

Thomas Guignard

Indexation automatique de documents avec

The logo for annif, featuring the word 'annif' in a bold, lowercase sans-serif font. The letter 'a' is white with a dark blue outline, and the letter 'i' has a small green dot above it. The rest of the letters are dark blue.

Conférence LibreABC

9 septembre 2025



FINNA.FI

annif

Contexte technique (et historique)

Digital Assistant
(ZB Zürich)

DA-3
(Allemagne)

extreme multilabel classification (XMLC)

annif
(prototype)

annif
v.0.45 (Omikuji)

annif
v.1.0

annif
v.1.1 (😊)

bag-of-words

word2vec

Parabel/Omikuji

2000

2013

2016

2017

2018

2019

2020

2021

2022

2023

2024

GPT

GPT-2

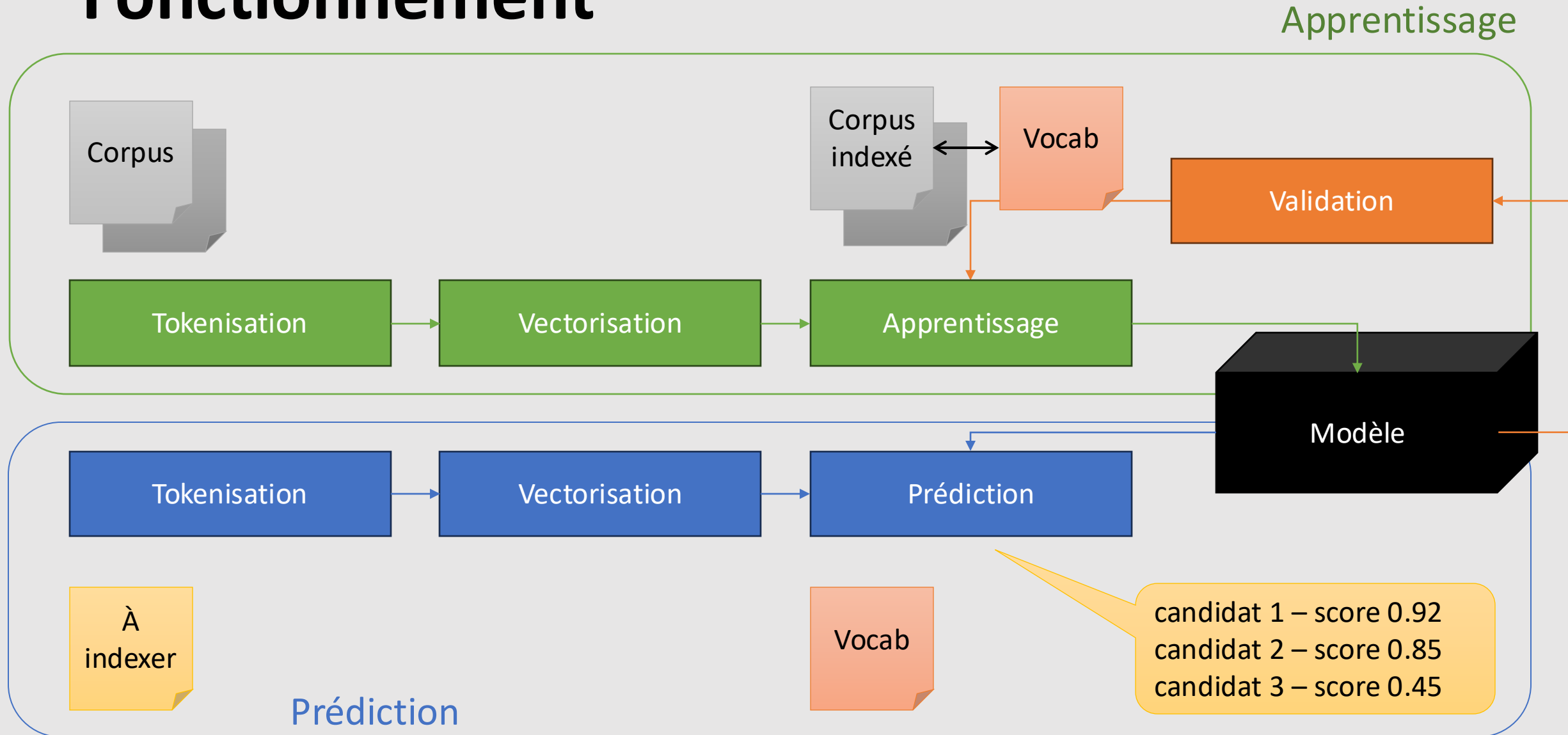
GPT-3

ChatGPT

BERT

Attention is All You Need
transformers

Fonctionnement



En pratique : Apprentissage

1. Configuration

```
[stw-tfidf-en]  
name=STW TFIDF project  
language=en  
backend=tfidf  
vocab=stw  
analyzer=snowball(english)
```

```
[stw-mllm-en]  
name=STW MLLM project  
language=en  
backend=mllm  
vocab=stw  
analyzer=snowball(english)
```

```
[stw-ensemble-en]  
name=STW ensemble project  
language=en  
backend=ensemble  
vocab=stw  
sources=stw-tfidf-en,stw-mllm-en:2
```

2. Chargement du vocabulaire

```
annif load-vocab stw datasets/stw-skos.ttl
```

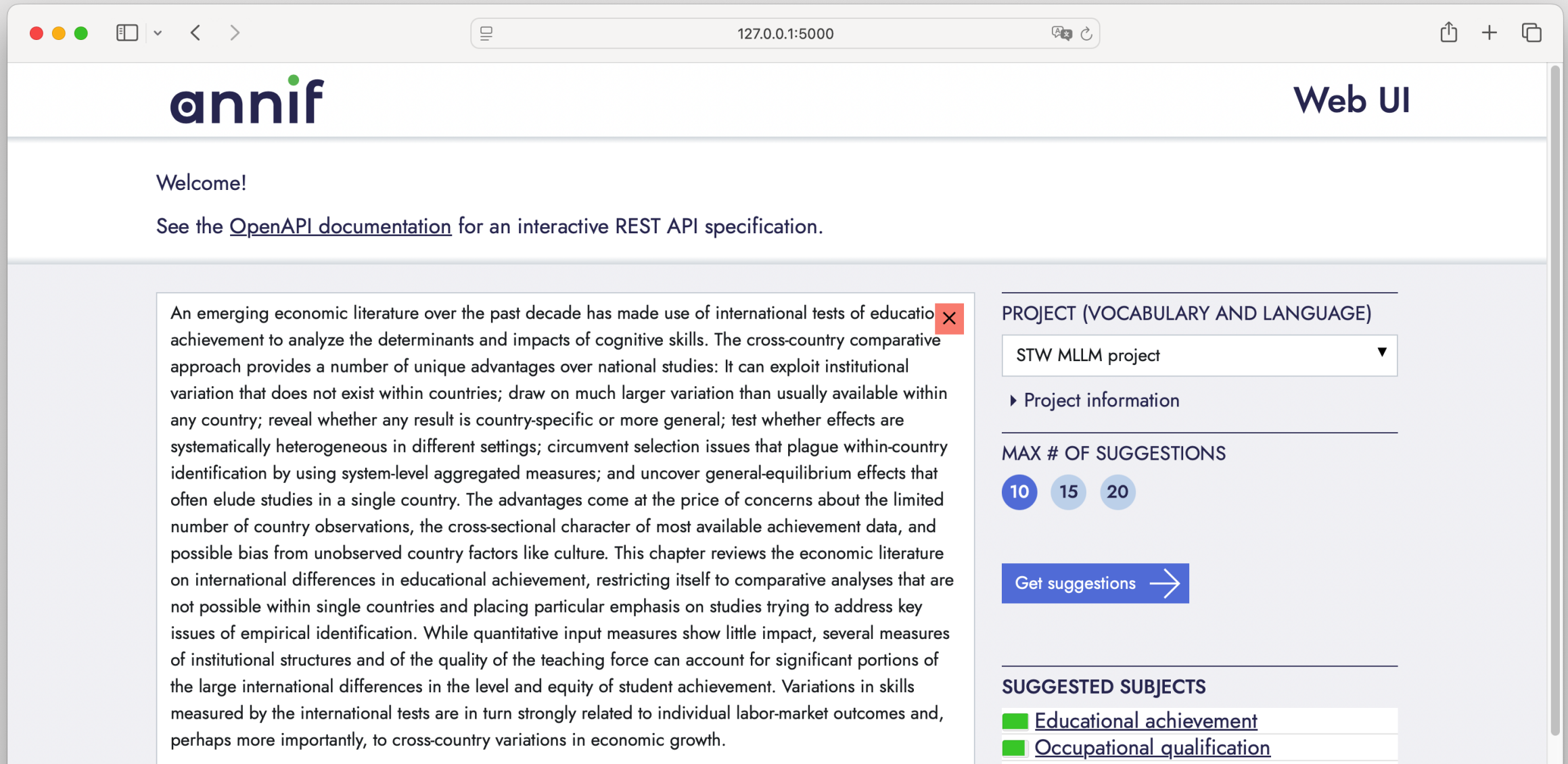
3. Apprentissage

```
annif train stw-tfidf-en datasets/stw-econbiz.tsv.gz
```

4. Validation

```
annif eval stw-tfidf-en datasets/stw-econbiz-test
```

Prédiction : Web



The screenshot shows a web browser window displaying the 'annif' Web UI. The browser's address bar shows '127.0.0.1:5000'. The page has a header with the 'annif' logo on the left and 'Web UI' on the right. Below the header, a 'Welcome!' message is followed by a link to the 'OpenAPI documentation'. The main content area is split into two columns. The left column contains a text block about international tests of education achievement, with a red 'X' icon in the top right corner. The right column features a 'PROJECT (VOCABULARY AND LANGUAGE)' section with a dropdown menu set to 'STW MLLM project' and a 'Project information' link. Below this is a 'MAX # OF SUGGESTIONS' section with three radio buttons for '10', '15', and '20', and a 'Get suggestions' button. At the bottom, a 'SUGGESTED SUBJECTS' section lists 'Educational achievement' and 'Occupational qualification' with green checkboxes.



Web UI

Welcome!

See the [OpenAPI documentation](#) for an interactive REST API specification.

An emerging economic literature over the past decade has made use of international tests of education achievement to analyze the determinants and impacts of cognitive skills. The cross-country comparative approach provides a number of unique advantages over national studies: It can exploit institutional variation that does not exist within countries; draw on much larger variation than usually available within any country; reveal whether any result is country-specific or more general; test whether effects are systematically heterogeneous in different settings; circumvent selection issues that plague within-country identification by using system-level aggregated measures; and uncover general-equilibrium effects that often elude studies in a single country. The advantages come at the price of concerns about the limited number of country observations, the cross-sectional character of most available achievement data, and possible bias from unobserved country factors like culture. This chapter reviews the economic literature on international differences in educational achievement, restricting itself to comparative analyses that are not possible within single countries and placing particular emphasis on studies trying to address key issues of empirical identification. While quantitative input measures show little impact, several measures of institutional structures and of the quality of the teaching force can account for significant portions of the large international differences in the level and equity of student achievement. Variations in skills measured by the international tests are in turn strongly related to individual labor-market outcomes and, perhaps more importantly, to cross-country variations in economic growth.

PROJECT (VOCABULARY AND LANGUAGE)

STW MLLM project

► Project information

MAX # OF SUGGESTIONS

10 15 20

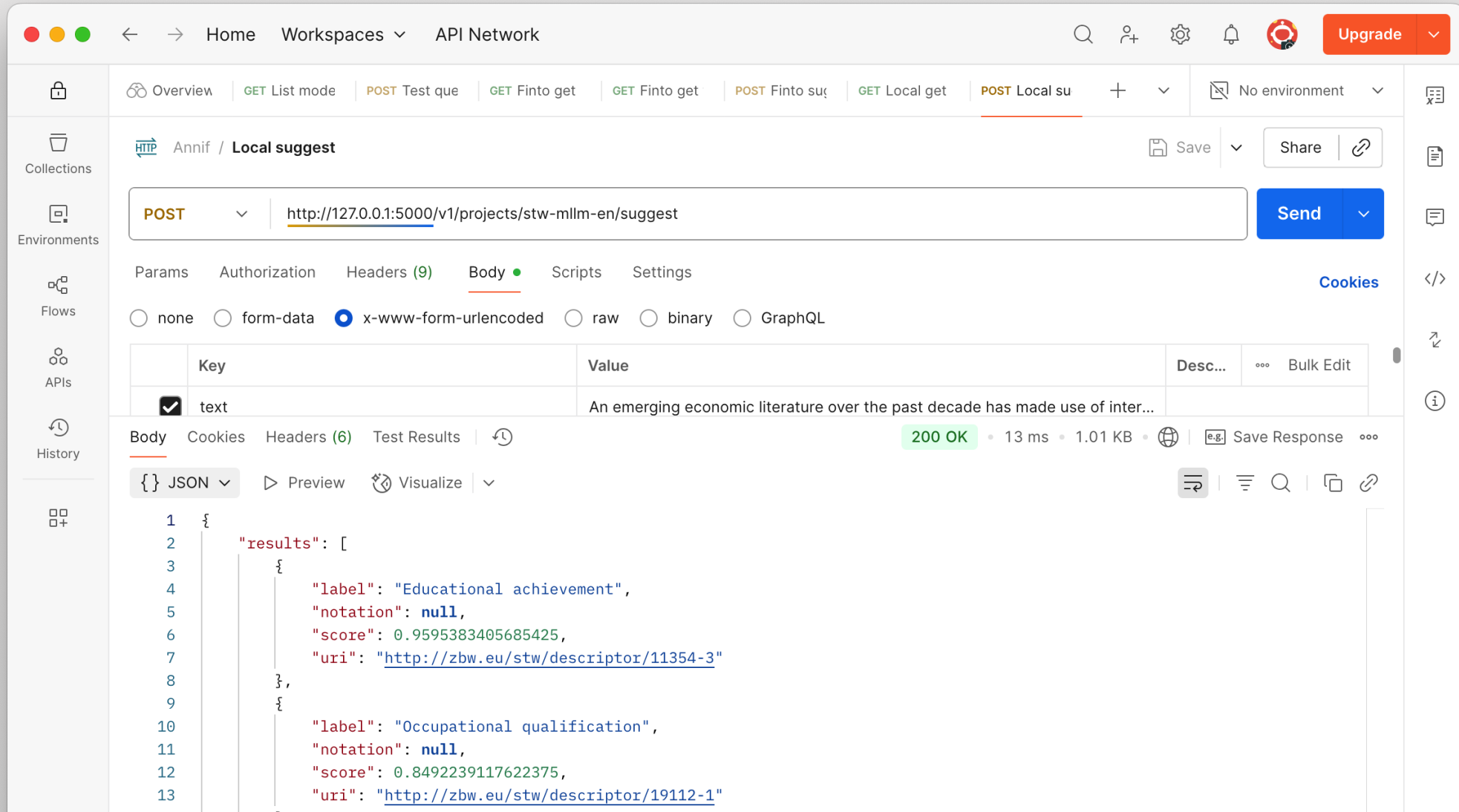
Get suggestions →

SUGGESTED SUBJECTS

☒ [Educational achievement](#)

☒ [Occupational qualification](#)

Prédiction : API



The screenshot shows the Annif API interface. The top navigation bar includes "Home", "Workspaces", and "API Network". The left sidebar contains icons for Collections, Environments, Flows, APIs, and History. The main area displays a POST request to the endpoint `http://127.0.0.1:5000/v1/projects/stw-mlm-en/suggest`. The request body is set to "x-www-form-urlencoded" and contains a text value: "An emerging economic literature over the past decade has made use of inter...". The response is a 200 OK status with a 13 ms response time and 1.01 KB of data. The response body is shown in JSON format, containing a list of results with labels, notations, scores, and URIs.

Key	Value	Desc...	Bulk Edit
☒ text	An emerging economic literature over the past decade has made use of inter...		

```

1  {
2    "results": [
3      {
4        "label": "Educational achievement",
5        "notation": null,
6        "score": 0.9595383405685425,
7        "uri": "http://zbw.eu/stw/descriptor/11354-3"
8      },
9      {
10       "label": "Occupational qualification",
11       "notation": null,
12       "score": 0.8492239117622375,
13       "uri": "http://zbw.eu/stw/descriptor/19112-1"

```

Prédiction : CLI

```
Annif-tutorial — -zsh — 88x29

Last login: Mon Sep  1 08:27:37 on ttys000
[(base) thomas@Corcovado ~ % cd /Users/thomas/Documents/dev/playground/Annif-tutorial ]
(base) thomas@Corcovado Annif-tutorial % conda activate annif
[(annif) thomas@Corcovado Annif-tutorial % annif suggest -L en -l 5 stw-ml1m-en data-sets]
[/stw-zbw/docs/validate/10008738211.txt ]
Suggestions for data-sets/stw-zbw/docs/validate/10008738211.txt
<http://zbw.eu/stw/descriptor/11354-3> Educational achievement 0.8382
<http://zbw.eu/stw/descriptor/11378-3> Students 0.5875
<http://zbw.eu/stw/descriptor/10414-2> Economic growth 0.3578
<http://zbw.eu/stw/descriptor/19112-1> Occupational qualification 0.3266
<http://zbw.eu/stw/descriptor/19037-3> Estimation 0.2791
(annif) thomas@Corcovado Annif-tutorial %
```

Démonstration !

Cas d'usage actuels

Via **fintoai**



Standalone



Getty



National Library
of Sweden



BIBLIOTEKA
NARODOWA



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

etc.

Et bientôt ...



Conclusion

- Flexible, open source
- Combine plusieurs approches de traitement du langage naturel
- Facile à installer (bibliothèque Python : `pip install annif`)
- Facile à intégrer (API, SaaS)
- Multilingue, s'adapte à divers corpus & référentiels
- Déjà adopté en production par plusieurs bibliothèques
- Complémentaire aux LLMs

<https://github.com/NatLibFi/Annif>

<https://doi.org/10.5281/zenodo.2578948>

thomas@tgconsulting.ca

@timtom@code4lib.social

Références

1. Alammar, J., & Grootendorst, M. (2024). *Hands-on large language models: Language understanding and generation* (First edition). O'Reilly Media, Inc. <https://www.oreilly.com/library/view/-/9781098150952/>
2. Beckmann, R., Hinrichs, I., Janßen, M., Milmeister, G., & Schäuble, P. (2019). Der Digitale Assistent DA-3: Eine Plattform für die Inhaltsschließung. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, 6(3), 1–20. <https://doi.org/10.5282/o-bib/2019H3S1-20>
3. *Indexation RAMEAU assistée par IA. Expérimentation réalisée avec 12 établissements testeurs. Rapport final.* (2025). Agence Bibliographique de l'Enseignement Supérieur. <https://abes.fr/etude-experimentation-indexation-rameau-assistee-par-ia/>
4. Inkinen, J., Lehtinen, M., & Suominen, O. (2025, March 24). *Annif Users Survey: Understanding Usage and Challenges*. Kansalliskirjasto. <https://www.doria.fi/handle/10024/190930>
5. Malits, A., & Schäuble, P. (2014). Der Digitale Assistent: Halbautomatisches Verfahren der Sacherschließung in der Zentralbibliothek Zürich: The Digital Assistant: a semi-automated system for subject cataloguing in the Zentralbibliothek Zurich. *ABI Technik*, 34(3–4), 132–143. <https://doi.org/10.1515/abitech-2014-0024>
6. Poley, C., Uhlmann, S., Busse, F., Jacobs, J.-H., Kähler, M., Nagelschmidt, M., & Schumacher, M. (2025). Automatic Subject Cataloguing at the German National Library. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 35(1), 1–29. <https://doi.org/10.53377/lq.19422>
7. Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., & Varma, M. (2018). Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 993–1002. <https://doi.org/10.1145/3178876.3185998>
8. Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 29(1), 1–25. <https://doi.org/10.18352/lq.10285>
9. Suominen, O., Inkinen, J., & Lehtinen, M. (2025). *Annif at SemEval-2025 Task 5: Traditional XMTC augmented by LLMs* (No. arXiv:2504.19675). arXiv. <https://doi.org/10.48550/arXiv.2504.19675>
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. <https://doi.org/10.48550/arXiv.1706.03762>