Guidelines for Full Text Annotations in the SoNAR (IDH) Corpus

Sina Menzel, Josefine Zinck, Hannes Schnaitter & Vivien Petras

•

Humboldt-Universität zu Berlin
Berlin School of Library and Information Science

10.5281/zenodo.5115933

Abstract:

This document presents guidelines for the manual annotations made on a representative sample of the full text corpus in the SoNAR (IDH) project (www.sonar.fh-potsdam.de).

Contents

1. Introduction

The following guidelines present the detailed ruleset and specifications for the manual annotation of named entities (NE) within a representative sample of the full text corpus in the SoNAR (IDH) project. "Annotation" means the manual tagging of appearances of predetermined semantic units within a text as well as the linking of these tags to a given knowledge base. Annotation is therefore an enrichment of full texts with metadata. Section 2 introduces the corresponding full text corpus, section 3 and 4 specify the environment and quality management for the annotation process.

The semantic units of interest for the SoNAR (IDH) full text corpus annotation are named entities of the following classes:

Persons (PER),
Organizations (ORG),
Locations (LOC),
Conferences (CONF),
Events (EVT),
Works and expressions (WORK).

Section 5 defines these classes and the corresponding annotation rules in detail, and also lists exceptions from these rules.

The ruleset of the present guidelines is being developed iteratively along with the ongoing annotation (please see version no. above). The purpose of the guidelines is to secure consistency and coherence in the annotation process, in order to achieve optimal quality of the annotation's outcome: The gold standard that supports the evaluation of an automated process of named entity recognition (NER) in the realm of the project. The guidelines build upon former work (Fort et al. 2009; Rosset et al. 2011; Reznicek 2013; Reiter 2017) as well as the German Integrated Authority File (GND) hosted by the German National Library (DNB). The latter will be one of the knowledge bases used for named entity linking (NEL) and therefore serves as orientation for ambiguous cases.

In a broader sense, annotation includes adjustments of the original text, such as character correction as described in section 7. Section 8 describes rules for the annotation of named entity linking (NEL).

2. Data set

The complete dataset of full texts in the SoNAR (IDH) annotation process consists of 2,078,127 historical German text documents derived from the *Zeitungsinformationssystem* repository (ZEFYS). The documents are newspaper pages from the following periodicals (late 19th and early 20th century, see table 1).

Title	Time span	# of documents	Shares in %
Berliner Börsenzeitung	1872-1931	642,480	30.92
Berliner Tageblatt	1877-1939	489,983	23.58
Berliner Volkszeitung	1890-1930	142,403	6.85
Deutsches Nachrichtenbüro	1936-1940	7,429	0.36
Neueste Mittheilungen	1882-1894	1,322	0.06
Norddeutsche Allgemeine Zeitung	1878-1918	120,362	5.79
Provinzial-Correspondenz	1863-1884	1,087	0.05
Teltower Kreisblatt	1856-1896	25,819	1.24
Vossische Zeitung	1857-1917	647,242	31.15

Table 1: Newspapers in the data set for SoNAR (IDH) full text annotations.

From this data set, a representative subset is derived, which is manually annotated over the course of the project.

3. Annotation environment

The annotations are made by a single human annotator using the project's browser based in-house-tool named *neat*[1] (named entity annotation tool in html). *Neat* is adjusted iteratively along the annotation process to any necessities that might occur due to specifics of the textual content.

4. Quality management

The quality of the annotation is secured by the present guidelines as well as sample checks of the annotated texts by the co-annotation of sample documents which allows for taking agreement measures. The latter is expected to bring forward disagreement cases that show loopholes in the guidelines. Additionally, we introduced the "TODO"-tag in *neat* which may be used for ambiguous or uncertain tokens in order to support discussion and clarification on the guidelines in regular meetings of the annotator and other project associates. After each completion of an annotated text document, a revision session by the annotator is required.

5. Annotation of Named Entity Recognition (NER)

[1] https://github.com/qurator-spk/neat

"For […] efficient NE annotation […], it is important to focus, not on how *to annotate, but rather on* what *to annotate […]."* – Fort et al. 2009, p. 147.

The following section defines characteristics of named entities as well as the different semantic entity classes considered in the annotation process. More examples, as well as exceptions and special cases can be found in appendix A.

5.1 General annotation rules

The following rules are partly extracted from Reznicek 2013, p. 2ff.

1. The value of precision is favored over recall in the annotation process. For this reason, ambiguous cases are not marked as named entities, but with the label "TODO" for discussion in the annotation meetings. Should a suspected NE not be decodable by the annotator (e.g. due to the historical origin of the corpus), it is not to be annotated.
   Example: A suspected organization which is unknown to the annotation team and not included in the GND nor on Wikipedia.

2. Named entities occur as proper nouns, full nominal phrases, as well as derivations and abbreviations of the former.
   Example: Die [[erfurter]LOCemb Innenstadt]LOC

3. Pronouns are not to be marked as named entities. See appendix B for exceptions.

4. Determiner (e.g. articles) are not part of named entities. See appendix B for exceptions.
   Example: Der [Polizeipräsident]PER

5. Named entities may include at least one and up to x tokens.

6. If named entities occur in the plural, they are to be treated the same way as in the singular.

7. Named entities might occur as part of a token, e.g. genitive case. In these cases, the entire token is to be labeled with the corresponding type of entity.
   Example: [Frankreichs]LOC Käsevielfalt
   [Kreuzberger]LOC Nächte sind lang
   [Lisas]PER Geburtstag

   > Note: An NE as part of a token is NOT the same as an embedded NE!
   > In the first case, the other parts of the token are no entities.
   > Nevertheless, the entire token is to be annotated in cases of embedded entities too.

8. Named entities may be embedded in other named entities (second level NE). This might also occur in compounds, if more than one component is a separate entity.
   Example: Die [[Heinrich Böll]PERemb-Stiftung]ORG
   Die [[SPD]ORGemb-Abgeordnete]PER

9. If one entity marks the entire (group of) token(s) while the other entity marks only parts of it/them or derives from it/them, the latter is the second level entity. If the order of levels is not clear, the annotator may choose, which class to mark on first and which on second level.
   Example: [Stonehenge]WORK/LOC
   (annotator decides which is embedded)

10. If more than one named entity is embedded in another named entity, the annotator chooses which entity is to be marked on second level by evaluating the nesting levels: Subject/object of the sentence is the first level entity, while its direct attribute is the second level entity. The third level component is to be left out.
    Example: Das [Attentat auf das [französische Königspaar]PERemb]EVT
    ([französische]LOC is to be left out in this case, because it refers to the second level entity PERemb)

incorrect

TOKEN	NE-TAG	NE-EMB
Attentat	B-EVT	O
auf	I-EVT	O
das	I-EVT	O
französische	I-EVT	B-LOC
Königspaar	I-EVT	B-PER

correct

TOKEN	NE-TAG	NE-EMB
Attentat	B-EVT	O
auf	I-EVT	O
das	I-EVT	O
französische	I-EVT	B-PER
Königspaar	I-EVT	I-PER

11. Enumerations of related entities are to be annotated separately. This also applies if one token does not represent the entire entity.
    Example: [Ost-]LOC und [Westdeutschland]LOC

12. Named entities may occur within metonymic references. In these cases, the referenced entity is to be annotated on the first level. If the referring token(s) may also mark a named entity, in which case they are annotated on the second level (embedded).
    Example: Der [[Kreml]ORG]LOCemb hat entschieden.

13. Named entities embedded in entities are only to be marked if unique or unambiguous. If a named entity is later referred to in an ambiguous way it is not to be annotated.
    Example: Die [[American]LOCemb Football Association]ORG
    Der [Dreibund]ORG […]. Der Bund […] (Bund is not to be annotated for it is ambiguous on its own)

14. Any exceptions to the aforementioned rules must be agreed upon by the annotation team and will be listed in appendix A or B.

15. After the completion of the annotation of a document, there is a mandatory revision of the same document by the annotator in order to secure the best possible gold standard.

5.2 Person (PER)

The following rules build upon Rosset et al. 2011, p. 21 and Reznicek 2013, p. 6. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 9, 10). See appendix A for a complete list of subclasses.

1. Named entities referring to definite individuals may be classified as “person” with the label “PER”.

2. The label may also be given to tokens referring to families.
   Example: Die Intrigen der [Borgia]PER

Note: Bands are considered organizations; see rule no. 3 in section 5.3.

3. Descriptors referring to unambiguous, exclusive family connections with information on the person they are referring to are to be annotated as person entities. This includes the temporal context of the source, e.g. the point in time a newspaper was published.

   Examples: [Max Mustermann]PER verhielt sich genau wie [Max Mustermanns Vater]PER.
   [Max Mustermann]PER verhielt sich genau wie sein Bruder.
   (in this case, the connection is not exclusive)
   [Max Mustermann]PER verhielt sich genau wie [seine Frau]PER.
   [Max Mustermann]PER verhielt sich genau wie seine Ex-Frau.
   (in this case, the connection is not exclusive)
   Exclusive family descriptions are also to be annotated, if the person they are referring to is not mentioned in the same sentence, but in the same section of the text.

   Example: [Jonas]PER ging von der Schule nach Hause. Er schaute auf sein Handy. So sah er nicht, dass [sein Vater]PER ihm von der anderen Straßenseite zuwinkte.

4. Populations are not to be marked as persons.
   Example: [Amerikaner]LOC
   [Sowjets]LOC
   Jüdische Gemeinden

5. The PER-class may refer to first names, middle names, family names, nicknames, fictional characters, pseudonyms. Nicknames do not have to be unique, but unambiguous.

6. Titles (academic titles, titles as Ms./Mrs. or Mr., as well as titles of nobility as Sir, Madame, Duke, Duchess, prince, princess, military titles as General or Lieutenant and the like) are not part of named entities. See appendix B for exceptions.

7. Job titles and functions are not to be annotated unless they describe unambiguous, exclusive positions of a definite individual in the context of the text section they appear. Descriptions of jobs (e.g. “Her Royal Highness”, “His Excellence”, “Her Majesty”) are to be treated the same way.
   Example: Die [Bundeskanzlerin]PER
   Der [englische Botschafter]PER
   (in the latter case, the article referred to the country of Turkey, which makes the position unambiguous)

If the job title and name of the person appear together, tokens in between are to be included in the entity.

Example: Die [Bundeskanzlerin Frau Dr. Angela Merkel]PER
(underlined tokens are not considered entities if they appear separately or at the margin of other entities, see rule 6 in this section)
Der [[russische]LOCemb Ministerpräsident Stolyvin]PER

Deputy and vice positions are to be annotated if unambiguous.

Example: Der stellvertretende Leiter der Filiale.
(ambiguous, there may be more than one deputy position)
Der [3. stellvertretende Leiter der Filiale]PER

8. Definite descriptions may be marked as persons in cases of referential unicity. Unclear cases are to be marked with the tag "TODO" for discussion in the annotation team.

Example: [The Iron Lady]PER
[The Rock]PER

5.3 Organization (ORG)

The following rules build upon Reznicek 2013, p. 6-7 and Rosset et al. 2011, p. 29ff. For orientation in ambiguous cases: Guidelines of the GND and RDA Toolkit (Section 11). See appendix A for a complete list of subclasses of this class.

1. Named entities that refer to bodies, companies and the like are to be classified as "organization" with the corresponding label "ORG".

2. ORG-entities may occur as acronyms or nominalizations.
   Example: Die [NATO]ORG
   Der [ADAC]ORG
3. Bands and similar professional collectives are not considered persons, but organizations.
   Example: Das Konzert der [Queens of the Stoneage]ORG

4. Generic descriptions in front of an ORG-entity are not to be annotated.
   Example: Firma [[G.H. Friedländer]ORG]PERemb
   Fleischereifachgeschäft [Wurstbasar]ORG

5. Religious attitudes itself are not considered organizations. Connected religious organizations are to be considered organizations only if they are unambiguous (e.g. connected to a location, see 5.4). For exceptions see appendix B.
   Examples: Freue, freue dich, o Christenheit! (no entities)
   Muslimische, jüdische und christliche Vereinsmitglieder. (no entities)
   Die [evangelische St. Lucas Gemeinde in [Pattensen]LOCemb]ORG

5.4 Location (LOC)

The following rules build upon Reznicek 2013, p. 7. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 16). See appendix A for a complete list of subclasses of this class.

1. Named entities referring to "politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)" (MUC-6 task definition 1995) are to be labeled with "LOC". Locations might be fictional.
   Example: Die Tür nach [Narnia]LOC

2. If more than one location is described in one token, one of them is to be marked as an embedded entity. See no. 9 under 5.1 for first and second level disambiguation. See Appendix B for any exceptions.
   Example: Die [[spanisch]LOC-deutsche]LOCemb Frau
   Das [[Deutsche]LOCemb Reich]LOC

3. Definite descriptions may be marked as locations in cases of referential unicity. Unclear cases are to be marked with the tag "TODO" for discussion in the annotation meetings.

   Example: [The Big Apple]LOC

4. Locations embedded in descriptors of populations are to be marked on first level for populations are not considered named entities (see also rule no. 4 under 5.2).
   Example: Die [Amerikaner]LOC
   [Sowjets]LOC

5.5 Conference (CONF)

For orientation in ambiguous cases: Guidelines of the GND. See appendix A for a complete list of subclasses and examples.

1. Named entities referring to uniquely named gatherings of individuals on a certain pre-defined scientific topic, goal or shared purpose as well as a pre-defined ending point are to be classified as "conferences" with the label "CONF".

2. CONF-entities may occur as acronyms or nominalizations.
   Example: Der diesjährige [CLEF-Task]CONF
   Die [DHd]CONF

3. If a CONF-entity holds a time tag, the latter is to be marked as part of the entity.
   Example: Der [Bibliothekartag 2018]CONF

5.6 Event (EVT)

1. Named entities referring to uniquely identifiable events apart from conferences are to be tagged with the label "EVT". This class is annotated for experimental purposes and therefore does not follow a strict definition.

2. In contrast to conferences, events may be of spontaneous nature.

3. Topics of events may vary (e.g. military, political, cultural…).

5.7 Works and expressions (WORK)

The following rules build upon Rosset et al. 2011, p. 39ff. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 6). Please make sure to check the token in question in the current version of the GND catalogue. See appendix A for a complete list of subclasses.

1. Named entities referring to titled human creations are to be classified as works or expressions. The corresponding label is “WORK”.

2. Separate parts of the works, such as acts in plays are to be annotated as named entities.
   Example: Der [zweite Akt von [Romeo und Julia]WORKemb]WORK.
   Die [Arie aus [Elias]WORKemb]WORK

6. Excluded full text sections

There is no exclusion of any sections in the full texts, the documents are to be completely annotated.

7. Full text correction

Since the annotation sample is based on print originals, the digitization process required the automated recognition of optical characters (OCR) within the scanned documents. Using current software solutions, this process still comes with an inevitable error rate (Kugler 2018, p. 42) which might affect the recognition of named entities by the human annotator and certainly affects the recognition of named entities by current learning algorithms for automated NER (Kettunen/ Ruokolainen 2017).

The following types of errors might occur (based on Zumstein/Baierer 2016, p. 74-75):

I. Character errors
   This is the most frequent and most relevant type of error in the annotation process. It includes mistakes in the recognition of characters.
II. Segmentation errors
   These errors are a special type of character error, where spaces between tokens are not recognized correctly. This leads to the incorrect splitting or merging of tokens.
III. Word errors
   Word errors are character errors of full words. This frequently occurs in correlation with shifting fonts or if automated post-OCR-normalizations apply. The latter are usually based on wordlists that might disimprove individual tokens.
IV. Sectional errors
   This type refers to formatting errors regarding the layout or other textual sections, e.g. sentence boundaries.

Corrections on the SoNAR annotation sample concentrate on error types I, II, and IV. They exclusively concern errors occurring in named entities. There is no OCR-correction of the entire full text! For this purpose, *neat* supports changes in the character strings as well as deleting, merging and splitting of tokens.

7.1 Marking of sentences

Since the data format in *neat* is based on the format used in the GermEval2014 Named Entity Recognition Shared Task, sentence boundaries are indicated by an empty line (position 0, see User Guide). For this reason, error type IV. is being corrected in the annotation process only if it concerns sentence boundaries or interferes with the correct annotation of named entities.

1. Colons do not mark the beginnings of sentences.

2. OCR-errors not correlating to a token as well as missing words (or parts of sentences) are marked as a new sentence.

7.2 OCR-correction

1. If a token is predicted to have an error, but the corresponding word is not recognizable neither by OCR results nor by the original scan, the token is not to be corrected, but to be annotated

2. There is no correction of orthography due to the historical context of the sample. The adjustment of a token's characters therefore has to follow the printed original on the scan, even if the spelling does not align with current orthography. This also applies to suspected spelling and printing mistakes within the original (ger.: Aufnahme nach Vorlageform). Ambiguous cases (spelling vs. OCR) are to be discussed by the annotation team, possible exceptions will be captured in the guidelines. This also applies to punctuation characters (e.g. "⸗" instead of "-" to mark compounds).

   Exception: Hyphenations of named entities over two lines in the original are to be counted as sectional errors. This also applies to composita that are divided into two lines in the original scan.

   Examples:

   incorrect

TOKEN
Herr
Gam-
betta

   correct

TOKEN
Herr
Gambetta

   incorrect

TOKEN
Vormittags-
Besuch

   correct

TOKEN
Vormittags-Besuch

3. Some newspapers in the corpus in gothic type do not distinguish between capital I and capital J. In these cases, the OCR interpretation is considered correct, since verification through checking the snippet is impossible.

4. Completely missing words due to OCR errors are to be manually reconstructed using the snippet, if the missing word(s) is/are recognized to be a named entity in the original scan.

5. Punctuation characters are to be counted as separate tokens each.
   Example:

TOKEN
dem
„
Jüngeren
“

   Exception: Punctuation characters as parts of abbreviations (e.g. “St.”) and numberings (e.g. “4.” for “fourth”) are part of the token and therefore not to be counted separately.

   Example:

TOKEN
Donnerstag
,
1.
Januar
.
Berliner
Tageblatt
.
Nr.
1
.
Seite
3
.

6. Should an entity be surrounded by punctuation characters, the latter are not to be included in the annotation of the entity.
   Example: **incorrect** **correct**

TOKEN	NE-TAG
Operette	O
"	B-WORK
Die	I-WORK
Wächter	I-WORK
der	I-WORK
Moral	I-WORK
"	I-WORK

TOKEN	NE-TAG
Operette	O
"	O
Die	B-WORK
Wächter	I-WORK
der	I-WORK
Moral	I-WORK
"	O

7. Should one or more punctuation characters be embedded between two or more tokens that mark a single entity, they are to be included in the annotation of the entity.
   Example: **incorrect** **correct**

TOKEN	NE-TAG
des	O
"	O
kleinen	B-PER
"	I-PER
Wilson	I-PER

TOKEN	NE-TAG
des	O
"	O
kleinen	B-PER
"	O
Wilson	I-PER

8. Extended dashes are to be corrected to a single dash in case of entity compounds (“–“ “-“).

9. If a sentence starts with a punctuation character, the first is to be considered a separate sentence.
   Exception: Quotation marks (“”)
   Example: **incorrect** **correct**

POSITION	TOKEN
0	
1	—
2	Das
3	morgen
4	erſchei

POSITION	TOKEN
0	
1	—
0	
1	Das
2	morgen
3	erſchei

10. Special characters are to be taken into account, if they are part of the basic Latin or extended German alphabet (“Ü”,”ü”,”Ö”,”ö”,”Ä”,”ä”,”ß”). Additionally, the following accents are to be taken into account: aigu (é), grave (è), circonflex (ê), as well as historical characters[ſ,ꝛ]

8. Annotation of Named Entity Linking (NEL)

8.1 Prerequisites

After completing the annotation of named entity appearances in full texts, their automated linking to records in knowledge bases can be annotated. In SoNAR (IDH), we link to Wikidata. This section describes rules for decision making in the annotation of links. As Ling et al. (2015) mark, there are very little prior attempts to define clear NEL annotation guidelines. After the preprocessing,

- links appear as an ID to Wikidata;
- only first level entities are being linked, unless the linking of the first level is not possible, then second level is preferred over no linking;
- all corresponding tokens of an entity correspond to the same link.

8.2 General annotation rules for entity linking

The decision, whether a link is correct or not builds on these general rules:

1. The step of NEL annotation concentrates on the linking, there is no further revision of the NER annotation outcome.
2. Universal links to specific entities may be considered correct.

   Example: [Berlin Kreuzberg]LOC (Link: Q64 refers to the city of Berlin in general) ist ein Stadtteil.

3. Specific links to universal entities may be considered incorrect.

   Example: [Berlin]LOC (Link: Q308928 refers to Berlin Kreuzberg) ist eine der größten Städte [Europas]LOC (Link: Q183 refers to Germany).

4. Alternative links to metonyms are being treated as universal links and are therefore considered correct (see rule no. 5 under 8.2).

   Example: Der [[Kreml]ORG]LOCemb hat entschieden (Link: Q133274 refers to the fortified complex in Moscow, not the group of people in the russian government).

5. Links may not refer to a different entity type than the entity type assigned in the NER annotation process.

6. As an addition to rule no. 4, a link may be considered correct if it refers to an entity type assigned on the second level.

8.3 Linking annotation for person entities (PER)

1. For person entities (PER), precision is favored over recall, meaning the link shall be as specific as possible. This includes the context of the full text, e.g. the publishing date.

2. In the process of manual disambiguation of person entities (PER), a Link is considered correct, if

   a. the name (variant) of the Wikidata record matches the **entity name** in the text

      AND

   b. the record contains **temporal information** (e.g. Schaffenszeitraum, Geburtsdatum), which matches the context of the text (e.g. date of publication of the corresponding newspaper).

      AND

   c. In addition to these two aspects, there is at least **one more information** in the Wikidata record, which is considered **decisive** in the context of the entity occurence.

      AND

   d. There is no information in the Wikidata record, which contradicts the context of the entity occurence by giving inconsistent further details on the described entity.

      Example: A person named “Hans Müller” is described in the newpaper text as “living in Berlin”. In the Wikidata link, name and temporal information are consistent with the context of the newspaper text, but the residence location described in the Wikidata record is “Paris” throughout the temporal lifespan.

8.4 Linking annotation for other entities

1. For all other entities {ORG, LOC, CONF, EVT, WORK}, recall might be favored over precision in cases that fall under rule no. 3 section 8.2. As long as the annotator identifies a link which is unambiguously connected to the entity, a specific link might also be accepted as correct. This applies, for example, to cases in which a successor or predecessor is linked.

Example: “Der [Kaufhof]ORG bei uns wird geschlossen.” (Link: Q80220059 refers to the predecessor “Galeria Karstadt Kaufhof” and is still considered correct, even if the merger had not taken place at the time the corresponding newspaper article was published.)

2. For combined location entities (LOC) within one token, the corresponding link must refer to the combined location, not one of its separate components.

   Example: “Die [[französisch-deutsche]LOC]LOCemb Freundschaft.” (Since there is no link for the combined locations, seperate links to “Deutschland” (Q183) or “Frankreich” (Q142) are both considered incorrect.)

   Exception: “Die [[österreichisch-ungarische]LOC]LOCemb Herrschaft.” (The link Q28513 refers to “Österreich-Ungarn” as a combined location and would therefore be considered correct.)

References

Zumstein, Philipp; Baierer, Konstantin (2016): Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. In: 027.7 Zeitschrift für Bibliothekskultur 4 (2). DOI: 10.12685/027.7-4-2-155

Fort, Karën; Ehrmann, Maud; Nazarenko, Adeline (2009): Towards a Methodology for Named Entities Annotation. In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), p. 142-145. Available at: https://www.aclweb.org/anthology/W09-3025.pdf

Kettunen, K.; Ruokolainen, T. (2017): Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017. Göttingen, Germany: ACM Press, S. 181–186. Available at: http://dl.acm.org/citation.cfm?doid=3078081.3078084

Kugler, Anna (2018): Automatisierte Volltexterschließung von Retrodigitalisaten am Beispiel historischer Zeitungen. 33-54 Seiten / Perspektive Bibliothek, Bd. 7, Nr. 1 (2018). DOI: 10.11588/PB.2018.1.48394.

Ling, Xiao; Singh, Sameer; Weld, Daniel S. (2015): "Design challenges for entity linking." *Transactions of the Association for Computational Linguistics* 3: 315-328.

Rosset, Sophie; Grouin, Cyril; Zweigenbaum, Pierre (2011): Entités Nommées Structurées : guide d'annotation Quaero (Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum), Technical report. Available at: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Rosset, Sophie; Grouin, Cyril; Fort, Karën; Galibert, Oliver; Kahn, Juliette; Zweigenbaum, Pierre (2012): Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In: Proceedings of the Sixth Linguistic Annotation Workshop, p. 40-48. Available at: https://www.aclweb.org/anthology/W12-3606.pdf

Reiter, Nils (2017): How to Develop Annotation Guidelines. Blog post. Available at: https://sharedtasksinthedh.github.io/2017/10/01/howto-annotation/

Reznicek, Marc (2013): Linguistische Annotation von Nichtstandardvarietäten —Guidelines und „Best Practices". Guidelines NER. Version 1.5. Available at:

https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5

Appendix

This is an extended and modified list based on the NoSta-D-TagSet (Rezincek et al. 2013, p. 6ff.).

Appendix A.1: Example tag-set PER

Entity class	Subclass	Example	Exceptions
	Berufliche Funktionen (exklusiv)	Finanzminister Kaiserin Leiter der Abteilung für die Butterindustrie (ggf. mit anschl. LOC) Seine königliche Hoheit	~~Staatssekretär~~ ~~Stellv. Staatssekretär~~
	Dynastie, Geschlecht, Fürstenhaus	Borgia Habsburger von Lippe-Biesterfeld	keine ORG
	Familienname	Feuerstein Winkler-Eversberg	
	Fiktiver oder religiöser Charakter	Buddha Harry Potter Heiliger Antonius Miss Piggy	
	Künstlername/Pseudonym	Felix Brummer Marilyn Monroe P!nk	
	Spitzname/Nickname	chatbotchatter123 honeylove86 Müller “Der Jüngere” Naddel	Der „Kleine“ Schatz
	Vorname/Mittelname	Emil Hannelore Nina Winfried	

Appendix A.2: Example tag-set ORG

Entity class	Subclass	Example	Exceptions
Organization (ORG)			~~Bildungseinrichtung~~ ~~nicht explizit: z.B. „die Schule“, „das Waisenhaus“~~ ~~Bundesgenossen~~ ~~Delegationen~~ ~~Der Feind~~ ~~Die Christen~~

			~~Die Großmächte~~ ~~Die kleinen Völker~~ ~~Die Verbündeten~~ ~~Schiffe~~ Adelshäuser = PER
	Bands, Musikgruppen, Orchester	The Beatles	
	Bildungseinrichtung (explizit)	Freie Universität Berlin	
	Institut	Dt. Inst. f. Menschenrechte	
	Kaufhäuser (unique)	Kaufhaus des Westens	Keine LOC!
	Krankenhäuser (unique)	Elisabeth-Krankenhaus (embedded LOC)	
	Kultureinrichtung (nicht explizit)	Cinemaxx SPK	Explizite Einrichtungen = LOC (Bsp. Pergamonmuseum)
	Militäreinheiten	2.Garde≠Feld≠Artillerie≠Regiment Blauhelme, Armeen, Heere, Sondereinsatzkommando Teltower Legion Besatzung Teruels	“Truppen” nicht ausreichend, Einheit hat mindestens eine Funktion, einen Anführer oder einen Stützpunkt
	Modelabel	Chanel	
	Öffentliche o. politische Organisation/Körperschaft	Armee Aufsichtsräte (WENN Firma ersichtlich!) Ausschuss Börse Deutscher Bundestag Die Pforte Eisenbahn EU Expertengruppe Feuerwehr Kabinett Kammer Kommission Krone Militär Ministerien NATO Parlament Polizei Präsidium Presse Regierung Schweizerische Westbahnen Thron Zoll	MinisterORGemb ~~ParlamentarierORGemb~~ StaatsachivarORGemb Metonymien, wie “Thron” oder “Krone” zählen hier ebenfalls
	Politische oder Militärbündnisse (unique)	Dreibund (evtl. embedded EVT)	

	Politische Parteien	Die Grünen Die Linke FDP SPD Opposition	~~Die Liberalen~~ ~~Zentrum~~ ~~Die Linken~~ ~~Kommunisten~~ ~~Sozialdemokraten~~ ~~Sozialisten~~
	Politische Bewegungen	Carlistische Bewegung	~~Kommunismus~~ ~~Bolschewisten~~
	Presse	Berliner Zeitung Die Pforte Tagesspiegel	
	Restaurants, Hotels	Adlon Sassella Zur Linde	Keine LOC!
	Ritterorden	Wasa-Orden	mögl. WORK oder WORKemb. (Kunstobjekt)
	Sender, Rundfunkanstalten	Arte Radio Bremen ZDF	
	Unternehmen	Microsoft VW	
	Vereine, Clubs	Füchse Berlin Lions Club VfB Stuttgart	Mannschaften = PER

Appendix A.3: Example tag-set LOC

Entity class	**Subclass**	**Example**	**Exceptions**
			[Reichs]tag (nur ORG) Ausland Deine Welt Die Frent Die ganze Welt Die Kolonie Feindesland Himmelsrichtungen Inland Inselreich International Ostfront Unsere Welt Ausdifferenzierte Angaben zu Flüssen: Elbe bei Dresden (in diesem Fall sind Elbe und Dresden jeweils gesondert auszuzeichnen)
	Gebäude	Bundestag	

		Kreml Pentagon	
	Geografische Räume (juristisch oder politisch) z.B. Kolonien	Deutsches Zollgebiet Fidschi Französischer Kolonialraum Frz. Hoheitsgebiet La Réunion	~~Reichsland~~ ~~Reich~~ *Embedded Entities bei doppelten Angaben:* Österreich-Ungarn Elsass-Lothringen
	Geografische Räume (kulturell)	Abendland Orient	~~Fernost~~
	Gewässer, Flüsse, Seen, Meere etc.	Spree Viktoriasee	
	Kontinente	Südamerika	
	Länder, Nationen, Staaten	Südafrika	Südamerika ≠ LOCemb
	Landschaften	Lüneburger Heide	
	Planeten, Galaxien	Erde Milchstraße	
	Schiffe	Gorch Fock Belgica Titanic	
	Sehenswürdigkeiten	Brandenburger Tor The Bean	
	Städte	(Hansestadt) Hamburg Kapstadt New York City	
	Stadtteile, Bezirke, Kieze	Köln Deutz Schöneberg	
	Straßen, Plätze	Alexanderplatz Bernauer Strasse	

Appendix A.4: Example tag-set CONF

Entity class	**Subclass**	**Example**	**Exceptions**
Conference (CONF)	Kongresse, Tagungen	CLEF Anatomie-Kongress	~~Expedition~~ ~~wissenschaftliche Reisen~~

Appendix A.5: Example tag-set EVT

Events (EVT)	Demonstrationen	Freitagsdemo von Fridays for Future	
	Festivals	Lollapalooza 2015	
	Firmenspezifische Treffen	Generalversammlung Aufsichtsratsitzung	
	Kriege	Zweiter Weltkrieg	

	Paraden	Christopher Street Day	
	Feste (wiederkehrende)	Weihnachten Ostersonntag	
	Expeditionen	belgische Süddpolexpedition	
	Parteitage	SPD Parteitag	

Appendix A.6: Example tag-set WORK

Entity class	**Subclass**	**Example**	**Exceptions**
Works and expressions (WORK)			~~Technische Serientypen, z.B. Automobiltypen (VW Käfer)~~ ~~Patente~~ ~~Software~~
	Aufführungen und Inszenierungen	Schwanensee	
	Filme	Some Like It Hot	
	Gemälde	Mona Lisa Guernica	
	Gerichtsurteile	Urteil des Reichsgerichts, 3. Strafsenat vom 20. Mai 1882 Manson-Urteil	Urteile haben mindestens eine Institution, Datum; möglichst Thema und Untergremium
	Gesetze, Verträge	Handelsvertrag zwischen Deutschland und Oesterreich vom [Datum (DD)-MM-YY] Meistbegünstigungsvertrag Hartz IV Gesetz vom [Datum (DD)-MM-YY] Staatsetat [YY/YY]	Etats haben mindestens eine Institution und Datum
	Literarische Werke	Die Räuber Frankenstein Der Abrogans	
	Musikstücke, Songs, Alben	Abbey Road Jingle Bells 9. Sinfonie	
	Orden	Orden vom Zähringer Löwen Kommandeur I. Klasse mit Eichenlaub und Schwertern	mögl. auch ORG bzw. ORGemb (Ritterorden)
	Plastiken und Skulpturen		
	Religiöse Werke	(Die) Bibel (Das) Alte Testament 95 Thesen	

Appendix B: List of Exceptions

Rule no.	**Exception**	**Explanation**
5.1 no. 3: Pronouns	Er (eg. “und so vergibt Er uns.”)	Capitalized Pronouns in the middle of a sentence refer to a religious entity and are marked as PER.
5.1 no. 4: Determiner	The Big Apple	This definite description of a location includes articles.
5.2 no. 6: Titles	Kaiser/in Minister/in des Inneren	This is an unambiguous title.
5.3 no. 5: Religious organizations	Katholische Kirche Evangelische Kirche Protestantische Kirche	Clearly identifiable organizations not connected to a location
5.4 no. 2: embedded LOC	Südamerika[LOC]~~[LOCemb]~~	