# Standardized Information on historical Proper Names in Digital Full Text Transcriptions. Crowdsourcing `ref="[ID]"`s for `<placeName>` and `<persName>` tags in the corpora of the German Text Archive / Deutsches Textarchiv

Christian Thomas, Matthias Boenig, Alexander Geyken, Susanne Haaf,
Bryan Jurish, Frank Wiegand, Kay-Michael Würzner

Berlin-Brandenburgische Akademie der Wissenschaften
Deutsches Textarchiv
Jägerstraße 22/23
D-10117 Berlin.
www.deutschestextarchiv.de
dta@bbaw.de

**Keywords.** Digitization, Full Text Transcription, Annotation, Text Encoding Initiative (TEI)-XML, Authority Records, Named Entity Recognition (NER), Collaborative Workflows, Crowdsourcing/Peer Sourcing.

**Abstract.**

In the context of the Deutsches Textarchiv (DTA), funded as a long-term project by the Deutsche Forschungsgemeinschaft (DFG), some 1,600 volumes of historical print-based publications, originally published between the early 17th and the early 20th century have been digitized and transcribed from 2007 to 2015 [01]. Due to the project's primary focus on the history of the German language, the full-text transcriptions document the original printed works, of which the earliest edition accessible was digitized. The transcriptions were acquired for the most part using the highly accurate double-keying method, guaranteeing an exceptionally high quality of the transcripts and the structuring information [03]. As substantial additions to this 'core corpus', highly accurate transcriptions of historical documents have been integrated into the DTA as sub-corpora to enhance the text basis [06]. These additions include another 1,000 documents from edition projects, researchers and libraries. At present (July 2015), the DTA 'core corpus' and its extensional corpora contain fictional and nonfictional first-edition prints (and a currently growing number of manuscripts), a total of more than 620,000 pages with more than 1 billion characters and roughly 150 million tokens.[1] The full text- and meta-data is available for free download under a Creative Commons-license via an API and/or an OAI-PMH interface.[2] The DTA serves as a basis for a large reference corpus for the New High German language, and is a valuable resource for linguistic and lexicographic studies, in contexts of literary analysis and research in the historic disciplines.

The DTA corpus texts are published via the Internet as digital facsimiles and as XML-annotated transcriptions together with comprehensive bibliographic meta-data. The annotation consistently follows the well-documented DTA Base Format (DTABf) [02], a fully interoperable subset of the Text Encoding Initiative's (TEI) P5 Guidelines[3] developed for the representation of (historical) written corpora. All corpus texts are also accessible via the web-based platform for collaborative quality assurance, DTAQ. Within DTAQ, the texts can be commented on, meta data

[1] Further resources are the complete issues of the Polytechnisches Journal (1820–1931; 370 volumes) and Die Grenzboten (1841–1922; 180 volumes). For further information on these important historical journals cf. the project's original web sites at Berlin's Humboldt University for the Polytechnisches Journal, http://dingler.culture.hu-berlin.de/, resp. the State and University Library (SuUB) Bremen for Die Grenzboten, http://brema.suub.uni-bremen.de/grenzboten. [Note: All URLs cited in this paper have been retrieved 2015-07-16.]

[2] Cf. www.deutschestextarchiv.de/download.

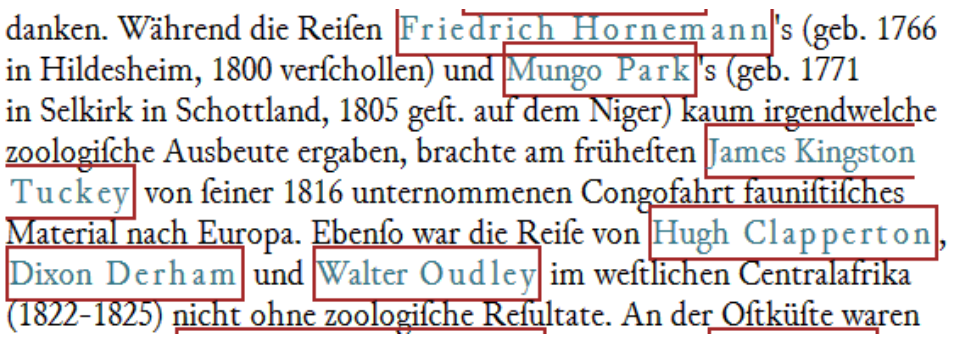
[3] Cf. http://www.tei-c.org/Guidelines/P5/.

can be checked, and any errors or inconsistencies can be reported as 'tickets'. As of July 2015, DTAQ has more than 800 registered users from ca. 130 different institutions in Germany and abroad contributing to the quality assurance process in a collaborative manner. Given the required permission, users can even edit the text and the XML annotation base, e.g. to correct misspellings, printing errors, or improve the annotation online with the implemented web-based editors. (cf. www.deutschestextarchiv.de/dtaq/about).

A further use case for collaborative annotations in DTAQ, which should be of interest for the DCH2015 community, is the possibility to enhance the referencing of proper names with stable URLs to authority records. For example, all persons mentioned in a text can be identified by using the TEI-element `<persName>`, and these entries can be linked to an appropriate database of information, e.g. the Gemeinsame Normdatei (GND)[4], by adding a `@ref` attribute with a unique ID or PURL as an identifier. As of July 2015, a total of 53,622 `<persName>` tags have been assigned to occurences of person's names in selected texts in the DTA corpora, about one third of these have been referenced with an authority date. For example, in J. V. Carus' "Geschichte der Zoologie bis auf Johannes Müller und Charles Darwin" from 1872[5], more than 5,000 persons have been identified and referenced with GND authority data represented in the XML base like this:

```
<p>[…] Während die Reiſen <persName ref="http://d-nb.info/gnd/118707094">Friedrich Hornemann</persName>'s […] und <persName ref="http://d-nb.info/gnd/118591746">Mungo <hi rendition="#g">Park</hi></persName>'s […] kaum irgendwelche zoologiſche Ausbeute ergaben, brachte am früheſten <persName ref="http://d-nb.info/gnd/100648282">James Kingston Tuckey</persName> von ſeiner 1816 unternommenen Congofahrt fauniſtiſches Material nach Europa. Ebenſo war die Reiſe von <persName ref="http://d-nb.info/gnd/119426943">Hugh Clapperton</persName>, <persName ref="http://d-nb.info/gnd/117632368">Dixon Derham</persName> und <persName ref="http://d-nb.info/gnd/1032293004">Walter Oudley</persName> im weſtlichen Centralafrika (1822-1825) nicht ohne zoologiſche Reſultate. […]</p>[6]
```

In the HTML representation, the tokens tagged with `<persName>` are highlighted and a link takes the reader directly to the respective authority information:


danken. Während die Reiſen Friedrich Hornemann's (geb. 1766 in Hildesheim, 1800 verſchollen) und Mungo Park's (geb. 1771 in Selkirk in Schottland, 1805 geſt. auf dem Niger) kaum irgendwelche zoologiſche Ausbeute ergaben, brachte am früheſten James Kingston Tuckey von ſeiner 1816 unternommenen Congofahrt fauniſtiſches Material nach Europa. Ebenſo war die Reiſe von Hugh Clapperton, Dixon Derham und Walter Oudley im weſtlichen Centralafrika (1822-1825) nicht ohne zoologiſche Reſultate. An der Oſtküſte waren

[4] Cf. http://www.dnb.de/gnd.

[5] Carus, Julius Victor: Geschichte der Zoologie bis auf Johannes Müller und Charles Darwin. München, 1872. In: Deutsches Textarchiv, http://www.deutschestextarchiv.de/carus_zoologie_1872.

[6] Ibid., p. 661, http://www.deutschestextarchiv.de/carus_zoologie_1872/672. For the purpose of this paper, the XML representation has been simplified by removing elements representing highlighting (<hi>) and line breaks (<lb/>).

It is obvious how this kind of information being annotated in the source can be used for cross-referencing to other digital resources, e.g. biographical databases or other virtual collections. The same is true for geographical names in the historical text corpus, of which 26,732 have been tagged with a `<placeName>` element (cf. Fig. 1), but many of these would still have to be referenced to be of even greater value for working with the corpus texts. Once the geographical names are being annotated in the standardized way the TEI suggests, and the XML base is enriched with references to authority records like the Getty Thesaurus of Geographical Names (TGN)[7] or GeoNames[8], these geographical names can be linked to databases or thesauri for (historical) places or to digitized (historical) maps.

The DTA supports the process of identifying proper names in the corpus basis with its expertise in computational linguistics [05], e.g. by automatically 'pre-tagging' proper names with Named Entity Recognition (NER) tools [04] (cf. Fig. 2). The quality control and indexing of these entities can comfortably be done online and collaboratively in DTAQ (Fig. 3). Given that the DTA's primary task is on text digitization and structural annotation of the corpus data, 'deeper' indexing like the process described here for persons and places, can only be carried out with the help of external scholars or public users of the texts, i.e. in cooperation projects or as a crowd- resp. peer-sourcing initiative. In our talk at the DCH 2015 we want to show the outlined workflow in more detail and elaborate on the benefits this brings for the wider community, and on the challenges we are facing especially when working on a large, heterogeneous historical text corpus.

**Selected references.**

[01] Geyken, A.: Wege zu einem historischen Referenzkorpus des Deutschen: das Projekt Deutsches Textarchiv. In: Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011, herausgegeben von Ingelore Hafemann, Berlin 2013, S. 221–234. [Full Paper: http://nbn-resolving.de/urn:nbn:de:kobv:b4-opus-24424; urn:nbn:de:kobv:b4-opus-24424]

[02] Haaf, S., Geyken, A., Wiegand, F.: The DTA “Base Format”: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources, In: Journal of the Text Encoding Initiative [Online], Issue 8 – PREVIEW | 2014–2015. [Full Paper: http://jtei.revues.org/1114; DOI: 10.4000/jtei.1114]

[03] Haaf, S., Wiegand, F., Geyken, A.: Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. In: Journal of the Text Encoding Initiative [Online] Issue 4 | March 2013. [Full Paper: http://jtei.revues.org/739; DOI: 10.4000/jtei.739]

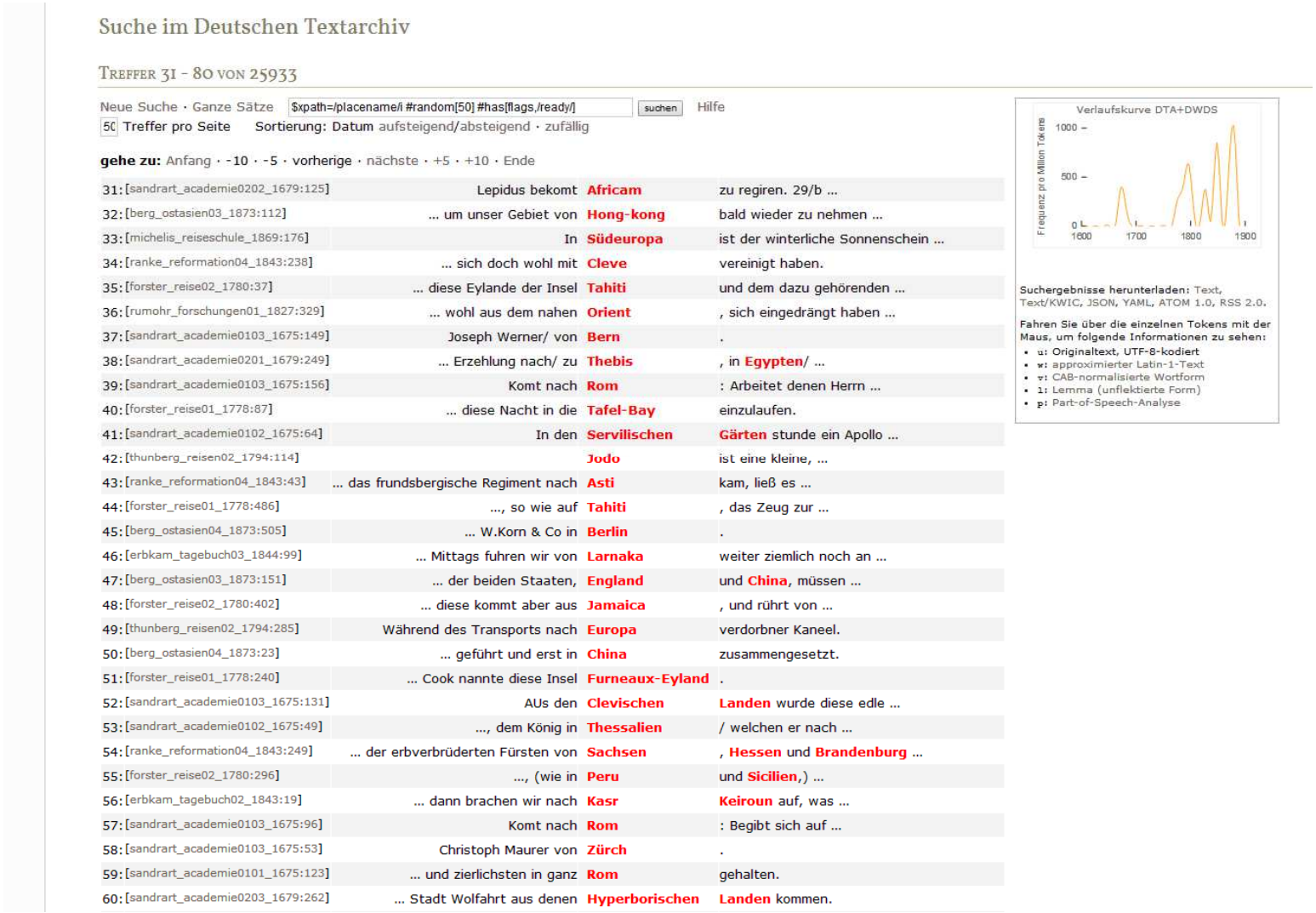
[04] Jurish, B., Thomas, C.: Named Entity Recognition (NER) im Deutschen Textarchiv – Computerlinguistisch gestützte Identifikation von Personen- und Ortsnamen in den Korpora des DTA. Workshop „Mehr Personen – Mehr Daten – Mehr Repositorien“, 4.–6. März 2013 in der Berlin-Brandenburgischen Akademie der Wissenschaften. [Abstract: http://www.deutschestextarchiv.de/files/Abstract_DTAE-NER_vortrag-2013-03-06.pdf, Slides: http://www.deutschestextarchiv.de/files/DTAE-NER_vortrag-2013-03-06.pdf]

[05] Jurish, B., Thomas, C. Wiegand, F.: Querying the Deutsches Textarchiv. In: U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Hg.): Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities (co-located with iConference 2014, Berlin, 4. März, 2014), S. 25–30, 2014. [Full Paper: http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf]

[06] Thomas, C., Wiegand, F.: Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th centuries) in CLARIN-D. In: Gippert, Jost / Gehrke, Ralf (Hrsg.): Historical Corpora. Challenges and Perspectives. Tübingen 2015, S. 181–196. [For an OA version, cf. Preprint of an earlier draft, 2012-10-31: http://edoc.bbaw.de/frontdoor/index/index/docId/2005, urn:nbn:de:kobv:b4-opus-23081]

[7] http://www.getty.edu/research/tools/vocabularies/tgn/.

[8] http://www.geonames.org/.



Suche im Deutschen Textarchiv

Treffer 31 - 80 von 25933

Neue Suche · Ganze Sätze $xpath=/placename/i #random[50] #has[flags,/ready/] suchen Hilfe

50 Treffer pro Seite Sortierung: Datum aufsteigend/absteigend · zufällig

**gehe zu:** Anfang · -10 · -5 · vorherige · nächste · +5 · +10 · Ende

| | | | |
|---|---|---|---|
| 31: [sandrart_academie0202_1679:125] | Lepidus bekomt | **Africam** | zu regiren. 29/b ... |
| 32: [berg_ostasien03_1873:112] | ... um unser Gebiet von | **Hong-kong** | bald wieder zu nehmen ... |
| 33: [michelis_reiseschule_1869:176] | In | **Südeuropa** | ist der winterliche Sonnenschein ... |
| 34: [ranke_reformation04_1843:238] | ... sich doch wohl mit | **Cleve** | vereinigt haben. |
| 35: [forster_reise02_1780:37] | ... diese Eylande der Insel | **Tahiti** | und dem dazu gehörenden ... |
| 36: [rumohr_forschungen01_1827:329] | ... wohl aus dem nahen | **Orient** | , sich eingedrängt haben ... |
| 37: [sandrart_academie0103_1675:149] | Joseph Werner/ von | **Bern** | . |
| 38: [sandrart_academie0201_1679:249] | ... Erzehlung nach/ zu | **Thebis** | , in **Egypten**/ ... |
| 39: [sandrart_academie0103_1675:156] | Komt nach | **Rom** | : Arbeitet denen Herrn ... |
| 40: [forster_reise01_1778:87] | ... diese Nacht in die | **Tafel-Bay** | einzulaufen. |
| 41: [sandrart_academie0102_1675:64] | In den | **Servilischen** | **Gärten** stunde ein Apollo ... |
| 42: [thunberg_reisen02_1794:114] | | **Jodo** | ist eine kleine, ... |
| 43: [ranke_reformation04_1843:43] | ... das frundsbergische Regiment nach | **Asti** | kam, ließ es ... |
| 44: [forster_reise01_1778:486] | ..., so wie auf | **Tahiti** | , das Zeug zur ... |
| 45: [berg_ostasien04_1873:505] | ... W.Korn & Co in | **Berlin** | . |
| 46: [erbkam_tagebuch03_1844:99] | ... Mittags fuhren wir von | **Larnaka** | weiter ziemlich noch an ... |
| 47: [berg_ostasien03_1873:151] | ... der beiden Staaten, | **England** | und **China**, müssen ... |
| 48: [forster_reise02_1780:402] | ... diese kommt aber aus | **Jamaica** | , und rührt von ... |
| 49: [thunberg_reisen02_1794:285] | Während des Transports nach | **Europa** | verdorbner Kaneel. |
| 50: [berg_ostasien04_1873:23] | ... geführt und erst in | **China** | zusammengesetzt. |
| 51: [forster_reise01_1778:240] | ... Cook nannte diese Insel | **Furneaux-Eyland** | . |
| 52: [sandrart_academie0103_1675:131] | AUs den | **Clevischen** | **Landen** wurde diese edle ... |
| 53: [sandrart_academie0102_1675:49] | ..., dem König in | **Thessalien** | / welchen er nach ... |
| 54: [ranke_reformation04_1843:249] | ... der erbverbrüderten Fürsten von | **Sachsen** | , **Hessen** und **Brandenburg** ... |
| 55: [forster_reise02_1780:296] | ..., (wie in | **Peru** | und **Sicilien**,) ... |
| 56: [erbkam_tagebuch02_1843:19] | ... dann brachen wir nach | **Kasr** | **Keiroun** auf, was ... |
| 57: [sandrart_academie0103_1675:96] | Komt nach | **Rom** | : Begibt sich auf ... |
| 58: [sandrart_academie0103_1675:53] | Christoph Maurer von | **Zürch** | . |
| 59: [sandrart_academie0101_1675:123] | ... und zierlichsten in ganz | **Rom** | gehalten. |
| 60: [sandrart_academie0203_1679:262] | ... Stadt Wolfahrt aus denen | **Hyperborischen** | **Landen** kommen. |

Verlaufskurve DTA+DWDS

Frequenz pro Million Tokens: 1000, 500, 0; 1600, 1700, 1800, 1900

Suchergebnisse herunterladen: Text, Text/KWIC, JSON, YAML, ATOM 1.0, RSS 2.0.

Fahren Sie über die einzelnen Tokens mit der Maus, um folgende Informationen zu sehen:

- u: Originaltext, UTF-8-kodiert
- w: approximierter Latin-1-Text
- v: CAB-normalisierte Wortform
- l: Lemma (unflektierte Form)
- p: Part-of-Speech-Analyse

*Fig. 1:* List of geographical names annotated as `<placeName>` within the DTA corpora (randomly sorted extraction).

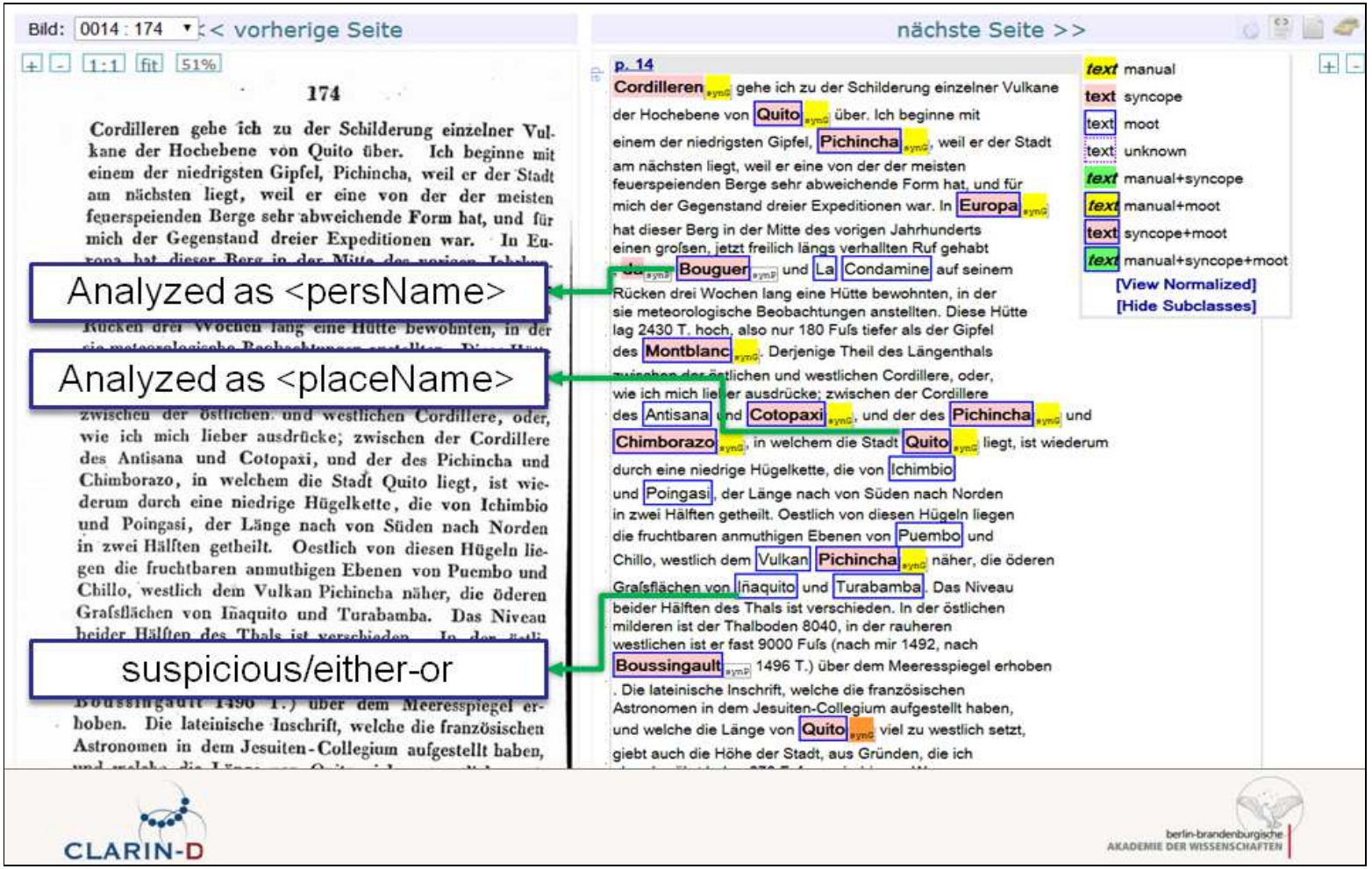



*Fig. 2:* Automated Named Entity Recognition (NER) in the DTAQ platform, supporting manual efforts to identify proper names (persons, places).

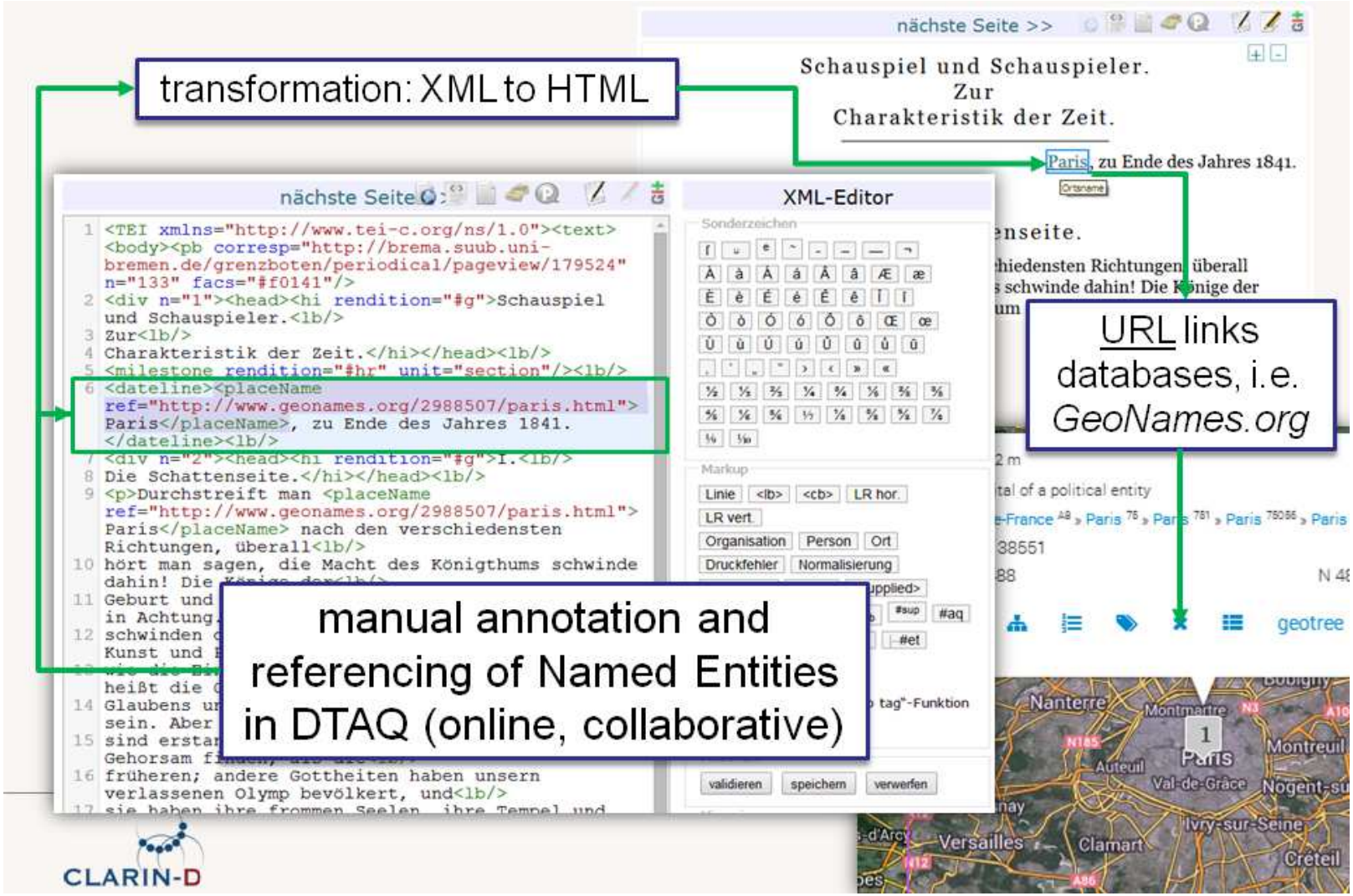



*Fig. 3:* Exemplary workflow for annotation and indexing of named entities.