

entity-fishing

a DARIAH entity recognition and disambiguation service

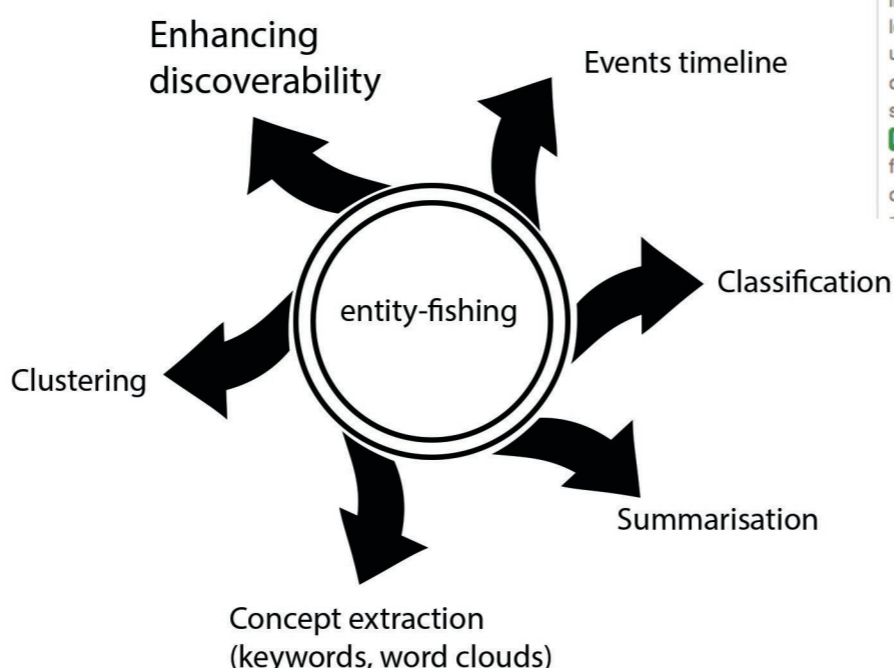
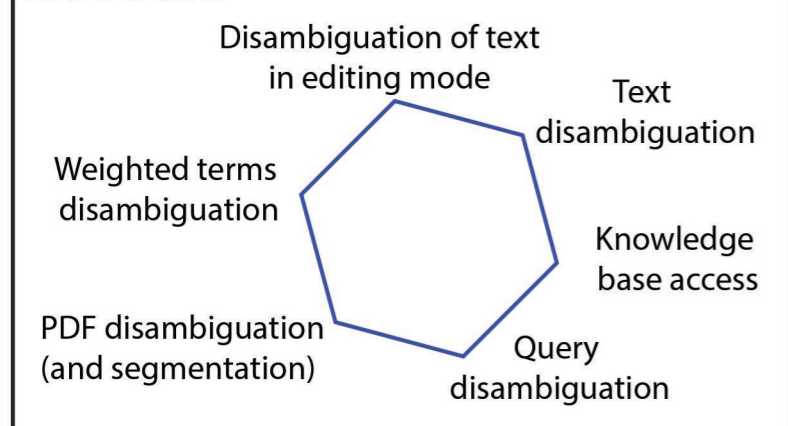
Luca Foppiano, Laurent Romary - ALMAnaCH - Inria Paris

Introduction

Entity extraction and disambiguation is the task of determining the identity of entities mentioned in a text against a knowledge base.

entity-fishing implements entity resolution and disambiguation against Wikipedia and Wikidata.

How is it used?



CHARLES PÉGU poète et **ÉPISTÉMOLOGUE** de l'histoire et de la politique. Au début du **XXE SIÈCLE**, alors que l'histoire universitaire s'efforçait de mettre en œuvre des méthodes de travail scientifiques, **PÉGU** mettait en cause des professeurs d'histoire de la **BORBONNE** et dénonçait une histoire qui lui paraissait étriquée, besogneuse, sans souffle. Outre une écriture terne, il reprochait également à cette école dite méthodique d'instiller une interprétation politisée de l'histoire. Nommément désigné à la vindicte des lecteurs des **CAHIERS DE LA QUINZAINÉ**, le professeur d'**HISTOIRE MODERNE** **CHARLES LANGLOIS** fit particulièrement les frais des diatribes péguystes. Le succès de l'**HISTOIRE MÉTHODIQUE** et la virulence des termes employés par **PÉGU** contribuèrent à discréditer sa critique, tandis que sa rupture avec l'**INTERNATIONALISME** socialiste et sa conversion au **CATHOLICISME** firent peser la suspicion sur sa démarche, considérée comme relevant de l'**APOLOGÉTIQUE**, ainsi que sur le bien fondé de ses interprétations. Si bien que, s'il est aujourd'hui admiré (ou détesté) comme **ÉCRIVAIN**, son apport à l'histoire est tenu pour négligeable. Or l'histoire tient une place importante dans l'œuvre de **CHARLES PÉGU**. Cet homme, qui vécut chaque moment de sa vie avec passion, de l'engagement **DREYFUSARD** à la **CONVERSION RELIGIEUSE**, accorda une grande importance à l'émotion que suscitait le passé et s'exprima souvent sur la nature de l'histoire et le rôle qu'elle devait tenir dans la société. Il ne nous paraît donc pas inopportun de dépasser une lecture convenue de **PÉGU** et de relire ses textes relatifs à l'histoire afin de comprendre sa démarche. Nous distinguerons trois facettes de **PÉGU HISTORIEN**, qui se complètent et s'éclairent mutuellement : il fut un critique de l'histoire-science, un poète du passé et un **PHILOSOPHE** méditant sur le temps. Ces trois aspects, étroitement entremêlés, sont fréquemment abordés par **PÉGU** même si, pour la commodité de l'exposé, nous nous contentons de dégager les traits saillants de certaines œuvres. Nous avons eu l'illusion, dans un

PÉGU

Type: **LOCATION**

Normalized: Charles Pégu

Domains: **Biology, Administration, Military**

conf: 0.3749



Charles Pierre Pégu, né le à Orléans et mort le à Villeroy (Seine-et-Marne), est un écrivain, poète et essayiste français. Il est également connu sous les noms de plume de Pierre Deloire et Pierre Baudouin.

sex or gender	Q6581097
place of birth	Orléans
VIAF ID	7395351
Library of Congress authority ID	n80038479
authority ID	
BnF ID	11918933f
SUDOC authorities	027062295
ISNI	0000 0001 2276 072X
ISNI	0000 0001 2276 072X
ISNI	0000 0001 2276 072X

entity-fishing within DARIAH

1. Available as a service in the whole DARIAH infrastructure
2. Flexible infrastructure
3. Integrated in OPERAS Open Access publishing platforms (financed by HIRMEOS, H2020 European project)
4. Ready for supporting more DH projects
5. Access to free Open Data via standardized REST API

Open Data

Entity fishing API provides streamlined and standardised access to the biggest and most complete Open Knowledge Base:



Coverage: 14M (en), 3.5M (de), 3.5M (fr), 2.2M (it), 3.3M (es) pages
Licence CC-BY



Coverage: 37M pages, 500+M statements
Licence CC-0

The year is 1935. The event, at least for literature in **Khasi**, is momentous. A man **diminutive** in stature but with a voice that cradled the vast soul of his people had decided to do what he knew best. He completed a classic in **Khasi** literature and the **Shillong Printing Works** published *The Old Days of the Khasis* (*Ki Sngi Barim U Hynñiew Trep*).³ **Soso Tham** came in from the wilderness to carve in words the identity of his people—he made us see, he made us hear, he made us feel and he made us fear.

In a land still under **British** rule **this** legendary schoolteacher expressed a weary frustration with the **English** texts he had taught his students year after year. He declared that from now on “he would do it himself”. And so he did. An **oral culture** for whom, in 1841, **Thomas Jones** of the **Welsh Presbyterian Mission** had devised a script, now had a **scribe** whose work expresses a profound love for his homeland and an **unwavering** pride in the history of his tribe—a history kept alive in

cial customs and in **fables** and legends handed down by storytellers.

refused to believe that a people with no evidence of a **was** without foundation or worth. He set out to compile **memories** of the ancient past—*ki sngi barim*—presenting

ideas in the Introduction have appeared in articles I submitted to the **es** (Meghalaya) and in a paper entitled ‘**Surviving Change**’ which I **t** a conference organised by **Lady Keane College, Shillong**, in August

in **Soso Tham**’s Preface to *Ki Sngi Barim U Hynñiew Trep*. **Shillong** in 1936.

KHASI

Type: **PERSON**

Normalized: **Khasi people**

Domains: **Sociology**

conf: 0.6636

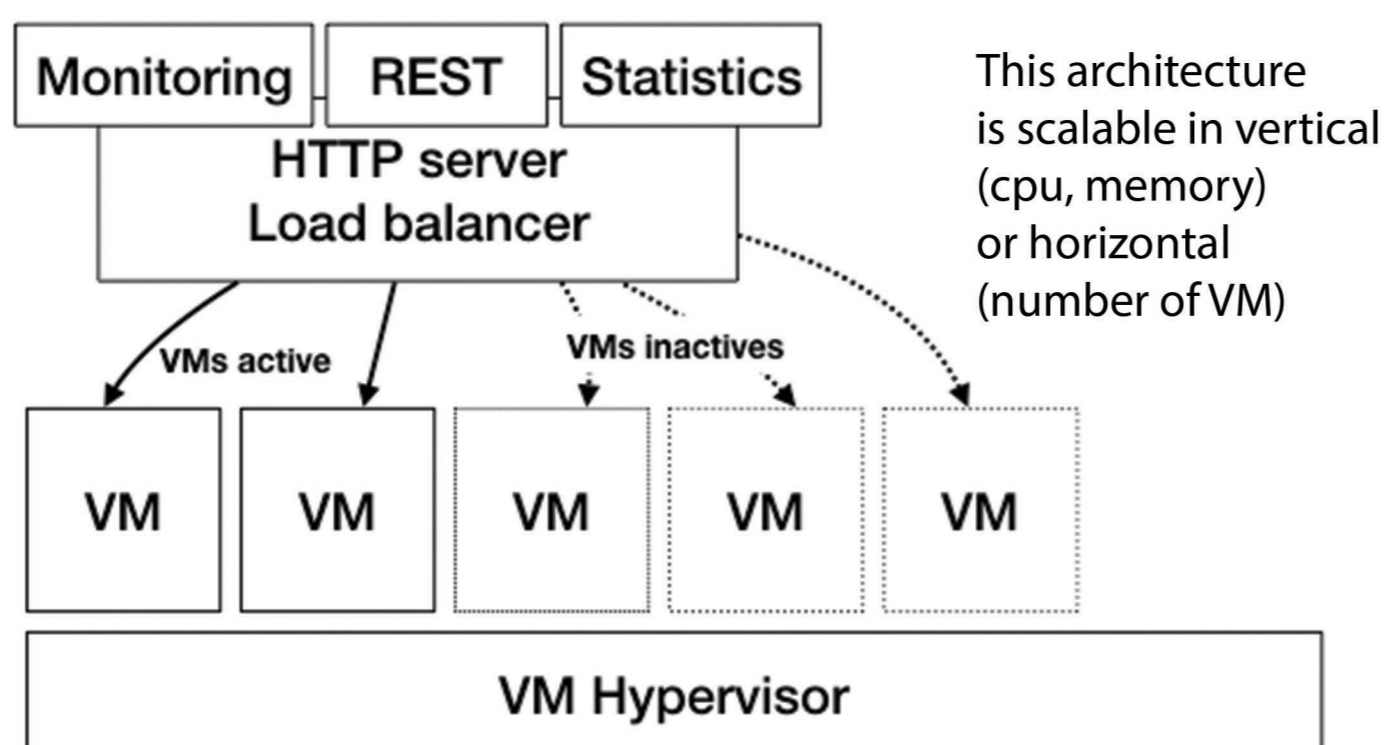


The **Khasi people**, endonym, (“Children of the Seven Huts”), are an indigenous tribe, the majority of whom live in the State of **Meghalaya** which is in the north eastern part of **India**, with a significant population in the border areas of the neighbouring state of **Assam**, and in certain parts of **Bangladesh**. The Khasi people are the native people of **Meghalaya** and forms the majority about 50.2% or 1.72 million of the state population. Their language, **Khasi**, is categorised as the northernmost **Austroasiatic** language. Primarily an oral language, the **Bengali script** was used to write **Khasi** after the arrival of Christian missionaries. Particularly significant in this regard was a Welsh evangelist, **Thomas Jones**, who transcribed the Khasi language into the Roman script. The Khasi people form the majority of the population of the eastern part of Meghalaya, and is the state’s largest community. Though around 85% of the Khasi populace have embraced Christianity, a substantial minority of the Khasi people still follow and practice their age old indigenous religion, which is known as “Ka Niam Khasi” and it is their belief that the

Why entity-fishing?

1. Generic approach but extensible to solve specific problems
2. Open Source software
3. 100% Machine Learning based
4. Native support for processing and structuring PDFs
5. Fast, resilient, designed for processing large amount of data
6. No required expertise in knowledge engineering
7. Multi language: en, it, fr, de, es
8. PDF processing throughput at 1.2 pages/seconds

Architecture



This architecture is scalable in vertical (cpu, memory) or horizontal (number of VM)

Future work

1. Support more languages
2. Better disambiguation results (accuracy and speed)
3. Integration of additional mention recognisers for DH
4. Integrating data from knowledge bases in Social Sciences and Humanities
5. Application for disambiguation of authors and affiliations

Acknowledgements

- We would like to thank:
- Patrice Lopez, the creator of entity-fishing, grobid and many other tools
 - Our colleagues at the European project HIRMEOS
 - Huma-num, for hosting and supporting the project

entity-fishing is **Open Source**, licence Apache 2.0



<http://github.com/kermitt2/entity-fishing>



<http://nerd.readthedocs.io>



<http://nerd.huma-num.fr>



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731102.