# N$^3$ - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format

**Michael Röder**[1,2,⋆], **Ricardo Usbeck**[1,2,⋆], **Sebastian Hellmann**[1], **Daniel Gerber**[1] **& Andreas Both**[1,2]

[1] Agile Knowledge Engineering and Semantic Web, University Leipzig, Germany
[2] R & D, Unister GmbH, Leipzig, Germany
michael.roeder@unister.de, usbeck@informatik.uni-leipzig.de

**Abstract**

Extracting Linked Data following the Semantic Web principle from unstructured sources has become a key challenge for scientific research. Named Entity Recognition and Disambiguation are two basic operations in this extraction process. One step towards the realization of the Semantic Web vision and the development of highly accurate tools is the availability of data for validating the quality of processes for Named Entity Recognition and Disambiguation as well as for algorithm tuning. This article presents three novel, manually curated and annotated corpora (N$^3$). All of them are based on a free license and stored in the NLP Interchange Format to leverage the Linked Data character of our datasets.

**Keywords:** Datasets, NLP Interchange Format, Named Entity Detection, Named Entity Disambiguation

## 1. Introduction

Automatically extracting and linking Named Entities (NEs) to a particular knowledge base (KB) from unstructured, natural language text is an extremely challenging task (Cucerzan, 2007). Leveraging Linked Data can help developing tools to automatically extract semantic data (Gerber et al., 2013; Hoffart et al., 2011; Mendes et al., 2011; Usbeck et al., 2014).

The tasks of Named Entity Recognition (NER) and Named Entity Disambiguation (NED) are part of the research area of Information Extraction (IE) (Ngonga Ngomo et al., 2011). NER is the task of identifying entities of certain types. NED is the task of disambiguating pre-identified named entities towards a certain KB. These IE steps depend on datasets which need human annotation, and therefore make it a time-consuming and expensive task. We publish three novel datasets (called N$^3$) in which named entities have been annotated manually. As KB for this annotation we used DBpedia (Auer et al., 2008) which is the central point of the Linked Open Data movement. These datasets have already been used to evaluate (Hoffart et al., 2011; Mendes et al., 2011) as well as in (Gerber et al., 2013; Usbeck et al., 2014). N$^3$ will be published using NLP Interchange Format (NIF) (Hellmann et al., 2013) ensuring a greater interoperability to overcome the need for corpus-specific parsers. The data can be downloaded from our project homepage[1].

Our main contributions are (1) the publication of three novel and freely available datasets for NER and NED, (2) an analysis of the underlying corpora, (3) and the transformation of these corpora to NIF providing provenance information. Finally, (4) our datasets also allow the analysis of coreference resolution (Singh et al., 2011) if the entity is not in the KB of the respective corpus.

## 2. State of the art

Recently, Steinmetz et al. (Steinmetz et al., 2013) published a statistical benchmark evaluation. Three datasets are described there of which two are freely available[2] in the NIF. The authors also analyze the aim of each dataset according to four baseline algorithms and respective underlying dictionaries for NED. Furthermore, the two available datasets, `KORE50` and `DBpedia Spotlight`, are published using NIF and have been part of the original benchmarks of (Hoffart et al., 2011; Mendes et al., 2011).

Unfortunately, both datasets do not inherit a clear license nor do they contain a large number of documents typically needed for optimization problems. `KORE50` comprises 50 sentences overall whereas `DBpedia Spotlight` consists of 58 sentences only. In contrast, our aim is to provide larger and more insightful datasets in NIF to leverage the possibility of optimizing NER and NED algorithms via Linked Data.

In the last couple of years, many more datasets for NER and NED evaluation have emerged. However, most of them are not freely available, e.g., the full CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) used in (Hoffart et al., 2011). Others are not yet annotated with Linked Data from DBpedia like the WePS (Web people search) evaluation dataset (Artiles et al., 2007).

## 3. Corpora

In the following section, we present the annotation process as well as specific features for each corpus of the N$^3$-Collection.

During the annotation, we focused on recognizing three main classes of NEs: persons, places and organizations. Each identified NE has been manually disambiguated to the DBpedia 3.9 if possible. In case there was no matching resource from a KB we created an URI[3] using the `http://`

[1] `http://aksw.org/Projects/N3nernednif`

[2] `http://www.yovisto.com/labs/ner-benchmarks/`

[3] `http://www.w3.org/TR/cooluris/`

| Corpus | Language | \|Documents\| | \|Words\| | Avg. Words/Doc. |
|---|---|---|---|---|
| News-100 | German | 100 | 48199 | 481.99 |
| Reuters-128 | English | 128 | 33413 | 261.04 |
| RSS-500 | English | 500 | 31640 | 63.28 |
| KORE50 | English | 50 | 1332 | 26.64 |
| DBpedia Spotlight | English | 10 | 3582 | 358.20 |

Table 1: Features of the corpora and their documents.

| Corpus | Entities | | Unique URIs | |
|---|---|---|---|---|
| | DBpedia | AKSW | DBpedia | AKSW |
| News-100 | 1547 | 108 | 315 | 57 |
| Reuters-128 | 650 | 230 | 299 | 145 |
| RSS-500 | 524 | 476 | 400 | 449 |
| KORE50 | 144 | 0 | 127 | 0 |
| DBpedia Spotlight | 331 | 0 | 249 | 0 |

Table 2: Number of single entities and unique URIs in the corpora.

aksw.org/notInWiki/ namespace (see Section 4.). Additionaly, we resolved coreferences for every named entity, especially, for entities that are not yet in the KB.

Furthermore, the collection of datasets is annotated by the version and language of the KB, hence any change of the underlying database can be analysed. In order to spread the corpora, we publish them under the Creative Commons BY-NC-SA license[4].

In general, our corpora contain more documents then any published NIF corpora (KORE50 and DBpedia Spotlight) so far. First order statistics for all datasets, such as the number of documents, words or average word count per document, can be found in Table 1.

The distribution of NEs over the texts is shown in Figure 1. RSS-500 has been left out because all of its documents comprise exactly two entities. As depicted in the diagram, all documents of the KORE50 corpus have less than six NEs. The documents from DBpedia Spotlight corpus reveal a larger context and thus more NEs. On the other side, DBpedia Spotlight comprises only 10 documents. The Reuters-128 corpus has most documents, although many of these documents are shorter and thus have less NEs.

### 3.1. News-100

This corpus comprises 100 German news articles from the online news platform news.de[5]. All of the articles were published in the year of 2010 and contain the word *Golf*. This word is a homonym that can have the following meanings:

- A gulf like the Gulf of Mexico or the Persian Gulf,
- the ball sport or
- a car model produced by the German manufacturer Volkswagen.

One researcher annotated the documents manually. Another researcher resolved occurring conflicts after supervising the corpus. Although the sport golf as well as the car are not within the class range of NER, they are kept for evaluation purposes.

### 3.2. Reuters-128

This English corpus is based on the well known Reuters-21578[6] corpus which contains economic news articles. In particular, we chose 128 articles containing at least one NE, as described in (Usbeck et al., 2014). Compared to the News-100 corpus the documents of Reuters-128 are significantly shorter and thus carry a smaller context, as can be seen in Table 1.

To create the annotation of NEs with URIs, we implemented a supporting judgement tool (see Figure 2)[7]. The input for the tool was a subset of more than 150 Reuters-21578 news articles sampled randomly. First, FOX (Ngonga Ngomo et al., 2011) was used for recognizing a first set of NEs. This reduced the amount of work to a feasible portion regarding the size of this dataset.

Afterwards, the domain experts corrected the mistakes of FOX manually using the annotation tool. Therefore, the tool highlighted the entities in the texts and added initial URI candidates via simple string matching algorithms. Two scientists determined the correct URI for each named entity manually with an initial voter agreement of 74%. This low initial agreement rate hints towards the difficulty of the disambiguation task.

In some cases judges did not agree initially, but came to an agreement shortly after reviewing the cases. While annotating, we left out ticker symbols of companies (e.g., *GOOG*

[4] http://creativecommons.org/licenses/by-nc-sa/4.0/

[5] http://www.news.de

[6] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[7] https://github.com/RicardoUsbeck/QRTool

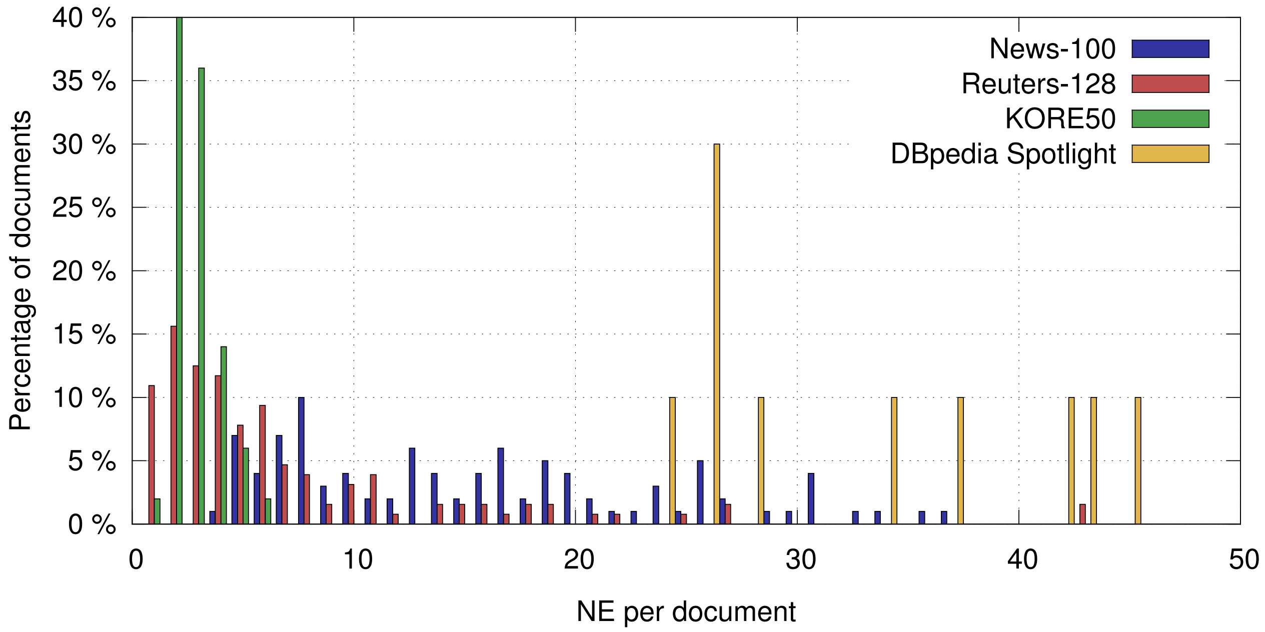
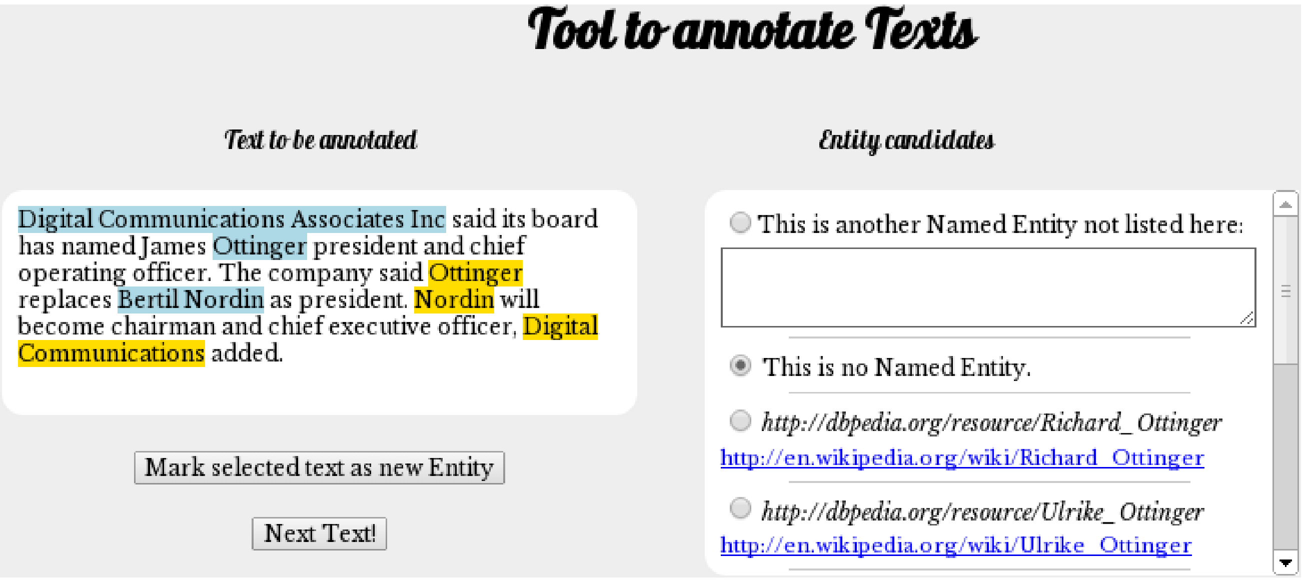


Figure 1: Distribution of NEs per document. We omitted two outliers from the News-100 corpus containing 55 and 63 NEs.



Tool to annotate Texts

Text to be annotated

Digital Communications Associates Inc said its board has named James Ottinger president and chief operating officer. The company said Ottinger replaces Bertil Nordin as president. Nordin will become chairman and chief executive officer, Digital Communications added.

Entity candidates

○ This is another Named Entity not listed here:

● This is no Named Entity.

○ *http://dbpedia.org/resource/Richard_Ottinger*
http://en.wikipedia.org/wiki/Richard_Ottinger

○ *http://dbpedia.org/resource/Ulrike_Ottinger*
http://en.wikipedia.org/wiki/Ulrike_Ottinger

Figure 2: GUI of our annotation tool.

for Google Inc.), abbreviations and job descriptions because those are always preceded by the full company name respectively a person's name.

### 3.3. RSS-500

This corpus has been created using a dataset comprising a list of 1,457 RSS feeds as compiled in (Goldhahn et al., 2012). The list includes all major worldwide newspapers and a wide range of topics, e.g., *World*, *U.S.*, *Business*, *Science* etc. The RSS list has been compiled using a 76-hour crawl, which resulted in a corpus of about 11.7 million sentences. A subset of this corpus has been created by randomly selecting 1% of the contained sentences.

Finally, one researcher annotated 500 randomly chosen sentences manually. These sentences were a subset of those which contained a natural language representation of a formal relation, like "..., who was born in..." for `dpo:birthPlace` (see (Gerber and Ngomo, 2012)). The relations had to occur more than 5 times in the 1% corpus. In case the mentioned entity is not contained in a new URI has been generated. This corpus has been used for evaluation purposes in (Gerber et al., 2013).

## 4. Using NIF for publishing corpora

For publishing our datasets we choose NIF because it is a Resource Description Framework (RDF)-based Linked Data serialization. NIF provides different advantages, e.g., *interoperability by standardization* (Hellmann et al., 2013) or *query-ability*. The *NIF-standard* assigns each document an URI as starting point and generates another Linked

```
@prefix nif:      <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix itsrdf:   <http://www.w3.org/2005/11/its/rdf#> .
@prefix xsd:      <http://www.w3.org/2001/XMLSchema#> .

<http://aksw.org/N3/Reuters-128/1#char=0,337>
      a       nif:String , nif:Context , nif:RFC5147String ;
      nif:beginIndex "0"^^xsd:int ;
      nif:endIndex "337"^^xsd:int ;
      nif:isString "Key Tronic corp said it has received contracts..."@en ;
      nif:sourceUrl <http://www.research.att.com/~lewis/Reuters-21578/15003> .

<http://aksw.org/N3/Reuters-128/1#char=0,15>
      a       nif:String , nif:RFC5147String ;
      nif:anchorOf "Key Tronic corp"^^xsd:string ;
      nif:beginIndex "0"^^xsd:int ;
      nif:endIndex "15"^^xsd:int ;
      nif:referenceContext  <http://aksw.org/N3/Reuters-128/1#char=0,337> ;
      itsrdf:taIdentRef <http://dbpedia.org/resource/Key_Tronic> ;
      itsrdf:taSource "DBpedia_en_3.9"^^xsd:string .
```

Listing 1: example of the resulting N3-triples.

```
Source = Reuters-21578
ID     = 15003
Text   = Key Tronic corp said it has
    received contracts...
```

Listing 2: example input text.

```
Select ?namedEntity {[] itsrdf:taIdentRef ?
    namedEntity }
```

Listing 3: SPARQL query to get all NEs.

Data resource per NE. Each document is a resource of type `nif:Context` and its content is the literal of its `nif:isString` predicate. Where possible, we added the source from which we got the document using the `nif:sourceUrl` predicate. Every NE is an own resource with a newly generated URI pointing to the original document via the `nif:referenceContext` predicate. Additionally the begin (`nif:beginIndex`) and end position (`nif:endIndex`) as well as the disambiguated URI (`itsrdf:taIdentRef`) and the respective KB (`itsrdf:taSource`) are stored. In contrast to (Steinmetz et al., 2013), mentioning the source of annotation serves as a more useful semantic background and is thus more valuable for further research. For instance, NIF document as depicted in Listing 1 has been transformed from the document from Listing 2.

The second advantage of using a corpus in NIF is that it is searchable using SPARQL. When the corpus is loaded in a Triple Store (e.g., Virtuoso[8]) one can easily find all NEs by posing a simple SPARQL query, as depicted in Listing 3.

## 5. Conclusion

Three different corpora are presented in this article. They can be used for named entity recognition and named entity disambiguation benchmarks as well as algorithm tuning. We compared these corpora with two already known datasets and showed the advantages of our datasets. The usability of these corpora for named entity disambiguation benchmark has been proven in (Usbeck et al., 2014; Gerber et al., 2013). Especially, we aim at providing a structured and standardized language resource for unstructured texts, enabling semantic querying. Our datasets link NLP algorithms to the Semantic Web by leveraging the power of NIF and Linked Data. In the future, the enlargement of the corpora and the improvement of their quality through the community are major issues that need to be worked on. Furthermore, converting more existing datasets to NIF and rebundle them in order to provide even more insightful NER and NED benchmarks is the next step of research in this area.

## 6. Acknowledgments

We thank Luise Erfurth and Didier Cherix for helping us creating annotations of the datasets and Jens Lehmann for his feedback. A special thanks goes to `news.de` for allowing us to use their articles. Parts of this work were supported by the ESF and the Free State of Saxony.

## 7. References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 64–69. Association for Computational Linguistics.

Sören Auer, Chris Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2008. DBpedia: A

[8] `http://sourceforge.net/projects/virtuoso/`

nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716.

Daniel Gerber and Axel-Cyrille Ngonga Ngomo. 2012. Extracting Multilingual Natural-Language Patterns for RDF Predicates. In *EKAW*, Lecture Notes in Computer Science. Springer.

Daniel Gerber, Axel-Cyrille Ngonga Ngomo, Sebastian Hellmann, Tommaso Soru, Lorenz Bühmann, and Ricardo Usbeck. 2013. Real-time RDF Extraction from Unstructured Data Streams. In *Proceedings of ISWC*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pages 782–792.

Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*.

Axel-Cyrille Ngonga Ngomo, Norman Heino, Klaus Lyko, René Speck, and Martin Kaltenböck. 2011. Scms–semantifying content management systems. In *The Semantic Web–ISWC 2011*, pages 189–204. Springer.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 793–803. Association for Computational Linguistics.

Nadine Steinmetz, Magnus Knuth, and Harald Sack. 2013. Statistical analyses of named entity disambiguation benchmarks. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October. CEUR Workshop Proceedings.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Sören Auer, Daniel Gerber, and Andreas Both. 2014. Agdistis - agnostic disambiguation of named entities using linked open data. In *Submitted to 11th edition of Extended Semantic Web conference,May 25th to May 29th, 2014 in Anissaras, Crete, Greece*.