The benefits of Linked Data for the social sciences

Analysing economic drivers and network effects of international migration based on semantically integrated data

Francesca Ceolan[1,2], Dimitris Alivanistos[3,4], Kathrin Dentler[3,5], Albert Meroño-Peñuela[3], Richard Zijdeman[1], Rinke Hoekstra[3], and Auke Rijpma[6]

Abstract. International migration is a phenomenon hard to analyse comprehensively, partly because of the lack of a single database that contains all the required data. Therefore, we converted three different datasets and recollected the data with the use of a single query. In this study, we demonstrate the utility of Linked Data as a flexible technology to manage data, and the possible use it can have in social sciences. Based on the selected data, we performed several regression analyses to validate two theories of the migration studies field: neoclassical economic theory and ethnic networks theory.

Keywords: linked data, migration theories, semantically integrated data, migrant integration, unified migration portal.

1 INTRODUCTION

Linked Data has been proven to be a flexible technology to describe, integrate, manage, expose and query many kinds of data in various domains, including the humanities [1], [2] and history [3]. However, Linked Data technology is rarely applied in social science. Therefore, we want to showcase with our study that Linked Data can be applied successfully to social science research. Specifically, we investigate international migration in several European countries. In our project, we use Linked Data technology to create an open, easily accessible and comprehensive data model with data on international migrants in the 28 European countries.

Much has been written on the subject of international migration, but in recent years it has become an even more widely discussed topic, holding many societal implications [4]. For this paper, we take European countries as a case study. Europe has been a popular immigrant destination for many decennia, and is nowadays experiencing one of the most significant inflows of migrants in history, both from outside the continent and within.

Our application of Linked Data to the field of migration studies shows that the benefits of this technology can be exploited also in social science. Our aim is not to research an innovative topic of migration studies, but to demonstrate that important, well-established analyses can be conducted based on Linked Data. The lack of literature on this topic [5] shows how this is an original and innovative study. Consequently, our contribution to the scientific community is the application of a recently developed technology to a different and differently established field.

1.1 RESEARCH QUESTIONS

- RQ1: *How can Linked Data technology help to understand international migration studies, a branch of social science?*
- RQ2: *How are the unemployment rate and the presence of immigrants in a country related to immigration to the country? And how does this relation differ between countries?*

We aim to answer these questions through the creation of a Unified Migration Portal (UMP), which consists of data about migration, refugees and countries converted to RDF (Resource Description Framework). Our UMP wants to be accessible and easily integrated through its shared vocabularies that make the expansion of variables easier.

2 DATASET DESCRIPTION

To answer our research questions, we use three different datasets, selected because they provide reliable data with indicators that have already been harmonized between countries [6], [7].

The first dataset is from the Organisation for Economic Co-operation and Development[7] and contains information regarding inflows of population between countries all over the world. Specifically, it contains the following columns: *Code* (referring to country code), *Country of birth/nationality, Variable* (in our case the *Inflows of foreign population by nationality*), *Gender, Country, Year, Value.*

The second dataset is from the Directorate-General of the European Commission, well known as Eurostat[8], and contains information about the population of European countries. Specifically, there are the *GEO* column containing the country the row corresponds to, the *CITIZEN* column which contains the type of population (e.g. Total population and Foreign-born population), the *TIME* column, corresponding to the time period that the population was measured and the *VALUE* where the population value is stored. Moreover, the *AGE, SEX, and UNIT* columns contain information about the age group (e.g. 12-25), gender (male, female, total) and the unit of the measurement (Number, Percentage etc.).

The third dataset is also from Eurostat[9], holding information about the unemployment rate of countries that are part of the

---

[1] International Institute for Social History (IISH), KNAW, the Netherlands
[2] email: f.ceolan@students.uu.nl
[3] Vrije Universiteit Amsterdam, the Netherland
[4] email: d.alivanistos@students.vu.nl
[5] email: k.dentler@vu.nl
[6] Utrecht University, the Netherlands, email: a.rijpma@uu.nl
[7] OECD.Stat, International Migration Database, Inflows of foreign population by nationality. Available at: https://stats.oecd.org/Index.aspx?DataSetCode=MIG.
[8] Eurostat, population database, “Population on 1 January by age, sex, and broad group of citizenship (migr_pop1ctz). Available at: http://ec.europa.eu/eurostat/data/database
[9] Eurostat, unemployment database, total unemployment rate (tsdec450). Available at: http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tsdsc330&plugin=1

Eurozone. As it follows the structure of Eurostat datasets, it contains the same columns as the population dataset.

3 ONTOLOGY DESIGN

Conceptualizing a Unified Migration Portal is a non-trivial task. A variety of multi-dimensional data has to be taken into account, such as unemployment rate and population, which in turn have more dimensions such as gender, time, and type. We modelled country data as *Indicators* of a country, indicating whether a country is or will be a possible destination for migration and why.

Furthermore, we modelled inflow and outflow of population and asylum seeking data as *Movement Of People*. The reason behind this design choice is that migration can be regarded as movement of a number of people from one place to another (in our case from country to country), in a specific time period (e.g. year).

This concept can be even more generic and extendible by adding other types of information like the gender of those who moved, their age group, etc. The unifying entity is *Country*, as it provides the link between movement_of_people and country_indicators. The resulting ontology is visualized in Figure 1.

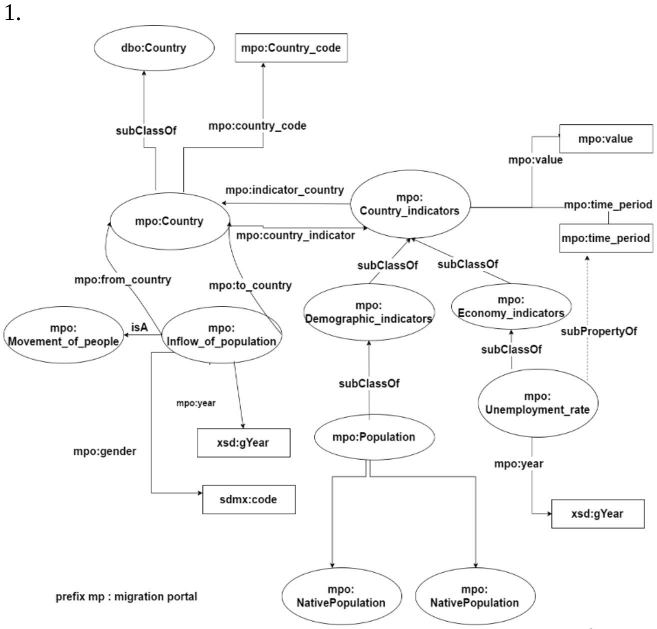

Figure 1. Schematic representation of our ontology[10].

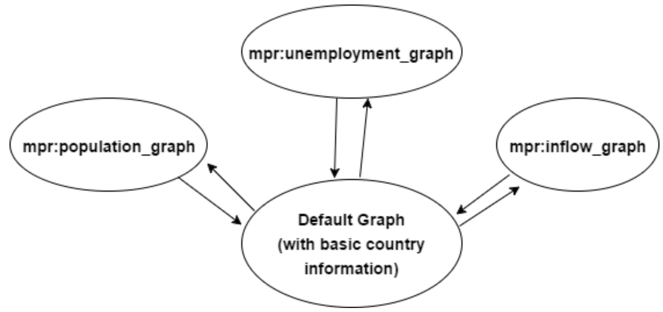

Figure 2. Visualization of the graphs in our triplestore.

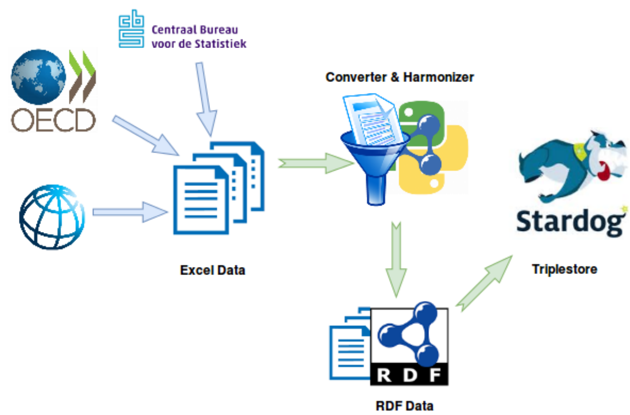

Figure 3. The pipeline for the data input to our Unified Migration Portal.

4 DATA CONVERSION AND STORAGE

We converted the data according to our ontology. The goal was to convert each dataset to RDF [8]. We used [9] RDFlib[11] to perform programmatic additions of class terms and individuals. During the conversion phase we also linked entities in our ontology with well-known online datasets, such as DBpedia [10] for countries, SDMX [11] for gender and geonames [12] for country codes.

To meet the requirements of the multi-dimensionality of the data, we made a specific design choice that needs to be emphasized. In our data model, we chose to have our main entity in our *default graph* and the different converted data sets in separate *named graphs* that connect to the default graph via the Country entity. This results in high extensibility of the Unified Migration Portal as it enables us to further link different types of data using the same infrastructure as described above. We present a visual representation of the resulting graphs in Figure 2.

5 THE TRIPLESTORE

For creating the endpoint where the migration portal can be reached, we chose an enterprise data unification platform

[10] A bigger version of Figure 1 and 2 can be found in the appendix.

[11] RDFLib is a pure Python package to work with RDF.

leveraging smart graph technology, called Stardog 4.2[12]. It is an OWL/RDF data store, which supports the RDF graph data model and the SPARQL [13] query language.

Stardog provides us with a query endpoint, where we tested the proper integration of each dataset after every conversion. This feature has also been proven useful for the evaluation of our work, which we will discuss in a later section.

6 SPARQL QUERIES

To ensure that we have indeed established the foundation for a Unified Migration Portal, we had to make sure in every step of the process that our triplestore would produce the desired results. Thus, we tested a variety of different SPARQL queries against the different datasets separately but also against the resulting unified graph. Some of the use cases we looked into were to gather the total population of European countries in a specific calendar year, getting countries that have a total population below a specific threshold (according to the user's research needs) and returning all countries that had an unemployment rate below 5 percent. However, the effectiveness of Linked Data lies in the combination of such query results. In the Appendix, we include the query that returns the multidimensional data from three graphs, upon which we investigate the correlations between population (native, foreign-born), unemployment rate and inflow of population per year for the past decade.

7 THEORIES AND HYPOTHESES

We used the queried data to find a pattern in the migration (inflow) to 28 European countries. In the following section, we test two previously established theories in the migration field.

According to the definition provided by the International Organization for Migration (IOM), international migration is "the movement of people, either across international borders or within a state"[13]. We will focus on the first part of the definition, namely international (as opposed to national) migration. What is of interest for this paper is studying whether there are some underlying factors, common to all European countries, which help to understand how the inflow of migrants works across countries. The focus is on two well-developed theories in the migration field, namely neoclassical economic theory and migration networks theory. Each one is measured with a main immigration predictor, respectively the economic situation of the host country and the immigrant networks already present.

7.1 NEOCLASSICAL ECONOMIC THEORY

This theory uses the economic concept of push and pull factors to explain international migration. The main assumption this theory makes is that individual migration happens based on cost-benefit calculations, and migrants will move where their wellbeing is maximized [14].

Here, the unemployment rate in the country of destination is taken as an indicator of how well a country is doing economically, when it is low it is assumed that it will be easier for the migrant to find employment, making it a more attractive destination [15]. The likelihood to find a job is thus considered a pull factor, as a low unemployment rate will "pull" people to a country as it means less competition on the labour market, and this results in more opportunities to find a job.

Unemployment is taken as a proxy of how well a country is doing economically, as previous literature has shown that migration tends to happen from "poorer" countries to "richer" ones. As Janissen [17] found, unemployment has a negative impact on net migration.

Therefore, we expect that: *the lower the level of unemployment in a country, the higher will the number of immigrants be (H1).*

7.2 NETWORKS THEORY

As Castels [18] states, "a purely economic understanding of migration can be misleading". Therefore, this paper also uses another driver to try to explain migration to European countries, namely network theory.

The networks we refer to in this paper are ethnic ones, and they are measured by the number of foreign-born people living in the host country (i.e. *number of foreigners of a specific nationality*).

The presence of an already existing co-ethnic network means that immigrants are assumed to be more likely to receive support from immigrants from the same origin country already established in the host country. This would reduce both migration costs and risks [19], incentivizing migrants to move to countries where there is already a big community of fellow citizens.

This is true even when the immigration policies of a country become rather restrictive, as it accounts for family reunification migration. Moreover, Pedersen and colleagues (2008) found that the presence of immigrants from one's own nationality already residing in a country has a large and significant positive impact on immigration inflow.

Therefore, we hypothesize that *the more foreigners reside in a given country, the higher will the number of immigrants to that country be (H2).*

According to the two theories, Figure 4 reports the conceptual model that we operationalized.

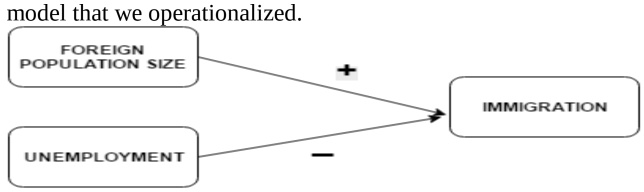

Figure 4. Conceptual model

[12] Stardog enterprise data unification platform. Available at: stardog.com

[13] IOM Key migration terms. Available at: http://www.iom.int/key-migration-terms

8 VARIABLES AND DESCRIPTIVE STATISTICS

8.1 MEASURES

The variables we use to assess the level of immigration to the 28 European countries are retrieved by querying our previously constructed triplestore, and measured as follows. The predicting variables foreign population size and unemployment rate are measured across countries, from 2006 to 2015. Due to the lack of data for most counties, the year 2015 was dropped from the analyses. The variable *foreign-born population* is used as a proxy to measure the size of ethnic networks as it reports the number of born-abroad people per country. *Unemployment rate* is used as a proxy for how well the country is doing economically, and it is reported for all the years. *Inflow of foreigners* is also reported for each country for the respective years and being the dependent variable, all countries with missing values were excluded from the analyses[14]. This resulted in the loss of 95 cases (see Table 1), most of which were clustered in countries that joined the EU more recently.

To account for the various size of the population, the analyses were conducted with relativized values for inflow and foreign population data. The relativized results do not differ in significance and direction of relation from the ones with absolute values, but are easier to interpret. Therefore, all the following results report values relative to the countries populations' size.

Furthermore, the variables *inflow* and *foreign population* were highly skewed, and to improve their distributions we imputed their natural logarithms. The variable *inflow* was also divided by 10, since its variance was otherwise too big to be manageable.

8.2 DESCRIPTIVE STATISTICS

Table 1 reports the descriptive statistics for the variables before the transformations used for the regressions. It is possible to observe that the countries report a wide range of inflow of foreigners, as well as population and unemployment rate.

Table 2 reports the correlations between the main variables in the study, in absolute values. Inflow is significantly correlated to foreign population and unemployment. As expected, with the first one it has a positive correlation, and with the latter a negative one.

9 ANALYSIS

To assess the impact of unemployment and ethnic networks on immigration to EU countries, we performed regression analysis in IBM SPSS Statistic 24[15] and in Mplus (version 7.4) [20]. As the variables were not normally distributed, we used maximum likelihood estimation with robust standard errors.

We proceeded in three different steps. First, we analysed all the countries together. Second, we selected Germany as case study, because of its size, central geographical position and long history as migration country. Third, we split the countries into four geographical groups (south, north, centre and east Europe[16], to test whether they present significant differences between them.

The first regression shows that inflow is negatively related to unemployment (*b*=-14.08, S.E.=6.11, *p*<.05)[17] and positively to ethnic networks (*b*=88.86, S.E.=26.47, *p*<.001), confirming both our hypotheses.

The second regression, performed on Germany, again confirms our hypotheses. Inflow is positively related to ethnic networks (*b*=363.34, S.E. = 3.7, *p*<.001) and negatively to unemployment (*b*=-7.02, S.E. = 2.766, *p*<.05).

The third regression, performed simultaneously with multiple group analysis on the four geographical groups, shows non-significant, mixed results. While the hypotheses are confirmed for the central and eastern European countries, the relations are reversed for the southern countries and for the Nordic ones they are both negative. This might be because of the contextual differences between countries with regard to immigration policies, welfare system and other further country-specific factors.

Table 1: Descriptive statistics

	Range	Mean	SD	N	Missings
Inflow	1107-1342529	140599.10	214339.11	194	95
Year	2006/2015			289	0
Total population	37366-82437995	18274133.63	23111486.29	289	0
Unemployment	2.5-27.5	8.97	4.479	289	0
Foreign population	12000-7289149	1147383.51	1790083.64	289	0
Relative foreign population	0-.86	.081	.097	289	0
Relative inflow	0-.04	.008	.007	194	95
Geo	1-4	2.53	1.19	239	50

SD; standard deviation, N: number of observations; Missings: missing values for each variable

[14] The countries with no data for the inflows: Croatia, Cyprus, Lithuania, Malta, Romania and Slovakia.

[15] IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp

[16] South=France, Greece, Italy, Portugal and Spain
North= Denmark, Finland, Norway and Sweden
Center= Austria, Belgium, Germany, Luxembourg, the Netherlands and the United Kingdom.

[17] Where *b* denotes the regression coefficient, and S.E the standard error.

Table 2. Correlations of relative values

	1	2	3
Inflow (1)			
Population (2)	-.222*		
Foreign population (3)	.247**	-.078	
Unemployment (4)	-.361***	.066	-.059

***p<.001; **p<.01, *p<.05

10 DISCUSSION AND FUTURE RESEARCH

With regard to our statistical analyses, some limitations should be noted. The data we used in the study has some restrictions regarding years, countries and information lacking some important predictors from immigration theories. However, future research could incorporate further data, as adding new databases is easy with Linked Data. This would enable researchers to answer more innovative research questions. For example, the addition of the migrants' educational level would allow investigating the brain drain phenomenon [21], i.e. the migration of highly skilled workers towards countries that offer them better economic possibilities. With regard to the hypotheses derived from networks and neoclassical economic theories, two clarifications should be made. Firstly, the performed analysis, even though based on previously proved theories, is purely illustrative. Its aim is not to predict migration but to show that this can be done starting from Linked Data. Second, an assumption was made regarding ethnic networks theory, namely that immigrants are able to engage with the ethnic networks present in the country they migrate to.

Moreover, the mixed results that appear when dividing the countries geographically demonstrate the need for further research to explain the substantial differences between the various countries. With Linked Data, this could be done with a query targeting the specific countries groups.

11 CONCLUSION

In this study, we integrated information from two trustworthy online databases, namely OECD and Eurostat, to demonstrate that important, well-established analyses can be reproduced based on Linked Data.

Our results demonstrate the benefits of Linked Data technology to integrate and prepare data commonly used in social science. Linked Data provides a flexible way of integrating information from different sources, and to access it directly from one single source to conduct research. This way, data becomes accessible, easy to query, analyse and visualize. Moreover, studies become more reproducible, allowing other researchers to duplicate and validate the findings. In our case, we published everything online: our original datasets, the converters, the converted datasets and the queries on GitHub[18], and all converted data ready to be queried in our SPARQL endpoint[19].

In the future, Linked Data can be game changing in allowing for collaborative efforts. Anybody can publish and link his or her own data to existing ontologies and other datasets in a sustainable way, and reuse data, queries and analyses published by others. Tools as qber [1] will hide the technicalities of Linked Data under the hood, while offering all its benefits. This will affect workflows: while traditionally, each researcher needs to integrate the required datasets himself, which is typically a time-consuming activity, integrated data will be available for use, allowing for research on a much broader scale. Such a unified portal might serve social scientists, data journalists or even (future) migrants themselves to gain access to more information.

In conclusion, we demonstrated how Linked Data can be applied to social sciences, in particular to successfully analyse international migration. We regard Linked Data as a flexible paradigm for gathering, implementing, analysing and describing data often used in social sciences.

ACKNOWLEDGMENTS

We would like to thank Amsterdam Data Science for funding our project.

REFERENCES

[1] R. Hoekstra, A. Meroño-Peñuela, K. Dentler, A. Rijpma, R. Zijderman, I. Zandhuis, „*An Ecosystem for Linked Humanities Data*. In Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016)," *ESWC*, 2016.

[2] A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, R. Hoekstra, S.Schlobach, „Linked Humanities Data: The Next Frontier?. A Case-study in Historical Census Data," *Proceedings of the 2nd International Workshop on Linked Science*, vol. 951, 2012.

[3] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Scholbah, F. Van Harmelen, „Semantic technologies for historical research: A survey," *Semantic Web*, vol. 6, nr. 6, pp. 539-564, 2014.

[4] BBC, „Why is EU struggling with migrants and asylum?," 3 03 2016. [Online]. Available: http://www.bbc.com/news/world-europe-24583286. [Geopend 13 01 2017].

[5] M. Devedžić & V. Devedžić, „Imagine: Using New Web Technologies in Demography," *Social Science Computer Review*, 2009.

[6] OECD, „ Data and metadata reporting and representation handbook.," 2007. [Online]. Available: http://www.oecd.org/std/37671574.pdf.

[7] „Where Eurostat data is coming from," Eurostat, Methodology, [Online]. Available: http://ec.europa.eu/eurostat/web/population-demography-migration-projections/methodology. [Geopend 13 01 2017].

[8] G. Klyne, J. J. Carrol, „Resource description framework (RDF): Concepts and abstract syntax," 2006. [Online].

[18] The migration portal repository

[19] http://stardog.clariah-sdh.eculture.labs.vu.nl/test#!/query

[9] Python Language Reference, version 2.7., Python Software Foundation, [Online]. Available: http://www.python.org. [Geopend 13 01 2017].

[10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R Cyganiak, Z. Ives, „Dbpedia: A nucleus for a web of open data,” in *The semantic web. Lecture Notes in Computer Science, vol 4825.*, Berlin Heidelberg, Springer, 2007, pp. 722-735.

[11] „Statistical Data and Metadata Exchange (SDMX),” International Organisation for Standardisation, [Online]. Available: https://sdmx.org/?page_id=2555. [Geopend 13 01 2017].

[12] B. Vatant, M. Wick, „Geonames ontology,” [Online]. Available: www.geonames.org/ontology/documentation.html.

[13] E. Prud’Hommeaux, A. Seabone, „SPARQL query language for RDF,” *W3C recommendation 15,* 2008.

[14] G. J. Borjas, „Economic theory and international migration,” *International migration review,* pp. 457-485, 1989.

[15] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, J. E. Taylor, „Theories of international migration: A review and appraisal.,” *Population and development review,* pp. 431-466, 1993.

[16] D. Ratha, C. Eigen-Zucchi, S. Plaza, „Migration and Remittances Factbook 2016,” *World Bank Publications.,* 2016.

[17] R. Jennissen, „Economic determinants of net international migration in Western Europe.,” *European Journal of Population/Revue Européenne de Démographie,* vol. 19, nr. 2, pp. 171-198, 2003.

[18] S. Castles, „The forces driving global migration,” *Journal of Intercultural Studies,* vol. 34, nr. 2, pp. 122-140, 2013.

[19] A. M. Mayda, „International migration: A panel data analysis of the determinants of bilateral flows,” *Journal of Population Economics,* vol. 23, nr. 4, pp. 1249-1274, 2010.

[20] L. K. Muthén & B. O. Muthén, Mplus User's Guide. Sixth Edition., Los Angeles, CA: Muthén & Muthén, 1998-2010.

[21] H. Docquier, H. Rapoport, „Globalization, brain drain, and development,” *Journal of Economic Literature,* vol. 50, nr. 3, pp. 681-730, 2012.

APPENDIX A

```
prefix mpo: <http://stardog.clariah-
sdh.eculture.labs.vu.nl/databases/the_migration_portal/vocab/>
prefix mpr: <http://stardog.clariah-
sdh.eculture.labs.vu.nl/databases/the_migration_portal/resource/>
PREFIX sdmx: <http://purl.org/linked-data/sdmx/2009/code#>
prefix dbr: <http://dbpedia.org/resource/>
prefix dbo: <http://dbpedia.org/ontology/>
# Select Countries, with their labels, a time period and the values
of their indicators at that timeperiod.
# ValueP = Population Value
# ValueU = Unemployment Value (rate)
select distinct ?Country ?Label ?Time_Period ?ValueTotalP
?ValueForeignP ?ValueU ?From_Country ?To_Country
?InflowValue ?Inflow_Time_Period ?country
where {
graph mpr:population_eu_graph{
select ?country ?Time_Period ?ValueForeignP where{
?Population mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:population_value ?ValueForeignP ;
mpo:population_type "Foreign country"^^xsd:string .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country)}}
graph mpr:population_eu_graph{
select ?country ?Time_Period ?ValueTotalP where{
?Population mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:population_value ?ValueTotalP ;
mpo:population_type "Total"^^xsd:string .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country)
}}
# graph mpr:population_eu_graph{
# ?Population mpo:country ?Country ;
# mpo:time_period ?Time_Period ;
# mpo:population_value ?ValueTotalF ;
# mpo:population_type "Foreign country"^^xsd:string .
# }
graph mpr:unemployment_eu_graph{
?Unemployment_rate mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:unit "Active population"^^xsd:string ;
mpo:indicator_value ?ValueU .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country) }
optional{
graph mpr:inflow_graph{
?Inflow mpo:from_country dbr:Total ;
mpo:to_country ?To_Country;
mpo:movement_time_period ?Inflow_Time_Period ;
mpo:movement_value ?InflowValue .
filter(str(?To_Country) = ?country) .
filter(str(?Time_Period) = str(?Inflow_Time_Period))
}}
# Somewhere here we need to say : if ?Country = ?From_Country
make it "country of origin" and To_Country
order by (?Country) (?Time_Period)
```