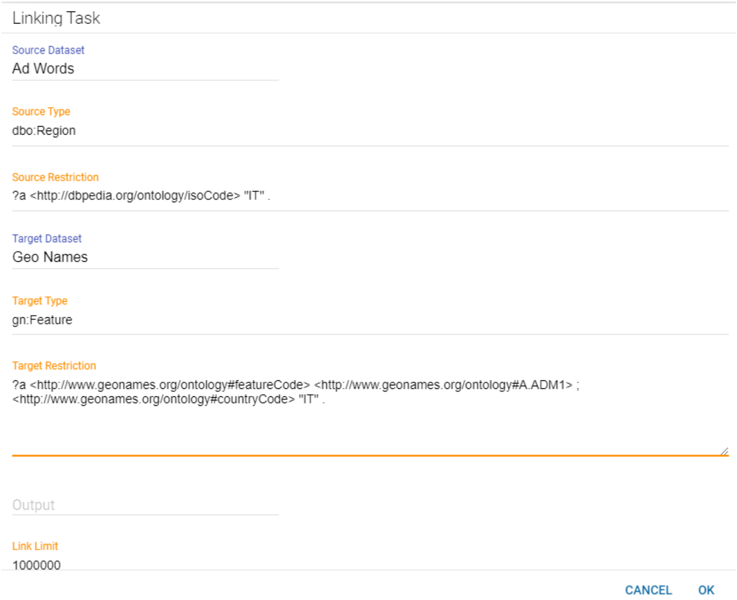




ASSIGNMENT 3: SILK

REGION LINKING

PROPERTIES:



Linking Task

Source Dataset
Ad Words

Source Type
dbo:Region

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

Target Type
gn:Feature

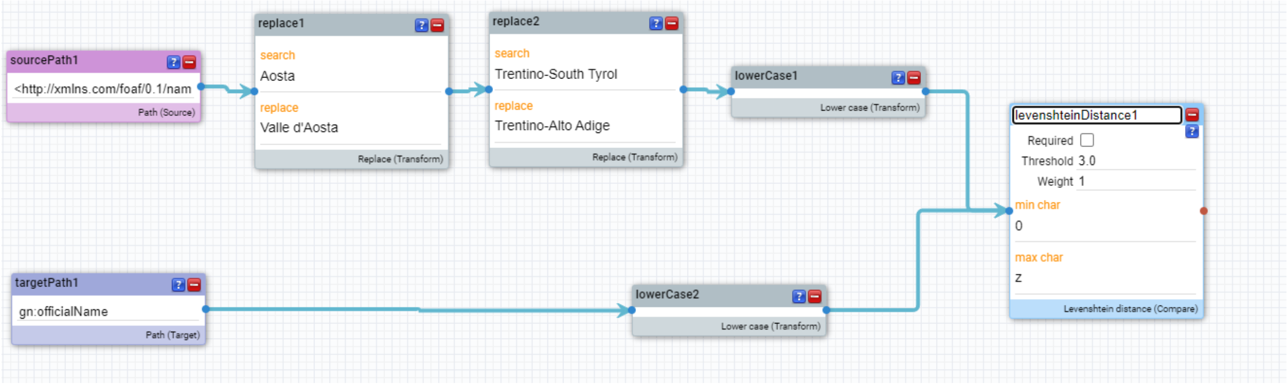
Target Restriction
?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.ADM1> ;
<http://www.geonames.org/ontology#countryCode> "IT" .

Output

Link Limit
1000000

CANCEL OK

PIPELINE:



sourcePath1
<http://xmlns.com/foaf/0.1/nam
Path (Source)

replace1
search
Aosta
replace
Valle d'Aosta
Replace (Transform)

replace2
search
Trentino-South Tyrol
replace
Trentino-Alto Adige
Replace (Transform)

lowerCase1
Lower case (Transform)

levenshteinDistance1
Required ☐
Threshold 3.0
Weight 1
min char
0
max char
z
Levenshtein distance (Compare)

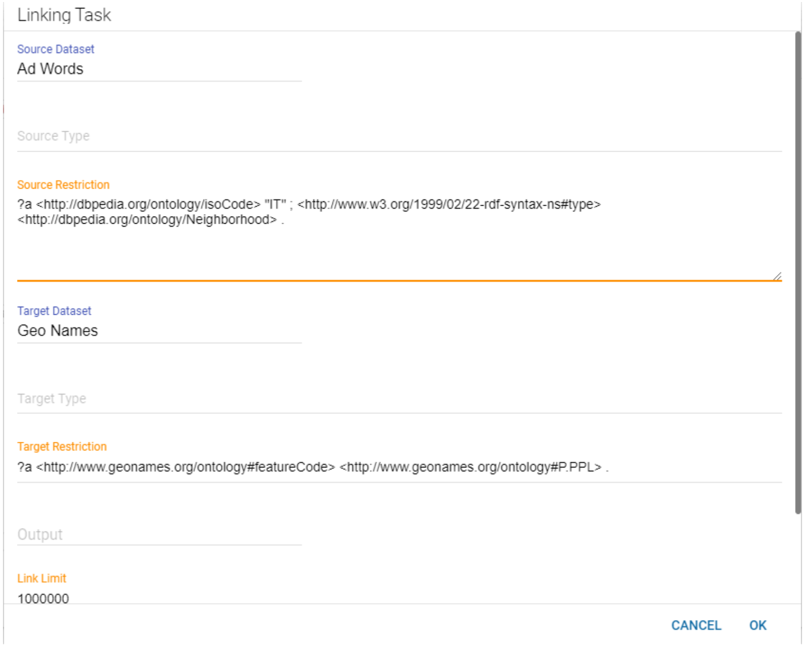
targetPath1
gn:officialName
Path (Target)

lowerCase2
Lower case (Transform)

NEIGHBORHOOD LINKING

PROPERTIES:

ASSIGNMENT 3: SILK



Linking Task

Source Dataset
Ad Words

Source Type

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" ; <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Neighborhood> .

Target Dataset
Geo Names

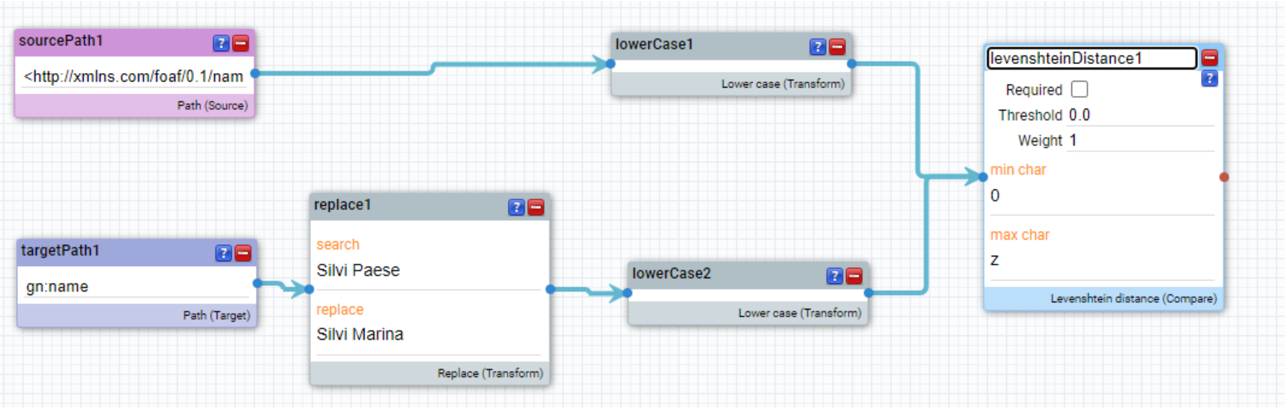
Target Type

Target Restriction
?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#P.PPL> .

Output

Link Limit
1000000

PIPELINE:



sourcePath1
<http://xmlns.com/foaf/0.1/nam
Path (Source)

targetPath1
gn:name
Path (Target)

replace1
search
Silvi Paese
replace
Silvi Marina
Replace (Transform)

lowerCase1
Lower case (Transform)

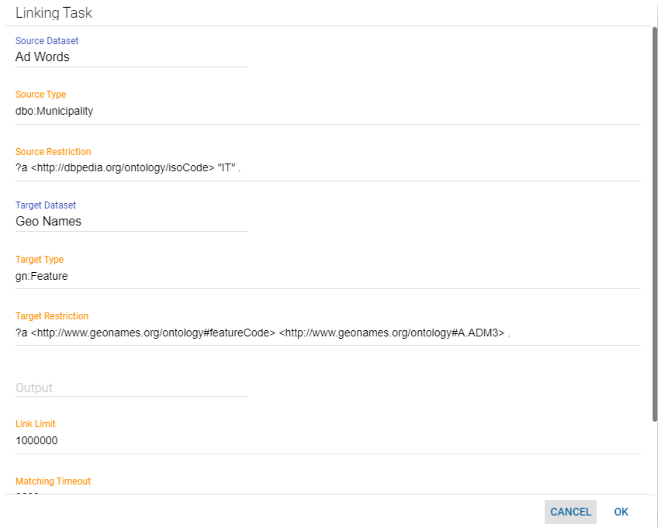
lowerCase2
Lower case (Transform)

levenshteinDistance1
Required □
Threshold 0.0
Weight 1
min char
0
max char
z
Levenshtein distance (Compare)

MUNICIPALITY LINKING

PROPERTIES:

ASSIGNMENT 3: SILK



Linking Task

Source Dataset
Ad Words

Source Type
dbo:Municipality

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

Target Type
gn:Feature

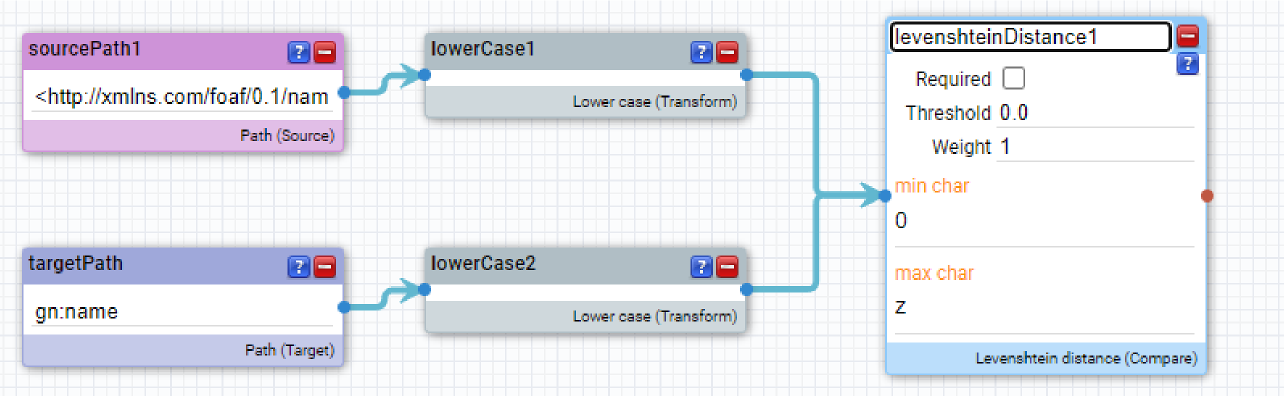
Target Restriction
?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.ADM3> .

Output

Link Limit
1000000

Matching Timeout

PIPELINE:



sourcePath1
<http://xmlns.com/foaf/0.1/nam
Path (Source)

targetPath
gn:name
Path (Target)

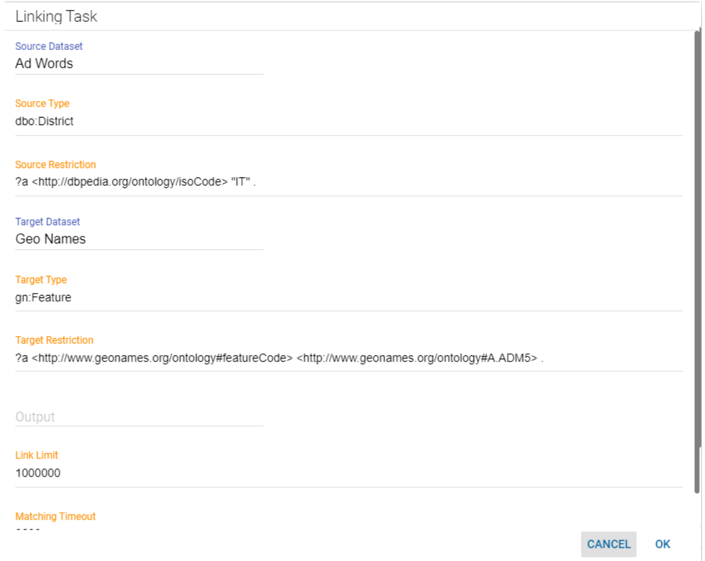
lowerCase1
Lower case (Transform)

lowerCase2
Lower case (Transform)

levenshteinDistance1
Required □
Threshold 0.0
Weight 1
min char
0
max char
z
Levenshtein distance (Compare)

DISTRICT LINKING

PROPERTIES:



Linking Task

Source Dataset
Ad Words

Source Type
dbo:District

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

Target Type
gn:Feature

Target Restriction
?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.ADM5> .

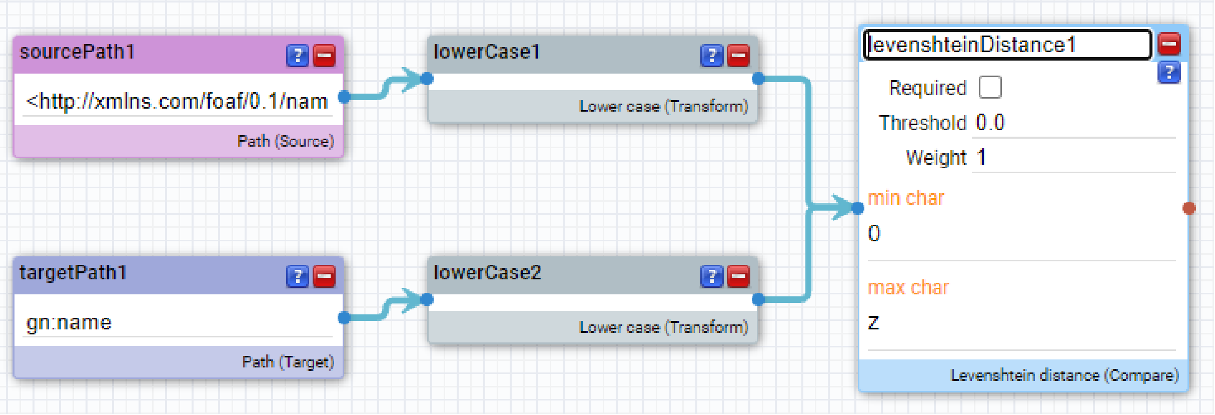
Output

Link Limit
1000000

Matching Timeout

ASSIGNMENT 3: SILK

PIPELINE:



Required □

Threshold 0.0

Weight 1

min char

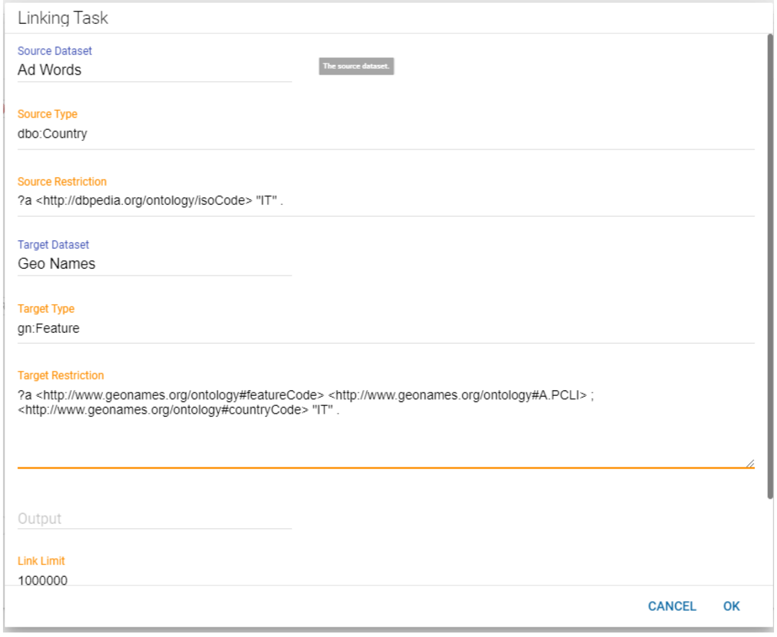
0

max char

z

COUNTRY LINKING

PROPERTIES:



Linking Task

Source Dataset
Ad Words

Source Type
dbo:Country

Source Restriction
?a <http://dbpedia.org/ontology/isoCode> "IT" .

Target Dataset
Geo Names

Target Type
gn:Feature

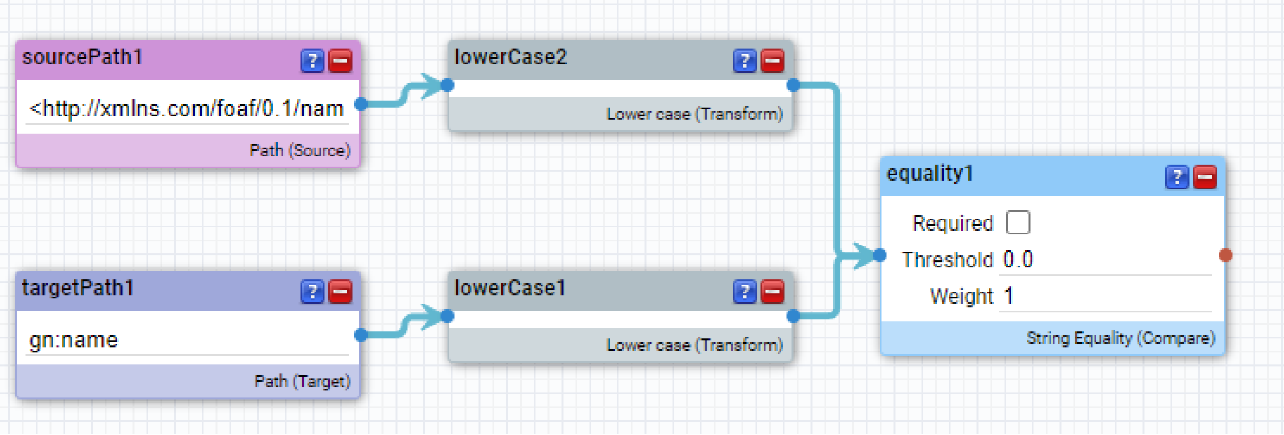
Target Restriction
?a <http://www.geonames.org/ontology#featureCode> <http://www.geonames.org/ontology#A.PCLI> ;
<http://www.geonames.org/ontology#countryCode> "IT" .

Output

Link Limit
1000000

PIPELINE:

ASSIGNMENT 3: SILK



sourcePath1

<http://xmlns.com/foaf/0.1/nam

Path (Source)

targetPath1

gn:name

Path (Target)

lowerCase2

Lower case (Transform)

lowerCase1

Lower case (Transform)

equality1

Required □

Threshold 0.0

Weight 1

String Equality (Compare)