# Tìpalo: a tool for automatic typing of DBpedia Entities

Andrea Giovanni Nuzzolese[1,2], Aldo Gangemi[1], Valentina Presutti[1], Francesco Draicchio[1], Alberto Musetti[1], and Paolo Ciancarini[1,2]

[1] STLab-ISTC Consiglio Nazionale delle Ricerche, Rome, Italy.
[2] Dipartimento di Scienze dell'Informazione, Università di Bologna, Italy.

**Abstract.** In this paper we demonstrate the potentiality of Tìpalo, a tool for automatically typing DBpedia entities. Tìpalo identifies the most appropriate types for an entity in DBpedia by interpreting its definition extracted from its corresponding Wikipedia abstract. Tìpalo relies on FRED, a tool for ontology learning from natural language text, and on a set of graph-pattern-based heuristics which work on the output returned by FRED in order to select the most appropriate types for a DBpedia entity. The tool returns a RDF graph composed of rdf:type, rdfs:subClassOf, owl:sameAs, and owl:equivalentTo statements providing typing information about the entity. Additionally the types are aligned to two lists of top-level concepts, i.e., Wordnet supersenses and a subset of DOLCE Ultra Lite classes. Tìpalo is available as a Web-based tool and exposes its API as HTTP REST services.

## 1 Introduction

DBpedia [6] and YAGO [8] are, de facto, the two reference ontologies for DBpedia resources. Unfortunately, a large number of DBpedia resources is still untyped, or has a very specialized type, and types are taken from ontologies that have heterogeneous granularities or assumptions (e.g., 272 infobox-based types in the DBpedia ontology (DBPO) against almost 290,000 category-based in YAGO). While it is reasonable to have limited semantic homogeneity on the Web, it is highly desirable to bring a more organized and complete typing to DBpedia entities. Knowing what a certain entity is (e.g., a person, organization, place, instrument, etc.) is key for enabling a number of desirable functionalities such as type coercion [5], data pattern extraction from links [6], entity summarization (cf. Google Knowledge Graph), automatic linking, etc. In this paper we demonstrate how Tìpalo and its Web API can be used for automatically typing DBpedia entities based on their natural language denitions as provided by their corresponding Wikipedia pages. Tìpalo relies on FRED [7], a tool that implements deep parsing methods based on frame semantics for deriving RDF and OWL representations of natural language sentences. On top of FRED, Tìpalo implements a set of graph-patterns-based heuristics that are used in order to extract the most appropriate types from the RDF representation derived from the natural language definition of a DBpedia entity. Additionally, Tìpalo gathers

the correct senses of the extracted types by exploiting a word-sense disambiguation (WSD) tool, i.e., UKB [1], which automatically links them to WordNet's synsets. Tìpalo finally provides foundational grounding to extracted types by using an alignment between WordNet and two top-level ontologies: WordNet super senses, and a subset of DUL+DnS Ultralite [3]. We refer interested readers to [4] for a detailed description of the algorithm and the evaluation of Tìpalo.

## 2 Related work

The DBpedia project [6] and YAGO [8] are the most relevant approaches at typing DBpedia entities. DBpedia provides an ontology extracted from Wikipedia infoboxes based on hand-generated mappings of infoboxes to the DBpedia ontology. The DBpedia ontology counts of 359 concepts (version 3.8) but only 2.3M entities over more than 4M are classified with respect to this ontology. YAGO types are extracted from Wikipedia categories and aligned to a subset of WordNet. The YAGO ontology is larger that the DBpedia one and counts of 290K concepts, with a larger, but still incomplete, coverage of DBpedia entities. Other relevant work related to our method includes Ontology Learning and Population (OL&P) techniques [2]. Typically OL&P is implemented on top of machine learning methods, hence it requires large corpora, sometimes manually annotated, in order to induce a set of probabilistic rules. Such rules are defined through a training phase that can take a long time. The method presented in this paper differs from existing approaches, by relying on a component named FRED [7], which implements a logical interpretation of natural language represented based on Discourse Representation Theory (DRT). FRED is fast and produces an OWL-based graph representation of an entity description, including a taxonomy of types. We parse FREDs output graph, and apply a set of heuristics, so that we can assign a set of types to an entity in a very efficient way.

## 3 Tìpalo

In this section we present an overview of the algorithm implemented by Tìpalo and a scenario for demonstrating how to concretely use Tìpalo. The purpose of this demo is to show the ability of Tìpalo in gathering the most appropriate types of Wikipedia entities emerging from the interpretation of natural language definitions available in Wikipedia abstracts.

***Our approach.*** Tìpalo is implemented as a pipeline of components and data sources as illustrated in figure 1. Each component in the pipeline implements a step of the computation: (i) extraction of an entity's definition from its corresponding Wikipedia abstract; (ii) natural language deep parsing provided by FRED whose output is a RDF representation of the entity definition; (iii) selection of candidate types by means of the application of graph-pattern-based

[3] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

heuristics to FRED's output; (iv) word-sense disambiguation of candidate types; and (v) type alignment to OntolWordNet [3], WordNet supersenses and to a subset of and DUL+DnS Ultralite.
More details about the pipeline and the evaluation of Tìpalo can be found in [4].

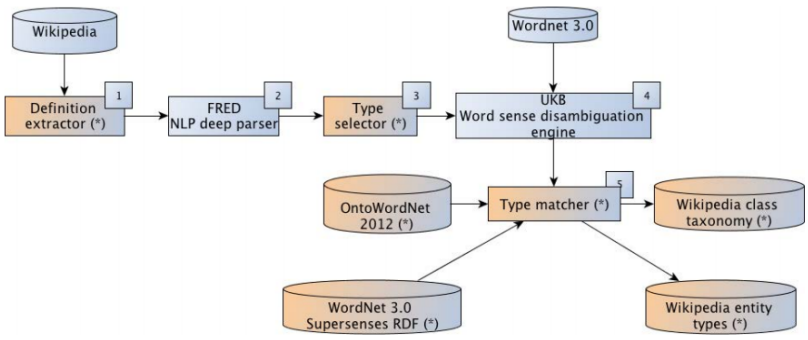


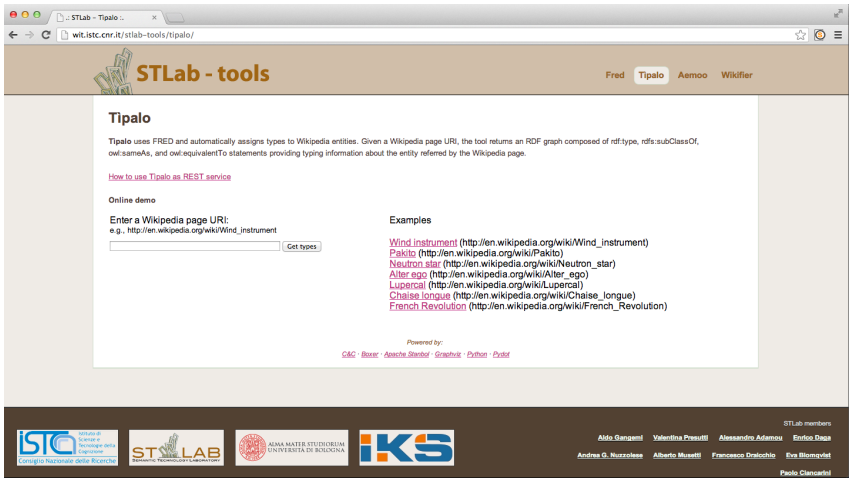

**Fig. 1.** The pipeline implemented by Tìpalo [4].

***Tìpalo at work.*** Tìpalo is available as a Web application. Figure 2(a) shows the on-line demo of Tìpalo [4]. Tpalo demo works only on DBpedia resources: its input is a Wikipedia page URL (see the text input in the right side of the on-line demo in figure 2(a)). On the right side of the interface, users can find a list of samples to start from.

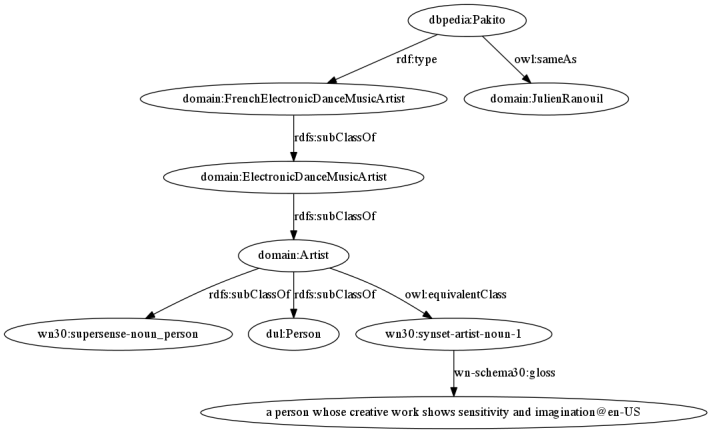
For demonstration purpose we show how to gather types for the entity `dbpedia:Pakito`. Once the Wikipedia URL of the entity has been provided in input to the system, Tìpalo returns (i) the natural language definition that has been used to infer the types of the entity by exploiting FRED and (ii) a RDF representation composed of rdf:type, rdfs:subClassOf, owl:sameAs, and owl:equivalentTo statements that provides typing information about the entity. In our example the definition that Tìpalo extracts from the Wikipedia abstract for `dbpedia:Pakito` is: "*Pakito is the alias of French Electronic Dance Music artist Julien Ranouil.*". As can be observed in figure 2(b) the entity `dbpedia:Pakito` is typed as `domain:FrenchElectoronicDanceMusicArtist` [5]. Tìpalo correctly recognizes the type `domain:FrenchElectoronicDanceMusicArtist` as subclass of the type `domain:ElectoronicDanceMusicArtist`, which is in turn

[4] The homepage of Tìpalo is available at `http://wit.istc.cnr.it/stlab-tools/tipalo`

[5] The prefix `domain` can be customized to any namespace desired by the user. The prefixes `dbpedia`, `dul`, and `wn30` stand for `http://dbpedia.org/resource`, `http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#`, and `http://www.w3.org/2006/03/wn/wn30/instances/` respectively.




(a) The home page of Tìpalo on-line at `http://wit.istc.cnr.it/stlab-tools/tipalo`.




(b) The output of Tìpalo for the resource `Pakito` derived from the natural language definition "*Pakito is the alias of French Electronic Dance Music artist Julien Ranouil.*"

recognized as subclass of `domain:Artist`. Tìpalo also disambiguate each extracted type through WSD. For this task Tìpalo uses UKB [1]. The IDs of WordNet synsets returned by UKB are resolved to corresponding OntoWordNet [3] URIs and are aligned to the extracted types by means of `owl:equivalentClass` axioms. In our example the type `domain:Artist` has been disambiguated with the OntoWordNet synset `wn30:synset-artist-noun-1`. This synset expresses the meaning "a person whose creative work shows sensitivity and imagination". Finally, the entity `dbpedia:Pakito` has been classified as `owl:sameAs`

`domain:JulienRanouil`.
Tpalo service is exposed as HTTP REST API [6].

## 4 Conclusions

We have presented Tìpalo, a tool that formalizes entity definitions extracted from Wikipedia for automatically typing DBpedia entities and linking them to other DBpedia resources, WordNet, and foundational ontologies. As ongoing work, we are deploying the tool on a more robust cluster for improving time performances and experimenting on a large-scale resource such as the whole Wikipedia. The medium-term goal is to incrementally build a Wikipedia ontology that reflects the richness of terminology expressed by natural language, crowd sourced definitions of entities.

## References


1. E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece, 2009. The Association for Computer Linguistics.
2. P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer, 2006.
3. A. Gangemi, R. Navigli, and P. Velardi. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer, 2003.
4. A. Gangemi, A. G. Nuzzolese, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini. Automatic Typing of DBpedia Entities. In *International Semantic Web Conference (1)*, volume 7649 of *Lecture Notes in Computer Science*, pages 65–81. Springer, 2012.
5. A. Kalyanpur, J. W. Murdock, J. Fan, and C. A. Welty. Leveraging community-built knowledge for type coercion in question answering. In *International Semantic Web Conference (2)*, volume 7032 of *Lecture Notes in Computer Science*, pages 144–156. Springer, 2011.
6. J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, 2009.
7. V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *EKAW: Knowledge Engineering and Knowledge Management that matters (to appear)*. Springer, 2012.
8. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.


[6] Details about how to use Tìpalo's API can be found at `http://stlab.istc.cnr.it/stlab/tipalo`