Object and Data properties:

Another problem that we need to take care of is related to the object properties and data properties. In the two ontologies we have the same properties which are expressed by the same value, written in a different manner.

For instance, the birth date in one ontology is expressed in the format “1973-08-19”, while in the second ontology the date format is the following: “August 19, 1973”. A choice we made was to change the format of the second ontology values to the same as the first ontology in order to compare them. Thus, “August 19, 1973” will be changed into “1973-08-19”.

Another example of property we dealt with, is the gender which is referred in one ontology as “Male” / ”Female” and in the second one as ”M/F”. Using one measure of similarity between these two strings, for instance “Male” and “M” will not give us a very good score. These two strings have just one letter in common but in fact, are referring to the same thing. So, we change the “M” into “Male” and “F” into “Female” and check the equality of two strings.

For the property religion, we found in the two ontologies the same format for the values. The only difference is that, in the second ontology, there are some blank spaces injected between the strings. For instance, “Roman Catholicism”, “Agnosticism”, “Buddhism” can be found also in the following forms “ Roman C at holicism” , “Agn ostic i sm “, “B u d d h is m”. We deleted all the spaces we found. Therefore, after applying this change, we compare for equality between “RomanCatholicism” from the first ontology and “RomanCatholisicm” from the second ontology.

Considering the form of the government, it was quite difficult to find some similarities between the values from the first ontology and the values from the second ontology. Even by computing a similarity measures considering the letters of these words, we have very low values, as shown in the following table. Thus, we decided to ignore this property.

For Malaysia (item6575592185453392182):

|  |  |  |
| --- | --- | --- |
| **First Ontology** | **Second Ontology** | **Jaccard Similarity** |
| Constitutional monarchy | CGn1tgtYaionwfjmon!Mczy | 0.3939 |
| Federation | Fedevation | 0.8181 |
| Monarchy | I5U5rc0t | 0.1428 |
| Parliamentary system | Par8lampDtary xystUm | 0.6 |

For the currency property, we can find either the same currency in the both ontologies, “USD” – “USD”, either in the second ontology we can find some injected characters between the letters of the currency, “RUB” – “RueB”. Broadly, we obtain a jaccard similarity between two similar strings greater or equal to 0.6, as shown in the following table. For two currencies that are completely different, we obtain a very small similarity measure, such as in the last line of the following table.

|  |  |  |
| --- | --- | --- |
| **First Ontology** | **Second Ontology** | **Jaccard Similarity** |
| MYR | MYR | 1.0 |
| RUB | RUeB | 0.75 |
| NZD | N4ZfD | 0.6 |
| USD | PUSyD | 0.6 |
| *USD* | *N4ZfD* | *0.1428* |

For properties such as “acted\_by”, “directed\_by”, “spoken\_in” our approach is to compare the number of such properties for each individual. The problem with these properties is that the URI is referred to another object, and not to a value:

<IIMBTBOX:acted\_by rdf:resource="http://oaei.ontologymatching.org/2010/IIMBDATA/en/item1665401821179624399"/>.

Thus, we keep only the pairs of subjects from both ontologies which have the same number of such a property.

For the property name, we have to cases to study:

* There are some names which are the same in the both ontologies:

“Lesotho” 🡪 “Lesotho”

“Leytonstone” 🡪 “Leytonstone”

* The majority of name in the second ontologies are just the initials from the first ontology, followed by a dot:

“Luke Skywalker” 🡪 “L. S.”

“New York City” 🡪 “N. Y. C.”

In this second case, we’ll change the first form to the second one and compare them.

Article is a property which contains the description of the individual. In this case we will compute a similarity measure.

The “calling code” for both ontologies are in the same form. In order to have a sameAs link between two countries, this code should be the same. In this case, we check for equality: for Russia, it’s 7 in both ontologies; for South Africa, 27; etc.

There is a property called iso\_639\_1\_code. This property can be found only for two subjects, and its value is either “ru”, either “en”, which comes from Russian Language or English Language. We say that if we find this property and if its value it’s equal in both ontologies, the SameAs link it’s valid.

With regard to the characteristic size, the values are very different:

|  |  |  |
| --- | --- | --- |
| **Individual Name** | **First Ontology** | **Second Ontology** |
| Auckland | 1086.0 | 6886.0 |
| Wellington | 290.0 | 560.6 |
| vancouver\_british\_columbia | 114.67 | 514.27 |

For the amount and estimated budget used, we have also very big differences:

1.8E7 🡪 35.4E19

1.15E8 🡪 37.40E09

…..

…..

…..

To resume this chapter, there are some properties which are more precise and important than others. The choice we make while choosing these properties will make a real impact on the computation and results of precision and recall. As we mentioned before, properties such as birthdate, religion, gender are by far a better choice than the properties we decided to ignore, such as the form of government, size, amount.

Similarity measures:

Brouillon

Similarity measure is just a function which qualifies the similarity between two strings.

Different measures of similarity used and tested:

Library: textdistance

Jaccard

Word 2 vec

Threasholds:

0.6 for currency

1 for birthday religion gender

……

Results using different measures of similarity:

Done:

form\_of\_government

starring\_in

/date\_of\_birth

name

acted\_by

article

currency

gender

directed\_by

religion

calling\_code

iso\_639\_1\_code only 2: en, ru

Dialect: just one: if both subj contains at least one prop dialect, add to result

size

amount

estimated\_budget\_used

To do:

**filmed\_in**

**mainly\_spoken\_in**

**has\_capital**

**born\_in**

**created\_by**

**shot\_in**

**spoken\_in**

<https://itnext.io/string-similarity-the-basic-know-your-algorithms-guide-3de3d7346227>