This document structures the development of a strategy for (semi-)automatic construction of ontologies (so-called ontology learning) m,.l,into logical work packages.

**WP 1: Definition of starting point for ontology construction**

**Owner:** Tim

* Why are we doing this? Background Knowledge Graphs, why not sticking to Data Catalogues?
* Experiences gathered at Bayer; how can we extend the portfolio but at the same time overcome the challenges present in the project? -> Automatization
* Goal setting

**WP 2: Develop criteria catalogue for evaluating the success of automatic ontology construction**

**Owner:** Gul / Tim

* If we want to measure the success of automatic creation of ontologies, we need to compare the results to the benchmark; benchmark is a human being creating the ontologies
* Any combination of APIs and other means of support need to be measured against output from humans
* So what makes a good ontology? The answer to this question is reflected in the criteria; later on, we need to see how close our solution fulfills this criteria (criteria should not be mapped to NLP APIs ***only*** as we are also want to highlight the shortcomings and don’t want to confirm what is already there out-of-the-box)
* Main criteria is therefore how good the quality of the relations is and how good (a part of) the ontology is in performing such a task.[[1]](#footnote-1) This can be further broken down into:
  + reuse of existing vocabulary (that is already being used in other ontologies on the web)
  + high breadth and depth of entity and relationship variance (“nested” constructs, amount)
  + Correct identification of domain
  + Accuracy of naming for entities and relationships (with the human choice being the benchmark) -> this could later be checked with precision / recall KPIs
  + The criteria we have so far provide value when we compare the capabilities of the APIs to each other
  + Check <http://wiki.opensemanticframework.org/index.php/Ontology_Best_Practices> for best practice criteria in ontology generation

**WP 3: Choose datasets and at best, existing ontologies that can reflect them**

**Owner:** Gul

* news dataset could be used
* but we need an ontology for the events happening in there

**WP 4: Choose NLP tools to apply ontology learning**

**Owner:** Gul

* Spacy, StanfordCoreNLP, OpenCalais
* Document the findings according to capabilities of APIs, i.e. tokenization output etc.
* Generate code snippets, jupyter notebooks to replicate findings (create repository on github)

**WP 5: Validate results against criteria as defined in WP 2**

**Owner:** Gul

* Criteria evaluation can be partially automated in code
* Matching of generated ontology against benchmarking ontologies from WP 3
* Enrichment of benchmarking ontology from WP 3?

**WP 6: Derive recommendations based on findings of WP 5**

**Owner:** Gul / Tim

* Document limitations
* discuss possible workarounds
* iterate previous steps

1. http://people.kmi.open.ac.uk/marta/papers/FernandezAswc2009.pdf [↑](#footnote-ref-1)