**Semantic Web Technology project**

Chris Pool S2816539

Ruben de Jong S2616726

Leonardo Velozo S1668501

**Introduction**

Project C seemed the best option for our group. For our project we aim to create a program that automatically generates family trees using the data from DBpedia using the properties “*Spouse, Children (of), mother (of), father (of) etc.”* and if necessary information extracted from the corresponded Wikipedia text*.* Our goal is to enrich the knowledge of family relationships covered in Wikipedia, giving inversed properties like mother-son and uncle-nephew relationships and also inferring subproperty relationships like son-of is a subporperty of child-of. By doing this, people can see the different relative and family relationships that an individual family member has with other relatives. A specific family tree and its family relationships will be summarized in a graphical representation.

**Topic / Research question**

A software generated family tree from existing Wikipedia (DBpedia) data, including and adding entities (persons) that do not have a direct link made to each other and adding the different kind of relationships between these entities.

**Method**

We want to develop our system using Python as programming language combined with some libraries like SparqlWrapper to create the family trees. In order to show them we want to create a HTML file. For version control we are going to use GitHub. Another method we are aiming to use, if the information can’t be extracted from the Wikipedia info boxes, is that of Natural Language Processing. For this we will mainly use Natural Language Processing with Python (NLTK).

**Possible Extension**

* Recognizing entities from plain text or other sources than DBpedia.
* Adding Family tree connections to the Dutch Wikipedia database (Project A)

**Literature**

Our project will be supported mainly by 2 articles. The article of Procházka & Smrž will support our project with recognition and disambiguation of entities in unstructured text. The article of Wu, Hoffman and Weld performs a similar project extracting information from Wikipedia info boxes and will therefore be useful for our project.

[Procházka, A. E. J. H. J., & Smrž, O. Entity Recognition Based on the Co-occurrence Graph and Entity Probability. Praque, ERD’14 ACM, 2014.](http://web-ngram.research.microsoft.com/ERD2014/Docs/submissions/erd14_submission_3.pdf)

Wu, F., Hoffman, R., Weld, D.S. Information Extraction from Wikipedia: Moving Down the Long Tail. New York, ACM 2008.