To extract named entities and relations between them from text and to convert them into an RDF representation which is linked to DBpedia (linked data).

The formalism underlying this “Web of Linked Data” is the Resource Description Framework (RDF) which encodes structured information as a directed labelled graph. Hence, in order to publish information as Linked Data, an appropriate graph-based representation of it has to be defined and created. While this task is of minor difficulty and can be easily automatized if the original information is already structured (as, e.g., in databases), the creation of an adequate RDF representation for unstructured sources, particularly textual input, constitutes a challenging task and has not yet been solved to a satisfactory degree.

To employs robust techniques from natural language processing (NLP) including named entity recognition (NER), word sense disambiguation (WSD) and deep semantic analysis. The RDF output is embedded in the Linked Open Data (LOD) cloud by using vocabulary from DBpedia.

The idea behind the project is to adapt IE methods to detailed user information needs in a completely automated way, with the objective of creating very large domain-dependent and task-dependent knowledge bases.

Finally for all other cases, we will learn shallow patterns. As opposed to approaches based on complex machine learning algorithms (random walks), we will focus on lexical-syntactic shallow pattern generalization algorithms. The patterns will be generalised from the textual context of each and based on features such as words (lexical), part of speech (syntactic) and expected semantics such as related entity classes. We want to enrich the pattern representation with semantics mined from external knowledge resources, such as fine-grained entity labels. The patterns are then applied to other web pages to create new candidate annotations.

The patterns are then clustered according to their semantic similarity and transform this information to RDF.

How to extract triple? OLLIE? …
How to publish new triples to LOD?
How to recognize entities? needs Disambiguation
How to recognize relations? needs Disambiguation
How to define URIs for the predicate and relation types provided by
How to link recognized named entities?
How to extract date strings?
How to map extracted triples onto the DBpedia namespace? Machine learning? pattern matching?

How to do an ontology mapping module that carries out the final mapping of predicates from the DBPedia, OLLIE, or any other extractor to the DBpedia namespace?
What happens to entities that there aren't existed in DBpedia?
How to generalize mappings? How to learn mappings automatically? How to map to DBpedia ontology?

Limitation:
we don't want to generate new linked data and we are going to integrate existing linked data like DBpedia.

Input:
The Wikipedia articles as input.

Output:
Output from the IE task will be both a set of instances to publish on the DBpedia RDF, as well as a set of annotations which will provide provenance for the generated instances.

Tools:
OLLIE, ReVerb, Semantic Parser, Semantic role labeler ,Wikifier (Illinois Wikifier), Stanford Parser, Open NLP, POS tags, Stanford CoreNLP, Coreference resolution, Named Entity Recognizer ...

Bootstrapping set: Wikipedia, infobox, Dbpedia. Do you have any better idea?

Hot to evaluate?

- Manually (It is not prefered), 200 random sample sentences from Wiki pages are selected and verified by existing triples in DBPedia
- Evaluation by using instance matching , Generating RDF from New York Times article and linking it to DBpedia.
- Comparing with old version of

Try this link!
http://dbpedia.org/snorql/?query=SELECT+DISTINCT+%3Fp+%3Fo%0D%0AWHERE+%7B%0D%0A+%7B%3Fs+%3Chttp%3A%2F%2Fdbpedia.org%2Fontology%2FwikiPageID%3E+534366+.%0D%0A+%3Fs+%3Fp+%3Fo%7D%0D%0A%7D