



Publikationen des Deutschen Archäologischen Instituts

Peter Baumeister

A Test Report on a Lightweight Pipeline for Named Entity Recognition: Using Machine Learning Models for Metadata Enrichment on Abstracts in iDAI.bibliography

Forum for Digital Archaeology and Infrastructure Faszikel 2026 1–38 (§)

<https://doi.org/10.34780/s6tar918>

Herausgebende Institution / Publisher:
Deutsches Archäologisches Institut

Copyright (Digital Edition) © 2026 Deutsches Archäologisches Institut
Deutsches Archäologisches Institut, Zentrale, Podbielskiallee 69–71, 14195 Berlin, Tel: +49 30 187711-0
Email: info@dainst.de | Web: <https://www.dainst.org>

Nutzungsbedingungen:

Mit dem Herunterladen erkennen Sie die [Nutzungsbedingungen](#) von iDAI.publications an. Sofern in dem Dokument nichts anderes ausdrücklich vermerkt ist, gelten folgende Nutzungsbedingungen: Die Nutzung der Inhalte ist ausschließlich privaten Nutzerinnen / Nutzern für den eigenen wissenschaftlichen und sonstigen privaten Gebrauch gestattet. Sämtliche Texte, Bilder und sonstige Inhalte in diesem Dokument unterliegen dem Schutz des Urheberrechts gemäß dem Urheberrechtsgesetz der Bundesrepublik Deutschland. Die Inhalte können von Ihnen nur dann genutzt und vervielfältigt werden, wenn Ihnen dies im Einzelfall durch den Rechteinhaber oder die Schrankenregelungen des Urheberrechts gestattet ist. Jede Art der Nutzung zu gewerblichen Zwecken ist untersagt. Zu den Möglichkeiten einer Lizenzierung von Nutzungsrechten wenden Sie sich bitte direkt an die verantwortlichen Herausgeber*innen der jeweiligen Publikationsorgane oder an die Online-Redaktion des Deutschen Archäologischen Instituts (info@dainst.de). Etwaige davon abweichende Lizenzbedingungen sind im Abbildungsnachweis vermerkt.

Terms of use:

By downloading you accept the [terms of use](#) of iDAI.publications. Unless otherwise stated in the document, the following terms of use are applicable: All materials including texts, articles, images and other content contained in this document are subject to the German copyright. The contents are for personal use only and may only be reproduced or made accessible to third parties if you have gained permission from the copyright owner. Any form of commercial use is expressly prohibited. When seeking the granting of licenses of use or permission to reproduce any kind of material please contact the responsible editors of the publications or contact the Deutsches Archäologisches Institut (info@dainst.de). Any deviating terms of use are indicated in the credits.



FORUM FOR DIGITAL ARCHAEOLOGY AND INFRASTRUCTURE

ABSTRACT (EN)

This article aims to encourage infrastructure-focused work units and teams with limited prior exposure to machine learning to actively explore current technologies and gain practical insight into their possibilities and limitations. The approach presented here can aid in the development of viable concepts and give a clearer understanding of technological feasibility, with a focus on practical solutions. To this end, the article presents a test scenario, applying Named Entity Recognition (NER) to abstracts in iDAI.bibliography – the catalogue of the DAI libraries. The approach uses a lightweight pipeline built on freely available models, a simple code base, standard hardware, and copyright-compliant methods, demonstrating how automated processing can meaningfully reduce human effort and improve the quality of the entries.

KEYWORDS

Named Entity Recognition, Metadata Enrichment, 'Hugging Face Transformers' pipeline

A Test Report on a Lightweight Pipeline for Named Entity Recognition

Using Machine Learning Models for
Metadata Enrichment on Abstracts in
iDAI.bibliography

Introduction

[1] In the last few years two phenomena seem to become increasingly important when it comes to the implementation of technical solutions for dealing with complex tasks:

[2] First, in the fields of programming, data science or machine learning we see that IT capacity and/or capabilities seem not to increase in proportion to the need.¹ This presumed shortage leads to the assumption that organizations have to decide more and more selectively into which projects to invest their resources and capacities. One possible effect might be that not every project considered by a department or work unit to be useful or desirable can be supported accordingly, especially when it comes to daily and low-level work tasks. Therefore, teams and work units that have not had much exposure to new technologies so far have to take more and more initiative. Or to put it more positively: the increased adoption and effective utilization of technology can enhance the autonomy of the teams, enabling them to operate more efficiently.

¹ On the problem of skilled labor shortage in IT departments at universities and research institutions in general see [Dreyer 2025](#), 34. This applies not only to the field of application, but also to research itself, see [Jurowetzki et al. 2025](#). For the economic sector this effect of skilled labor shortage is documented as well, see <https://www.bitkom.org/Presse/Presseinformation/Deutschland-fehlen-IT-Fachkraefte>, also [Büchel et al. 2023](#). – If a slight decline in the relevance of this question in 2026 (see [Dreyer 2026](#), 15) is linked to the possibilities that large language models increasingly offer in the field of programming or other IT tasks, remains to be seen.

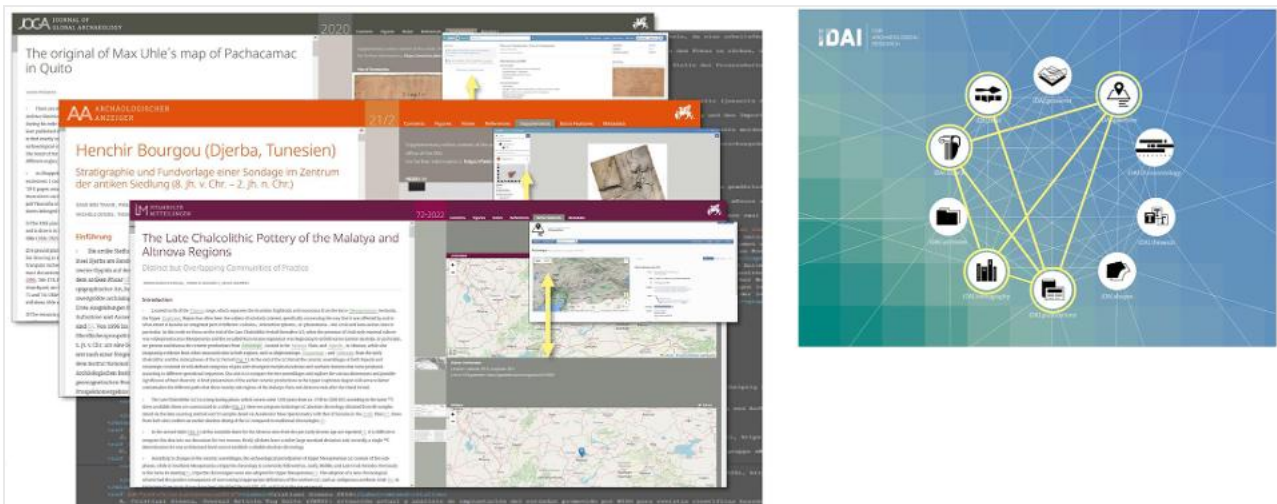


Fig. 1: DAI Journal Viewer Concept: Integrating additional research data into journal articles



Fig. 2: DAI Journal Viewer: Preview of an iDAI.gazetteer entry

[3] Second, and in parallel, one can observe a remarkable increase of new technological possibilities, especially in the field of available machine learning models, and – equally important – an improved accessibility to those technologies in the form of easily applicable frameworks.

[4] Given these circumstances, it seems more than sensible for teams that need automated solutions to actively seek points of contact with these technologies, not necessarily to build reliable solutions themselves in the first place, but to gain experience with possibilities and limitations, that helps to outline concepts and develop an understanding of their feasibility and – no less important – to reflect on the decision making processes and methods used in their daily tasks.

[5] This article² describes a test that pursues comparable objectives, encouraging teams working in the field of infrastructure to explore the possibilities that current machine learning technology already offers³.

2 I would like to thank Federico Buccellati, Sabine Thänert and especially Fabian Riebschläger and Lisa Steinmann, for critically reviewing the text and for their helpful suggestions.

3 Albeit a different focus, this article is based on a lecture “Szenarien zur Metadatenanreicherung durch den Einsatz von Machine Learning-Modellen in der Archäologie” given by Sabine Thänert and the author at the BiblioCon 2025.

[6] The starting point was the DAI's new concept for publishing scientific journals: alongside the print and PDF versions, the journal articles are published within the DAI Journal Viewer, a customized version of the 'elife lens viewer'⁴, based on JATS .xml files ([fig. 1](#)).⁵ One central aspect of this approach is to enrich the machine-readable article text with authority data by linking selected entities with entries of specific authority data systems, as well as other iDAI.world services⁶ to integrate content 'on the fly' from other information resources. Location names are, for example, linked with entries in the iDAI.gazetteer⁷, a service connecting toponyms with coordinates and which represents the DAI authority instance for geographical information about archaeological places. Finally, a tab within the viewer shows a preview version of iDAI.gazetteer's main information on the location, including a map section view ([fig. 2](#)).

[7] Because neither the marking of entities in the article text nor the extraction and linking is usually carried out consistently beforehand, this leads to an additional workload for the copy editors. Therefore, the idea was to support this process by using machine learning methods. For this purpose, a Named Entity Recognition (NER) plugin was created for the so called TagTool_WiZard application (ttw), a tool that supports the pre-structuring and pre-processing of DAI Journal Viewer article documents.⁸ The newly added NER plugin extracts location names using the 'Hugging Face Transformers' pipeline, queries the iDAI.gazetteer and returns suggestions of gazetteer IDs for the extracted place names ([fig. 3](#)).⁹

[8] Although the plugin itself is still experimental, it was decided to test it without much preparation for other use cases. One of the DAI's information resources that contains textual and text related content is iDAI.bibliography ([fig. 4](#)).¹⁰ It represents not only the catalogue of the DAI libraries but is also one of the largest databases with additional bibliographic entries on ancient studies. Due to its history (see below), the quality of the entries varies – primarily due to the heterogeneous nature of the descriptive information – so that any additional metadata is a welcome addition to improving and expanding the entries in iDAI.bibliography. This resulted in the idea to modify ttw's NER plugin slightly to make it run also on the entries in iDAI.bibliography to see whether sustainable added value for the entries can be generated.¹¹ The so called BibPip pipeline.¹²

4 <https://lens.elifesciences.org> resp. <https://github.com/elifesciences/lens> (26.03.2026).

5 For the concept in general see Baumeister 2022. Articles illustrating the concept can be found, for example, in the Archäologische Anzeiger (<https://publications.dainst.org/journals/aa/issue/archive>), e. g. Ben Tahar et al. 2021; Kose et al. 2022 or Tzochev 2021 and so on. This concept has now been adopted by almost all DAI journals, see <https://publications.dainst.org/journals>.

6 The iDAI.world is a digital research environment providing services to store, explore, analyse, visualise and publish research data worldwide, see <https://idai.world>.

7 <https://gazetteer.dainst.org>.

8 https://github.com/pBxr/TagTool_WiZard. The NER plugin was added in version 2.1.0 (2025-03-11).

9 see in detail 'README.md' and 'ttw_help.html', https://github.com/pBxr/TagTool_WiZard.

10 <https://zenon.dainst.org>.

11 This is, of course, a relatively vaguely described goal, but it is clear that the test does not claim to achieve automatic subject indexing or anything similar, see e. g., Kasprzik 2025; Kasprzik 2023. Very informative is Balnaves et al. 2025 containing a wide range of articles. See also footnote 29 on Named Entity Linking. – A good overview of automated content indexing in libraries in Germany is provided by Weers 2025, 17–32.

12 see <https://github.com/pBxr/BibPip> resp. <https://doi.org/10.5281/zenodo.19238548>. On how to set up the environment see the ttw documentation (see footnote 8).

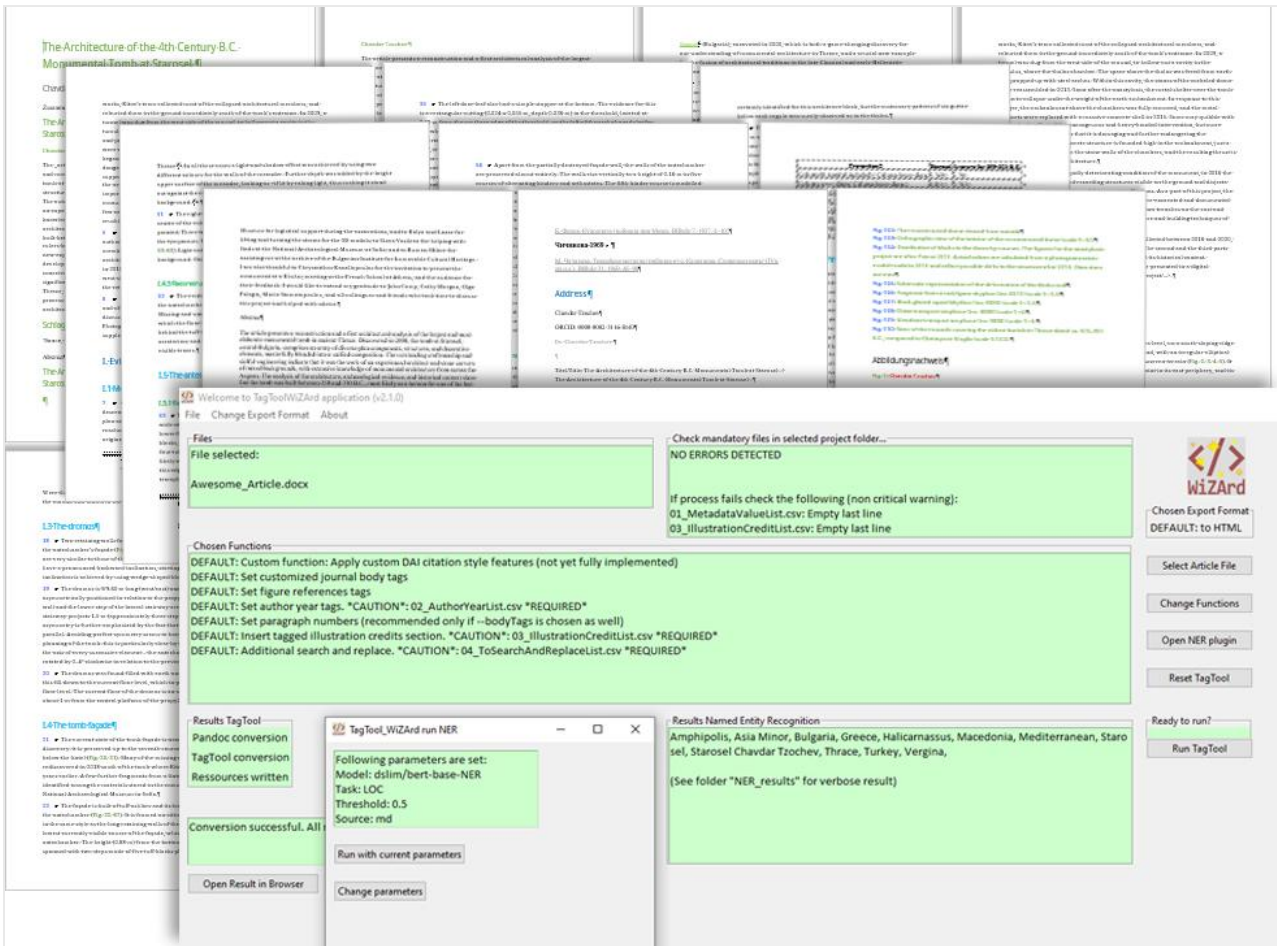


Fig. 3: A tool for semi-automatic formatting journal article documents



Fig. 4: IDA.bibliography/Zenon

[9] Thus the aim was explicitly not to outline and implement a data science project including model trainings etc. but to test a pipeline with a) existing and freely available models, b) on basis of a relatively simple code base with c) standard hardware that d) complies with copyright requirements and – most important – e) that achieves a beneficial outcome in that sense that the human effort required can be reduced.

Approach

General characteristics

[10] In abstract terms, BibPip has the following characteristics:¹³

- The aim was originally to create a pipeline that is based on a relatively simple code base. BibPip is therefore written in Python and uses only common Python libraries.
- For the core feature, i. e. the machine learning tasks, the 'Hugging Face Transformers' pipeline¹⁴ is used. In many aspects this platform can be considered as one possible helpful entry point into the world of machine learning. This corresponds with one goal, namely to encourage non-specialists to engage with machine learning technologies.¹⁵
- A modular approach offers the possibility to experiment with models, settings and parameters.¹⁶
- Detailed logs help to monitor how the pipeline works, make the results traceable and give an impression of the quality and the effects of the selected options on it.¹⁷
- The current setup allows to run the models on a local machine, making it compliant with copyright protection regulations¹⁸.
- It can be run on normal desktop machines without a dedicated GPU, nevertheless, with some limitations that are described below.

The source: iDAI.bibliography entries

[11] As mentioned above, BibPip is working on iDAI.bibliography's entries.¹⁹ The technological basis behind iDAI.bibliography is Koha, a widely used open source library management system.²⁰ iDAI.bibliography contains ca. 1,482,000 bibliographical entries²¹ on books/monographs, journals, articles, anthologies, map collections, e-Resources and so on. For these entries a number of abstracts was recorded over the years, either by data harvesting from other resources or by entering the abstracts, especially in the case of DAI publications. In total 30,979 datasets with abstracts were used for this experiment.

13 The BibPip repo contains the usual documentation about technical aspects etc.

14 <https://huggingface.co>. The Stanford AI Report 2024 provides a good definition of what Hugging Face is: "Hugging Face: a platform and community dedicated to machine learning and data science [...] serves as a one-stop destination for building, deploying, and training machine learning models [...] offering a GitHub-like hub for AI code repositories, models, and datasets [...] has attracted significant attention from industry giants" (Maslej et al. 2024). See also Pol et al. 2024. For articles that highlight also critical aspects, see Osborne et al. 2024; Suryani et al. 2025.

15 In addition to extensive documentation, there are also a variety of relevant tutorials, <https://huggingface.co/learn>.

16 see the 'NER_Parameter's variable in 'settings.py', e. g. model, aggregation strategy or confidence threshold.

17 01_log.txt, 02_input-texts.txt, RESULT.csv, Statistics.csv, Statistics.html (plus file name extensions).

18 Especially when used only for experimental use cases, another relevant argument in favour of using a model on a local machine is the more limited environmental impact, especially when used only for experimental use cases. For a general overview see Maslej et al. 2025, 71–74.

19 This paragraph including the following footnotes was contributed by Sabine Thänert.

20 see <https://koha-community.org>.

21 May 2025.

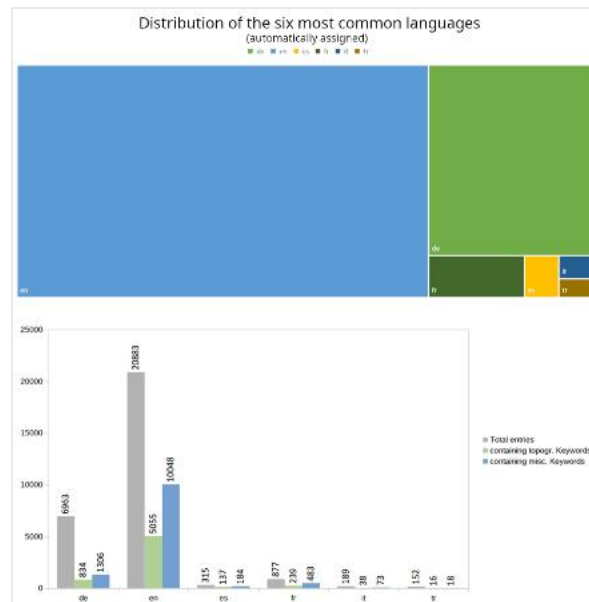


Fig. 5: Automatically assigned languages with the 'langdetect' library

[12] Looking more in depth, the construction of a pipeline is met with considerable challenges due to the very heterogeneous nature of the data itself – and this is probably representative of many information resources with a similar history:

1. Data history: The current content of the DAI's Koha instance is a result of the DAI libraries' long history – and of the history of its digital library management. Correspondingly, the entries' properties are not homogenous. There were a total of 12 data migrations involving various merge processes, and one can easily imagine the data's heterogeneity resulting from these different sources and transitions.²²
2. Multiple languages, varying transliterations: In line with the DAI's research focuses, the entries cover a wide range of languages. Thus the necessity for transliterations. Although strict formatting and data standards apply, variations could not be completely avoided and deviations occur.²³
3. Range of 'Domains': The data sets reflect the broad thematic and scientific spectrum of DAI's work. The entries span an exceptionally broad range of discipline-specific terminologies, geographic regions, and entity types, making named entity recognition far more complex.

22 With a long history beginning in 1829, the DAI today runs 16 libraries in different locations, the largest of which are the libraries of the Rome department (ca. 240,000 titles), the Roman-Germanic Commission (RKG), Frankfurt (ca. 200,000 titles), and the Head Office in Berlin (ca. 100,000 titles). Some of the DAI's libraries began using library management systems like BIS-LOK or LIDOS over 25 years ago, but at that time as more or less autonomously acting libraries. In 2002 the DAI began to structure and consolidate the former decentralized libraries as a single network, introducing Aleph as central library management system (see <https://exlibrisgroup.com>), with (at that time) seven data-contributing libraries (Athens, Istanbul, Madrid, the Eurasia and Orient department, the Head Office in Berlin and the Commission for Archaeology of Non-European Cultures [KAAK] in Bonn) on the basis of ca. 35,000 datasets, initially contributed by the Berlin libraries and the KAAK. The departments in Rome and Cairo and the Romano-Germanic Commission (RGK) followed. Almost 20 years later, Aleph was replaced by the aforementioned Koha.

23 e. g. the Library of Congress' ALA-LC Romanization Tables (see <https://www.loc.gov/catdir/cpsol/roman.html>) respectively the ex DIN 1460 tables for Cyrillic titles.

Preparation

[13] For further processing, the Koha datasets were exported as a .csv file with only minimal manual preparation.²⁴ Because a substantial part of the Koha entries lacks language information, the pipeline starts also with language detection and assignment.²⁵ In total, 53 languages are detected, with six most common languages ('de', 'en', 'es', 'fr', 'it', and 'tr') ([fig. 5](#)).²⁶

Named Entity Recognition

Models and modularity

[14] As mentioned previously, the core functionality is provided by the 'Hugging Face Transformers' pipeline and the models available on the 'Hugging Face' platform.

[15] Here we encounter the main challenge: models that can be considered to be specifically pretrained or suitable for the subject area or 'domain(s)' discussed here are more or less missing, at least in the sense of the whole range of disciplines, topics, geographic regions and languages, that is covered by the DAI's research focus. Nevertheless, there are models focused on archaeology; one very helpful example is the ArchaeoBERT-NER model by Alex Brandsen and Leiden University, which is available for the 'Hugging Face Transformers' pipeline and therefore integrable.²⁷

[16] The lack of specifically pretrained models leads one to ask if there is an option to deal with this challenge without incurring enormous effort. Besides the need for a vast and high-quality dataset that accurately represents the specialized knowledge and terminology, the fine-tuning of the model involves iterative testing, which can be both time-consuming and costly – making it far too complicated for a simple test. Since it was therefore clear from the beginning that no training will be conducted, one possible way to have a positive impact on the quality of the results can be a modular approach, allowing different setups and adjustments and the observation of the subsequent effects. BibPip's options allow to run a model only on entries with a selected language so that different models with different focuses on specific languages can be run sequentially. Other settings enable, for testing purposes, different options that might have an impact on the result as well e. g. the so called aggregation strategy or the threshold.

[17] The broad spectrum of languages offers an additional challenge that NER has to face, and it raises the question if there is a possibility to deal with this variety without specific training actions using only simple mechanisms. For this reason, the possibility to call a translation model was integrated. Using an

24 For a better understanding, some of the MARC21 codes denoting the column names were changed to more readable terms e. g. from '520a' to 'Abstract' and so on.

25 Using the 'langdetect' library that returns ISO 639-1 language codes.

26 See below for the error rate.

27 <https://huggingface.co/alexbrandsen/ArchaeoBERT>. See Brandsen 2022; Brandsen et al. 2022. – If other models are integrated, minor alterations of the code might be necessary; e. g. other models may have different class definitions/designations which must be taken into account in order to correctly produce the desired results.

exemplary and widely used many-to-many model²⁸ for this test, the default entries ('Title', 'Subtitle', 'Topical_Term', 'Geographic_Name' and the 'Abstract') in 'de', 'fr', 'es', 'it' were translated into 'en'.

Plausibility

[18] The simplest way to check the plausibility of a result is to query an authoritative information resource.²⁹ BibPip comes with the option to run the extracted location names through the aforementioned iDAI.gazetteer webservice.³⁰

Tests and results

Preliminary remark

[19] In respect of reproducibility, BibPip was tested with an exemplary and generic model, in this case dslim/bert-base-NER.³¹ The results below are based on this configuration.³² For more specific scenarios, this model, as well as the translation model, can be – up to a certain extent – replaced.

[20] It is to be noted that BibPip is focusing on locations and place names (denoted as 'LOC') for this experiment, other recognized entities (e. g. persons ['PER'] and so on) are being recorded in the results as 'MISC' at the moment.

Quantity and quality

[21] For this test, the complete dataset was processed in several runs with different settings and options, although not all combinations were tested. As stated above, the aim was not to outline and implement a data science project but to see if it is possible to setup a working pipeline in general. For this reason, the goal was not to measure the specific result accuracy as would be expected within a normal data science context. This is not only because the effort involved would be very high, but also because it would be necessary to work on all stages of the process, and in particular on the preparation of the source data, in order to create a valid basis on which to reliably measure result quality. Instead the goal was to quickly generate a rough and sufficient impression of the results and the quality. A set of log documents was additionally an aid in assessing the results.³³

28 Also on the local machine, in this case for this test 'facebook/m2m100_418M', see https://huggingface.co/facebook/m2m100_418M.

29 To give an impression of what is usually necessary to perform Named Entity Linking, see e. g. [Menzel et al. 2021](#). See also footnote 11.

30 Differing from ttw's NER plugin, which returns suggested gazetteer IDs to prepare the semi-automatic linking of the article files, BibPip's iDAI.gazetteer query checks only for hits in general.

31 <https://huggingface.co/dslim/bert-base-NER>.

32 The tests on which the results are based were carried out in May 2025, so this is the reference point; changes may have occurred since then and results may differ. – Due to copyright reasons, it is not possible to make the whole test dataset available.

33 In order to quickly detect possible errors during the compilation of the source texts, these are additionally stored in a separate log file ("02_input-texts.txt").

Language	Entries	... with entries in kwLOC	... with entries in kwMISC	... newly added kwLOC	... newly added kwMISC
0 af	8	1	5	8	31
1 ar	5	0	0	8	14
4 ca	29	8	9	44	118
5 cs	5	1	2	5	17
6 cy	2	0	1	0	2
7 da	12	3	5	10	31
8 de	6969	835	1312	13837	39410
9 el	42	19	22	59	149
10 en	20868	5047	10037	46982	86294
11 es	318	140	183	607	1070
12 et	6	1	2	12	17
14 fi	8	1	3	24	26
15 fr	886	236	484	1580	2919
17 he	1	0	1	0	0
19 hr	27	3	3	42	62
21 id	26	11	14	70	30
22 it	181	40	72	356	523
25 ko	5	4	5	4	5
26 lt	5	0	0	3	17
27 lv	3	1	1	4	2
28 mk	1	1	1	1	1
32 nl	48	9	26	68	104
33 no	48	0	5	31	140
35 pl	4	0	0	0	9
36 pt	50	5	4	33	152
37 ro	9	2	8	8	11
38 ru	16	13	15	8	30
39 sk	1	1	1	0	1
40 sl	58	0	54	147	279
41 so	2	1	1	1	4
43 sv	35	3	10	48	140
44 sw	13	7	13	3	32
48 tl	1056	35	43	36	903
49 tr	145	15	16	228	689
50 uk	1	1	1	0	1
52 vi	2	1	2	4	0
53 zh-cn	11	6	10	4	15

Fig. 6: Example of statistical evaluation of results

[22] To start with the quantity, the numbers might seem impressive at first glance, especially the amount of extracted 'MISC' elements: Depending on the configuration, the numbers reach from 98,655 to a maximum of 134,490 newly extracted keywords (fig. 6). The amount of 'LOC' hits is also comparatively high, between 44,252 and 65,230.

[23] The fact that not all possible combinations of options were tested makes the results difficult to interpret. Taken individually, some settings seem to have only a minor effect.³⁴ But in combination they lead to the range between the lower and upper figures mentioned above.

[24] Interestingly, the experiment of the automated translations doesn't seem to have a positive impact on the quantity produced – translations to English do not generate more or better results. If we take for example the 6,963 abstracts that have been labeled as 'de', the model extracts 39,689 keywords from the original version, in the translated version only 39,410 are detected (fig. 7). Speaking of the total of all 10,096 non-English abstracts the range is even broader, it reaches from 46,954 extracted keywords in the original versions to only 29,669 in the automatically translated texts (fig. 8).

[25] Apart from the quantities, a look at the detailed log files helps to get an impression of the quality. Unsurprisingly, high error rates accompany every step of the process, starting from the language detection, the translation, the NER process itself and the plausibility check.

34 For example, the so called Aggregation Strategy. Roughly speaking 'first' leads in this test to the most results, 'max' to ca. 5% and 'medium' to ca. 20% less hits.

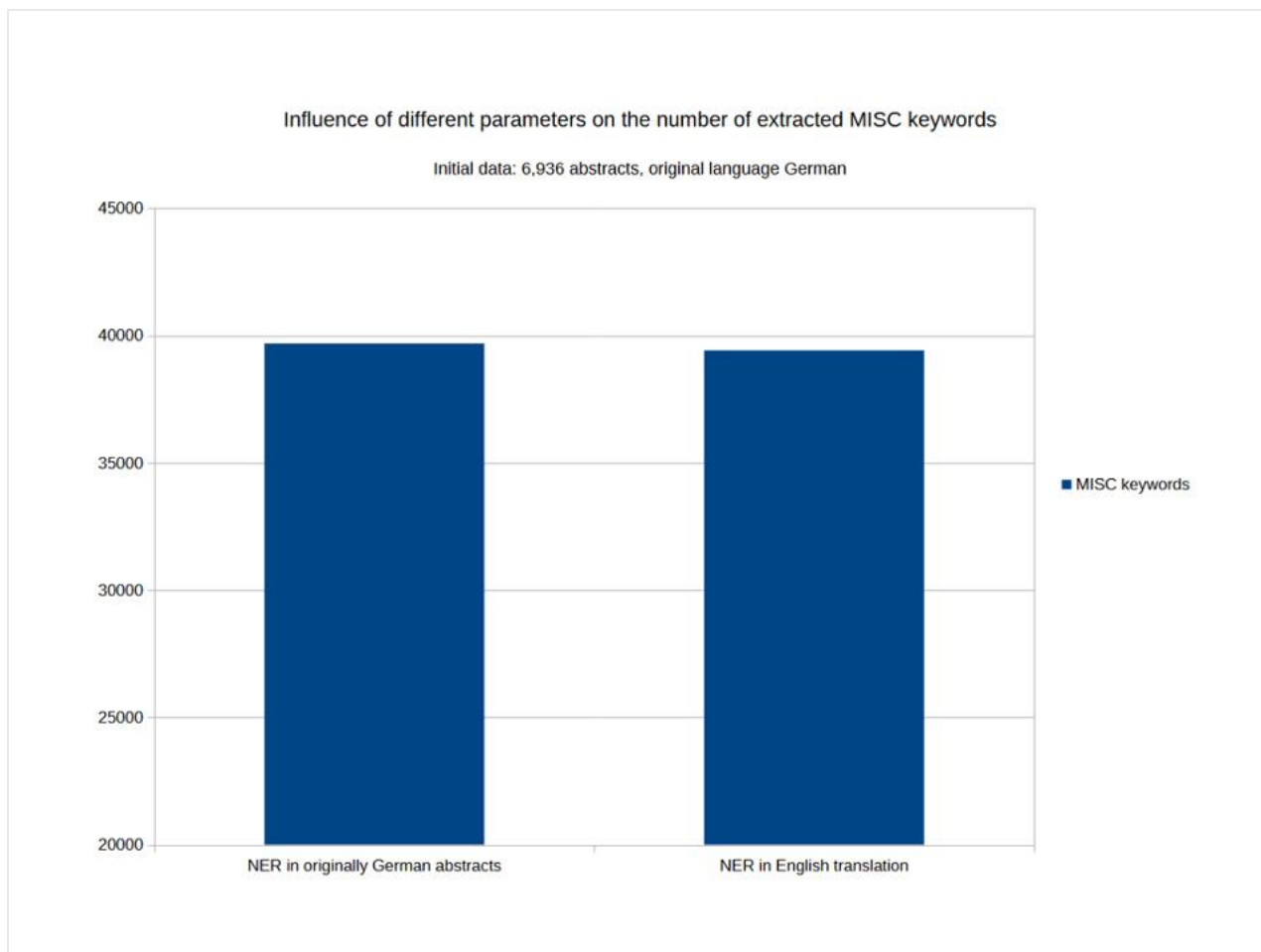


Fig. 7: Translation effects on NER results on German abstracts

[26] Regarding the NER results, all the well-known and therefore expected strengths and weaknesses clearly become apparent: If we have an abstract for which we can assume that there were many (English) texts on this topic available to train the model, the hit rate is high (fig. 9),³⁵ this also seem to apply to summaries with more simple wording (fig. 10).³⁶ Of course, in this test case the newly found keywords alone would only be of limited use in describing the content precisely. However, they can be a useful addition when combined with the existing information and thus contribute to better indexing. Examples of entries that lie at the borders of our own area of expertise, either thematically or geographically, show how helpful such suggestions can be (fig. 11),³⁷ not to mention the cases where no keywords exist in Koha at all. But it needs to be stressed that these examples are by no means representative of the general quality of the output. For more specialized topics and less commonly known languages, the rate decreases rapidly.

35 Biblionumber 1265501 = <https://zenon.dainst.org/Record/001541223>. This entry does not contain keywords, but the parallel title does (<https://zenon.dainst.org/Record/001558576>).

36 Biblionumber 29551 = <https://zenon.dainst.org/Record/000048515>.

37 Biblionumber 1233684 = <https://zenon.dainst.org/Record/001508878>. Notifications in BibPip's log indicate when keywords are already existing in Koha: "(...) already in Koha)".

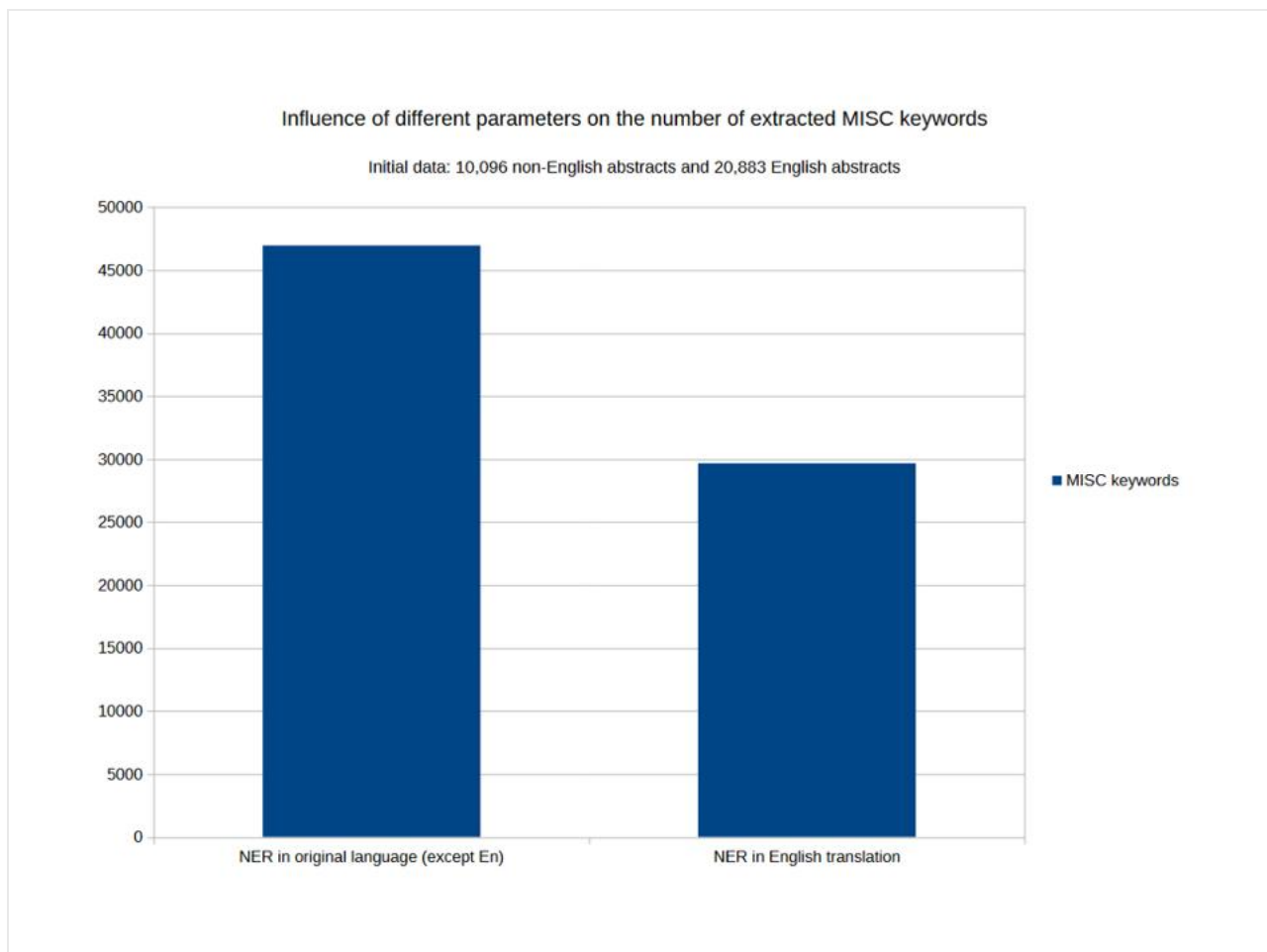


Fig. 8: Translation effects on NER results on all non English abstracts

[27] The plausibility check option for 'LOC' keywords when querying the iDAI.gazetteer is a main factor for a reduced number of hits. BibPip returns a log with negative hits that can help to identify missing location names in the iDAI.gazetteer (fig. 12). But the query fulfills its purpose in eliminating false positive hits only in part, despite the fact that a certain number of terms were filtered out correctly (fig. 13). The multilingualism and transliteration pose a fundamental challenge, because of the amount of additional non-German entries. The number of English versions of the entries in the iDAI.gazetteer is limited, resulting in many false negatives.³⁸ Additionally, we face the basic problem of missing entries and inconsistencies that is an inevitable phenomenon of any information resource that has been built up heterogeneously over many years.

38 In case of negative non-German 'LOC' hits one way to obtain slightly more reliable results is to start a new query with an automatically translated German version. This function/option (see 'translateKWforGazetteerCheck') was implemented in BibPip after the test runs described here, therefore a statistical evaluation is omitted here; in a separate run 1.526 translated 'LOC' hits were recorded - however, analysis of the results shows that a certain error rate is to be expected here as well. – Interestingly, this function can apparently also help in case of incorrect or unusual spellings, as some place names were found after their (correct) translation (e. g. "Mozaambique" -> "Mosambique", etc.).

```

381850 Biblionumber 1265501:
381851 1. Filtered entities
381852
381853 Entity type ['LOC']:
381854 Alexandria
381855 Dead Sea
381856 Egypt
381857 (Judaism not in gazetteer)
381858 (Khirbet Qumran not in gazetteer)
381859 Lake Mareotis
381860 Qumran
381861
381862 Entity type ['MISC']:
381863 Classical
381864 Dead Sea Scrolls
381865 Egyptian
381866 Essenes
381867 Graeco
381868 Hellenistic
381869 Hellenistic Jews
381870 Jewish
381871 Jewish Pythagoras
381872 Palestinian
381873 Pythagoreanism
381874 Roman
381875
381876 Entity type ['ORG']:
381877 Library of Alexandria
381878
381879 Entity type ['PER']:
381880 Philo
381881
381882 2. Verbose NER result:
381883 {'entity_group': 'LOC', 'score': 0.9996476, 'word': 'Alexandria', 'start': 0, 'end': 10}
381884 {'entity_group': 'LOC', 'score': 0.99960417, 'word': 'Qumran', 'start': 15, 'end': 21}
381885 {'entity_group': 'MISC', 'score': 0.9837672, 'word': 'Dead Sea Scrolls', 'start': 112, 'end': 128}
381886 {'entity_group': 'LOC', 'score': 0.99925387, 'word': 'Khirbet Qumran', 'start': 132, 'end': 146}
381887 {'entity_group': 'LOC', 'score': 0.9991151, 'word': 'Dead Sea', 'start': 154, 'end': 162}
381888 {'entity_group': 'MISC', 'score': 0.984291, 'word': 'Dead Sea Scrolls', 'start': 176, 'end': 192}
381889 {'entity_group': 'LOC', 'score': 0.9984164, 'word': 'Qumran', 'start': 331, 'end': 337}
381890 {'entity_group': 'MISC', 'score': 0.813287, 'word': 'Dead Sea Scrolls', 'start': 359, 'end': 375}
381891 {'entity_group': 'LOC', 'score': 0.99967813, 'word': 'Alexandria', 'start': 420, 'end': 430}
381892 {'entity_group': 'LOC', 'score': 0.99971837, 'word': 'Egypt', 'start': 432, 'end': 437}
381893 {'entity_group': 'LOC', 'score': 0.99966514, 'word': 'Alexandria', 'start': 439, 'end': 449}
381894 {'entity_group': 'MISC', 'score': 0.9930871, 'word': 'Hellenistic Jews', 'start': 469, 'end': 485}

```

Fig. 9: Example: Log entry biblionumber 1265501 (Zenon-ID 001541223)

```

8458 Biblionumber 29551:
8459 1. Filtered entities
8460
8461 Entity type ['LOC']:
8462 Aegean
8463 Africa
8464 Australia
8465 Caucasus
8466 Central Asia
8467 China
8468 Egypt
8469 Europe
8470 Far East
8471 Indonesia
8472 Japan
8473 (Latin America not in gazetteer)
8474 Levant
8475 Mesopotamia
8476 (Michaelstein not in gazetteer)
8477 Middle East
8478 Near East
8479 (Parthians not in gazetteer)
8480
8481 Entity type ['MISC']:
8482 Archaeology
8483 Celts
8484 English
8485 European Prehistory
8486 Greek
8487 Hallstatt Period
8488 Hittites
8489 Nordic Bronze Age
8490 Roman Antiquity
8491 Situlae
8492
8493 Entity type ['ORG']:
8494 International Study Group
8495 (Kloster already in kont)
8496
8497 Entity type ['PER']:
8498 Hans Hickmann
8499
8500 2. Verbose NER result:
8501 {'entity_group': 'ORG', 'score': 0.46826103, 'word': 'I', 'start': 51, 'end': 52}
8502 {'entity_group': 'ORG', 'score': 0.74872, 'word': 'International Study Group', 'start': 59, 'end': 94}
8503 {'entity_group': 'MISC', 'score': 0.7632287, 'word': 'Archaeology', 'start': 104, 'end': 115}
8504 {'entity_group': 'ORG', 'score': 0.42193357, 'word': 'Kloster', 'start': 119, 'end': 126}
8505 {'entity_group': 'LOC', 'score': 0.9813735, 'word': 'Michaelstein', 'start': 127, 'end': 139}
8506 {'entity_group': 'MISC', 'score': 0.9993983, 'word': 'English', 'start': 222, 'end': 228}
8507 {'entity_group': 'PER', 'score': 0.9997215, 'word': 'Hans Hickmann', 'start': 325, 'end': 338}
8508 {'entity_group': 'LOC', 'score': 0.99916935, 'word': 'Far East', 'start': 344, 'end': 352}
8509 {'entity_group': 'LOC', 'score': 0.99975265, 'word': 'China', 'start': 358, 'end': 363}
8510 {'entity_group': 'LOC', 'score': 0.99978399, 'word': 'Japan', 'start': 365, 'end': 378}
8511 {'entity_group': 'LOC', 'score': 0.99977285, 'word': 'Indonesia', 'start': 375, 'end': 384}
8512 {'entity_group': 'LOC', 'score': 0.9992627, 'word': 'Middle East', 'start': 398, 'end': 401}
8513 {'entity_group': 'LOC', 'score': 0.99939959, 'word': 'Mesopotamia', 'start': 409, 'end': 418}
8514 {'entity_group': 'LOC', 'score': 0.99959975, 'word': 'Caucasus', 'start': 424, 'end': 432}
8515 {'entity_group': 'LOC', 'score': 0.9985838, 'word': 'Central Asia', 'start': 434, 'end': 440}
8516 {'entity_group': 'LOC', 'score': 0.9841615, 'word': 'Parthians', 'start': 455, 'end': 464}
8517 {'entity_group': 'LOC', 'score': 0.99981724, 'word': 'Near East', 'start': 470, 'end': 479}
8518 {'entity_group': 'LOC', 'score': 0.9994, 'word': 'Levant', 'start': 489, 'end': 495}
8519 {'entity_group': 'LOC', 'score': 0.99970114, 'word': 'Australia', 'start': 497, 'end': 505}
8520 {'entity_group': 'MISC', 'score': 0.839957, 'word': 'Hittites', 'start': 514, 'end': 522}
8521 {'entity_group': 'LOC', 'score': 0.99956673, 'word': 'Latin America', 'start': 526, 'end': 537}
8522 {'entity_group': 'LOC', 'score': 0.99956086, 'word': 'Africa', 'start': 542, 'end': 548}
8523 {'entity_group': 'LOC', 'score': 0.99980809, 'word': 'Egypt', 'start': 550, 'end': 553}
8524 {'entity_group': 'LOC', 'score': 0.9977831, 'word': 'Aegean', 'start': 564, 'end': 578}
8525 {'entity_group': 'MISC', 'score': 0.9978502, 'word': 'Greek', 'start': 572, 'end': 577}
8526 {'entity_group': 'MISC', 'score': 0.9936254, 'word': 'Roman Antiquity', 'start': 582, 'end': 597}
8527 {'entity_group': 'MISC', 'score': 0.9932566, 'word': 'European Prehistory', 'start': 599, 'end': 618}
8528 {'entity_group': 'MISC', 'score': 0.9972518, 'word': 'Nordic Bronze Age', 'start': 628, 'end': 643}
8529 {'entity_group': 'MISC', 'score': 0.9688857, 'word': 'Situlae', 'start': 651, 'end': 658}
8530 {'entity_group': 'MISC', 'score': 0.9836559, 'word': 'Hallstatt Period', 'start': 578, 'end': 600}
8531 {'entity_group': 'MISC', 'score': 0.9978018, 'word': 'Celts', 'start': 703, 'end': 708}
8532 {'entity_group': 'LOC', 'score': 0.9996799, 'word': 'Europe', 'start': 740, 'end': 752}

```

Fig. 10: Example: Log entry biblionumber 29551 (Zenon-ID 000048515)

```

333952  Biblionumber 1233684:
333953  1. Filtered entities
333954
333955  Entity type ['LOC']:
333956  An Son
333957  Ban Non Wat
333958  Bàu Tró
333959  Bền
333960  Bình Đa
333961  Cầu Sắt
333962  Cái Vạn
333963  Cù Lao Rùa
333964  Long An Province
333965  Lum Khao
333966  Lộc Giang
333967  Mán Bạc
333968  Rạch Lá
333969  Rạch Núi
333970  Southeast Asia
333971  Southern Vietnam
333972  Suối Linh
333973  (Vietnam already in Koha)
333974  Xóm Rền
333975  Đa Kái
333976  Đình Ông
333977
333978  Entity type ['MISC']:
333979
333980  Entity type ['ORG']:
333981  (no result)
333982
333983  Entity type ['PER']:
333984  Ban
333985  Ban Chiang
333986  Khok Charoen
333987  Khok Phanom Di
333988  Laang Spean
333989  Nong Nor
333990  Samrong Sen
333991  Tha Kae
333992
333993  2. Verbose NER result:
333994  {'entity_group': 'LOC', 'score': 0.99009556, 'word': 'Southern Vietnam', 'start': 44, 'end': 60}
333995  {'entity_group': 'LOC', 'score': 0.99560773, 'word': 'An Son', 'start': 98, 'end': 104}
333996  {'entity_group': 'LOC', 'score': 0.98493963, 'word': 'An Son', 'start': 133, 'end': 139}

```

Fig. 11: Example: Log entry
biblionumber 1233684
(Zenon-ID 001508878)

```

488  486|89904|Wadi Musa
489  487|89959|Diauehi
490  488|89959|East Anatolian
491  489|89959|Erzurum Museum
492  490|89959|Sinoria
493  491|89960|Balkans
494  492|89960|Konya Plain
495  493|90514|Branchidai
496  494|90514|Gulf of Iasos
497  495|90514|Gulf of Mandalya
498  496|90517|Nemea Valley
499  497|90517|Tsoungiza Hill
500  498|90518|Lake Tabqa
501  499|90518|Upper Tabqa Area
502  500|90520|Stratum VI
503  501|90524|Mycenaean Pylos
504  502|90525|Northeast Greece
505  503|90528|Hellenistic Halos
506  504|90528|Sourpi Plains
507  505|90528|Sourpu
508  506|90531|Pseira Island
509  507|90532|Hagia Photia Cemetery
510  508|90710|Sumeria
511  509|91064|Ksanthos
512  510|91087|Isauria
513  511|91087|Maeander
514  512|91087|Pisidia
515  513|91088|Gulf Sheikhdoms
516  514|91283|Sea of Marmara
517  515|91283|Trojan Plain
518  516|91571|Kizzuwatna

```

Fig. 12: Log Excerpts: Extracted
'LOC' entities that were not found
in iDAI.gazetteer

```

4879  Biblionumber 13260:
4880  1. Filtered entities
4881
4882  Entity type ['LOC']:
4883  (Haustrind of not in gazetteer)
4884  (Kratochvil not in gazetteer)
4885  Mikulcice
4886  (Zdenek not in gazetteer)
4887
4888  Entity type ['MISC']:
4889  English
4890  Russian
4891
4892  Entity type ['ORG']:
4893  (no result)
4894
4895  Entity type ['PER']:
4896  (no result)
4897
466702  Biblionumber 1315322:
466703  1. Filtered entities
466704
466705  Entity type ['LOC']:
466706  Abydos
466707  Old Kingdom
466708  (Osiriskult not in gazetteer)
466709  Umm el - Qaab
466710  Agypten
466711
466712  Entity type ['MISC']:
466713  Osiride
466714
466715  Entity type ['ORG']:
466716  Alten Reich
466717  Die Arbeiten der Jahre
466718  German Archaeological Institute
467358  Biblionumber 1315937:
467359  1. Filtered entities
467360
467361  Entity type ['LOC']:
467362  (Bellapais Abbey not in gazetteer)
467363  Byzantium
467364  (Christendom not in gazetteer)
467365  (Cyprus already in Koha)
467366  Europe
467367  Famagusta
467368  Islam
467369  (Lusignan Cyprus not in gazetteer)
467370  Nicosia
467371
467372  Entity type ['MISC']:
467373  Armenians
467374  Cypriot Gothic

```

Fig. 13: Log Excerpts: False positive 'LOC' results filtered by iDAI.gazetteer query

Conclusions

Framework and usability

[28] As shown, it has become possible to set up such lightweight frameworks to use machine learning technologies with a limited investment of resources.³⁹ Python in general has become a widely used language in the machine learning world. It offers many relevant libraries to process and analyze the results, and most platforms in the field of machine learning allow Python-based interactions. Furthermore, questions regarding the complexity of the code base will be of decreasing importance in the near future, because code creation is becoming a much more manageable task thanks to large language models – even though one must always be aware that a sufficient understanding of the core concepts in general and code base in particular will be inevitable in the future as well.⁴⁰

[29] Nevertheless, major hurdles remain. Apart from the fact that the necessary environment must first be set up, ensuring a certain level of usability involves a considerable amount of effort. Not only must the content be prepared for the pipeline, but the returned results must also be further processed. These are serious obstacles when it comes to rolling out comparable solutions. It is therefore worth considering the extent to which more application-oriented frameworks should be used, even if this limits one's own flexibility. There are some very helpful approaches to facilitate the use of machine learning technologies in a very intuitive way by following a visual programming concept, much like the Orange Data Mining toolkit.⁴¹ Although its strength lies with data analysis and visualization, Natural Language Processing tasks are also coming into focus.⁴²

39 Although BibPip itself is not representative in this context due to the fact that some features incorporated for the test scenario have increased the complexity of the code (logs and so on). Furthermore, some of the options and settings proved to be only of limited use in retrospect as stated above.

40 BibPip is extending the original ttw plugin that was written without support of large language models. Had it been written from the beginning with the help of ai models, the result would certainly be much leaner.

41 see <https://orangedatamining.com>.

42 see the plug-ins / add-ons for Text and Data Mining or Natural Language Processing. (<https://pypi.org/project/orange3-nlp>) that provides a collection of widgets for nlp tasks).

[30] The number of specialized and freely available models that can be run on a local machine is still limited at the moment – especially when standard computer equipment (one of our prerequisites) limits performance – but the number is growing rapidly, because many research institutions and projects publish their results and models on open platforms.⁴³ This means that we can expect significantly improved models to be available very soon.

[31] When it comes to performance, processing larger datasets is of course an issue. But even with a standard computer without a dedicated GPU, representative slices of the dataset can be processed within a reasonable timeframe for testing the main configuration in advance. If we speak of daily and low-level work tasks, such as those in the editorial office which require the repetitive processing of short texts, the performance with desktop machines and standard CPUs can be considered sufficient.

Added value – and conceptual reasoning

[32] But what is the conclusion with respect to the expected added value – especially when taking into account the fact that the machine learning landscape is developing rapidly and major players and the industry are developing new applications and methodologies with considerable effort, specialists and expertise?

[33] If similar, simple lightweight pipelines are used to make suggestions within continuous and repetitive processes, suggestions which require feedback and verification from a user, they can in fact help to reduce the users' effort to a certain extent. Formatting articles or cataloguing books or media might be such use cases. Even if only a certain percentage is covered, search time will be reduced.

[34] Finally, one can also imagine scenarios in which comparable pipelines already assist in structuring information at the moment when it is entered into special applications.

[35] Nevertheless, the challenge is to integrate them smoothly into preexisting ways of working, both technically and organizationally, as too many breaks in current practices undermine the gained efficiency. This is not yet necessarily foreseeable for use cases like editorial processes, which also involves a larger number of stakeholders. For a retrospective use on larger datasets, e. g. in a library scenario, the effort involved must be weighed against the added value, but since entries without keywords can be filtered easily, this could at least be a good starting point for gradually enriching entries. In any case such pipelines can help to alleviate bottleneck situations, caused usually by the need for human intervention.⁴⁴

[36] Finally, however, attention shall be drawn to another aspect: What is probably even more important is the fact that even such low-level projects help to improve one's own concepts (how to deal with multilingualism, the broad

⁴³ See e. g. models of the Berlin State Library (<https://huggingface.co/SBB>), the Annif models (<https://huggingface.co/NatLibFi>), or models by national initiatives like the Swiss AI Apertus LLM collection (<https://www.swiss-ai.org/apertus>) to name just a few.

⁴⁴ I thank Lisa Steinmann for this aspect.

ranges of specialist disciplines and so on), as the results may reveal insufficiencies or inconsistencies in one's own information management immediately. This raises awareness about the need to equip information resources at generation with relevant information or encoding so that they can be used for machine processing. This applies to content, form, and structure, especially with respect to the use or integration of authority data and the requirements that arise within this context, as we have seen here e. g. with respect to lacking multilingualism, completion, precision and so on. Meaningful use of machine learning applications is only possible if the reference data is of sufficient quality and uniformity.

[37] For these reasons and in view of the challenges that new technologies pose for work units or teams in the field of information infrastructure, it is important that they engage actively with them. Even though this simple test dealt with questions relating to the handling of text, the approaches and procedures with regard to multimodal applications are very similar.

[38] In any case, it is hoped that this article can show the value of even low-threshold scenarios – and that it will as well encourage people from those teams to explore the possibilities and take advantage of the opportunities.

References

- Balnaves et al. 2025** E. Balnaves – L. Bultrini – A. Cox – R. Uzwyszyn (ed.), *New Horizons in Artificial Intelligence in Libraries* (Berlin 2025), <https://doi.org/10.1515/9783111336435>
- Baumeister 2022** P. Baumeister, Das neue maschinenlesbare Zeitschriftenmodell des Deutschen Archäologischen Instituts. Ein Werkstattbericht, *FdAI* 2022/1, § 1–67, <https://doi.org/10.34780/cf6e-prcf>
- Ben Tahar et al. 2021** S. Ben Tahar – Ph. von Rummel – K. Mansel – H. Möller – T. Mukai – M. Aoudi – M. Dinies – Th. Lappi – J. Peters – S. Trixl – S. Büchner, Henchir Bourgou (Djerba, Tunesien). Stratigraphie und Fundvorlage einer Sondage im Zentrum der antiken Siedlung (8. Jh. v. Chr. – 2. Jh. n. Chr.), *AA* 2021/2, § 1–178, <https://doi.org/10.34780/fo5j-59fd>
- Brandsen 2022** A. Brandsen, Digging in Documents: Using Text Mining to Access the Hidden Knowledge in Dutch Archaeological Excavation Reports (2022, February 15), <https://hdl.handle.net/1887/3274287>
- Brandsen et al. 2022** A. Brandsen – S. Verberne – K. Lambers – M. Wansleben, Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain, *Journal on Computing and Cultural Heritage* 15, 3, Article 51 (September 2022), <https://doi.org/10.1145/3497842>
- Büchel et al. 2023** J. Büchel – J. F. Engler – A. Mertens, Gesuchte Datenkompetenzen in Deutschland, in: *Vierteljahresschrift zur empirischen Wirtschaftsforschung*, Jahrgang 50, 2023, Nr. 2, 3–17, <https://doi.org/10.2373/1864-810X.23-02-01>
- Dreyer 2025** M. Dreyer, Results of the ZKI Top Trends Survey Conducted by the ZKI Strategy and Organisation Working Group for the Year 2025, <https://doi.org/10.5281/zenodo.14904518>
- Dreyer 2026** M. Dreyer, ZKI Top Trends Survey 2026, <https://doi.org/10.5281/zenodo.18520259>
- Jurowetzki et al. 2025** R. Jurowetzki – D. S. Hain – K. Wirtz et al., The Private Sector is Hoarding AI Researchers: What Implications for Science?, *AI & Soc* 40, 4145–4152 (2025), <https://doi.org/10.1007/s00146-024-02171-z>
- Kasprzik 2023** A. Kasprzik, Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW – ein Status- und Erfahrungsbericht, *o-bib – Das offene Bibliotheksjournal* 10, 1, (2023), 1–13, <https://doi.org/10.5282/o-bib/5903>
- Kasprzik 2025** A. Kasprzik, Transferring Applied Machine Learning Research into Subject Indexing Practice, in: Balnaves et al. 2025, 199–212, <https://doi.org/10.1515/9783111336435-015>
- Kose et al. 2022** A. Kose – B. Engels – M. Trümper, Die Basilike Stoa an der Agora von Thera. Rekonstruktion ihrer Entwicklung auf Basis stratigraphischer Grabungen, *AA* 2022/1, § 1–109, <https://doi.org/10.34780/tiq0-a22r>
- Maslej et al. 2024** N. Maslej – L. Fattorini – R. Perrault – V. Parli – A. Reuel – E. Brynjolfsson – J. Etchemendy – K. Ligett – T. Lyons – J. Manyika – J. C. Niebles – Y. Shoham – R. Wald – J. Clark, *The AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024, <https://doi.org/10.48550/arXiv.2405.19522>
- Maslej et al. 2025** N. Maslej – L. Fattorini – R. Perrault – Y. Gil – V. Parli – N. Kariuki – E. Capstick – A. Reuel – E. Brynjolfsson – J. Etchemendy – K. Ligett – T. Lyons – J. Manyika – J. C. Niebles – Y. Shoham – R. Wald – T. Walsh – A. Hamrah – L. Santarlasci – J. Betts Lotufo – A. Rome – A. Shi – S. Oak, *The AI Index 2025 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025, <https://doi.org/10.48550/arXiv.2504.07139>
- Menzel et al. 2021** S. Menzel – H. Schnaitter – J. Zinck – V. Petras – Cl. Neudecker – K. Labusch – El. Leitner – G. Rehm, Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten, in: M. Franke-Maier – A. Kasprzik – A. Ledl – H. Schürmann, *Qualität in der Inhaltserschließung* (Berlin 2021) 229–258, <https://doi.org/10.1515/9783110691597-012>
- Osborne et al. 2024** C. Osborne – J. Ding – H. R. Kirk, The AI Community Building the Future? A Quantitative Analysis of Development Activity on Hugging Face Hub, *Journal of Computational Social Science* 7, 2024, 2067–2105, <https://doi.org/10.1007/s42001-024-00300-8>
- Pol et al. 2024** U. R. Pol – P. S. Vadar – T. T. Moharekar, Hugging Face: Revolutionizing AI and NLP, *International Journal for Research in Applied Science and Engineering Technology* 12(8), 2024, 1121–1124, <https://doi.org/10.22214/ijraset.2024.64023>

Suryani et al. 2025 M. A. Suryani – S. Karmakar – B. Mathiak, Exploration of Hugging Face Models by Heterogeneous Information Network and Linking Across Scholarly Repositories, in: L. M. Aiello – T. Chakraborty – S. Gaito (eds), Social Networks Analysis and Mining. ASONAM 2024. Lecture Notes in Computer Science, vol 15213 (2025), https://doi.org/10.1007/978-3-031-78548-1_27

Tzochev 2021 C. Tzochev, The Architecture of the 4th Century B.C. Monumental Tomb at Starosel, AA 2021/2, § 1–120, <https://doi.org/10.34780/d5bi-h53t>

Weers 2025 B. S. Weers, Automatisierte Inhaltserschließung an der Bibliothek des Max-Planck-Instituts für Mathematik in den Naturwissenschaften (Leipzig 2025), <http://doi.org/10.33968/9783966270786-00>

Source of Illustrations

Fig. 1: Graphic: Peter Baumeister

License: CC BY 4.0

Fig. 2: Screenshot: Peter Baumeister

License: CC0

Fig. 3: Screenshot: Peter Baumeister

License: CC0

Fig. 4: Screenshot: Peter Baumeister

License: CC0

Fig. 5: Screenshot: Peter Baumeister

License: CC0

Fig. 6: Screenshot: Peter Baumeister

License: CC0

Fig. 7: Screenshot: Peter Baumeister

License: CC0

Fig. 8: Screenshot: Peter Baumeister

License: CC0

Fig. 9: Screenshot: Peter Baumeister

License: CC0

Fig. 10: Screenshot: Peter Baumeister

License: CC0

Fig. 11: Screenshot: Peter Baumeister

License: CC0

Fig. 12: Screenshot: Peter Baumeister

License: CC0

Fig. 13: Screenshot: Peter Baumeister

License: CC0

ABSTRACT (DE)


In dem Artikel wird ein Testszenario beschrieben, das den sehr niedrigschwelligen Einsatz von sogenannter Named Entity Recognition (NER) in Abstracts in iDAI.bibliography – dem Katalog der DAI-Bibliotheken – beschreibt. Vorrangiges Ziel des Beitrags ist es, Arbeitseinheiten bzw. Teams aus dem Bereich Informationsinfrastruktur, die bisher nur eingeschränkt mit Technologien wie dem Maschinellen Lernen in Berührung kamen, zu ermutigen, sich mit diesen Technologien auseinanderzusetzen und praktische Einblicke hinsichtlich deren Möglichkeiten und Grenzen zu gewinnen. Der Ansatz nutzt eine schlanke Pipeline, die auf frei verfügbaren Modellen, einer einfachen Codebasis, Standardhardware und einer urheberrechtskonformen Vorgehensweise basiert und die zeigt, wie automatisierte Verarbeitung händischen Aufwand reduzieren und die Qualität der Einträge bis zu einem gewissen Grad verbessern kann. Da der Fokus auf pragmatischen Einsatzmöglichkeiten hinsichtlich der Bewältigung einfacher Alltagsaufgaben liegt, kann ein solches Szenario bei der Entwicklung eigener diesbezüglicher Vorhaben und Konzepte helfen, auch hinsichtlich eines besseren Verständnisses der technischen Machbarkeiten.

SCHLAGWORTE

Named Entity Recognition,
Metadatenanreicherung, 'Hugging Face
Transformers' pipeline

AUTHORS

 Peter Baumeister

 German Archaeological Institute, Head
Office

METADATA

DOI (url):

<https://doi.org/10.34780/s6tar918>

Title:

A Test Report on a Lightweight Pipeline for
Named Entity Recognition

Subtitle:

Using Machine Learning Models for Metadata
Enrichment on Abstracts in iDAI.bibliography

Volume:

Faszikel 2026

Created on:

2026-04-23

Copyright:

© 2026 Deutsches Archäologisches Institut

License:

<https://creativecommons.org/licenses/by/4.0/>

The text of this article is licensed under the license stated above. Unless otherwise indicated, figures and other non-text content are not covered by this license. For reuse of such materials, see specific license terms for each individual figure.

IMPRINT

Forum for Digital Archaeology and
Infrastructure

ISSN:

2748-8861

Publisher:

Deutsches Archäologisches Institut
Zentrale
Podbielskiallee 69-71
14195 Berlin
Deutschland

Journal-Editors:

Benjamin Ducke, Friederike Fless, Fabian
Riebschläger, Peter Baumeister, Lisa
Steinmann

Publishing-Editor:

FdAI, DAI Zentrale, Arbeitsstab
Kommunikation

Editing:

Elianor Sket, Marcel Riedel

Typesetting:

Elianor Sket, Marcel Riedel

Layout:

Corporate Design: LMK Büro für
Kommunikationsdesign, Berlin

This PDF was created with [JatSinForm](#).