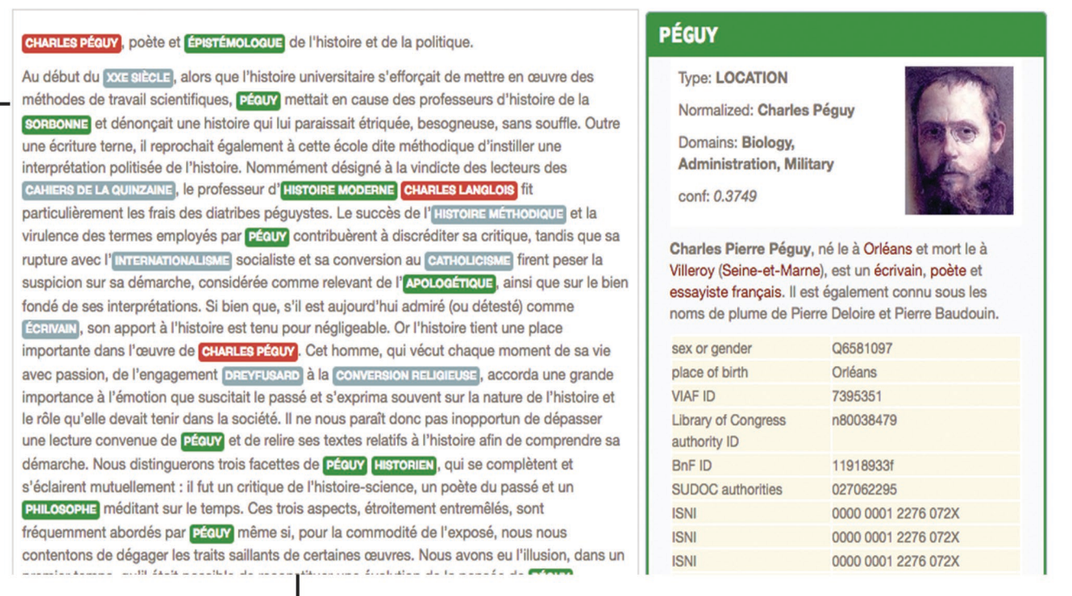
entity-fishing a DARIAH entity recognition and disambiguation service

Luca Foppiano, Laurent Romary - ALMAnaCH - Inria Paris

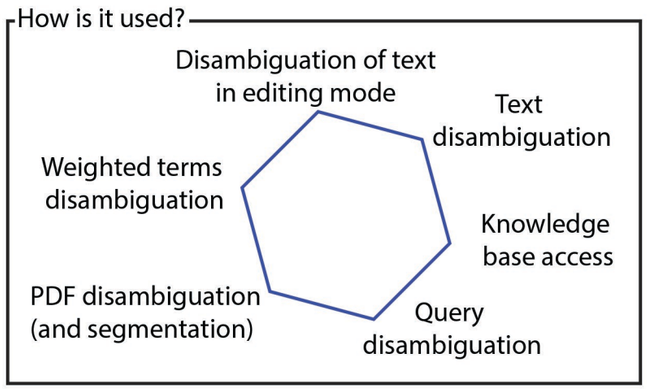
Introduction

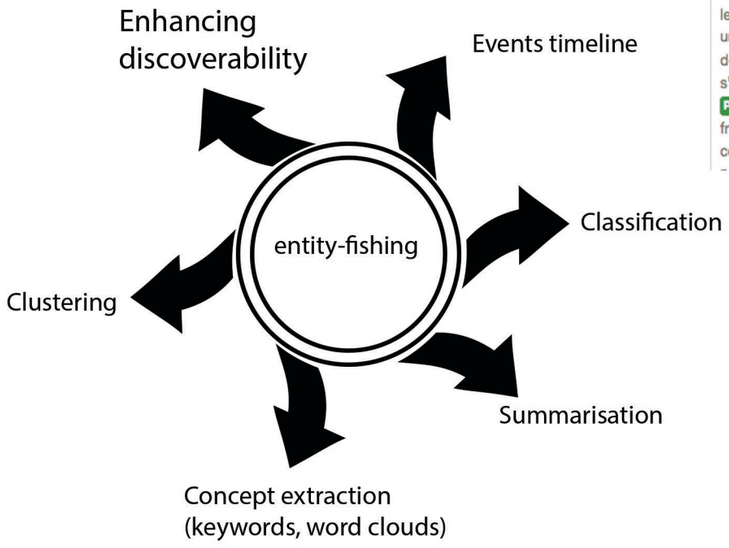
Entity extraction and disambiguation is the task of determining the identity of entities mentioned in a text against a knowledge base.

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



How is it used?





entity-fishing within DARIAH

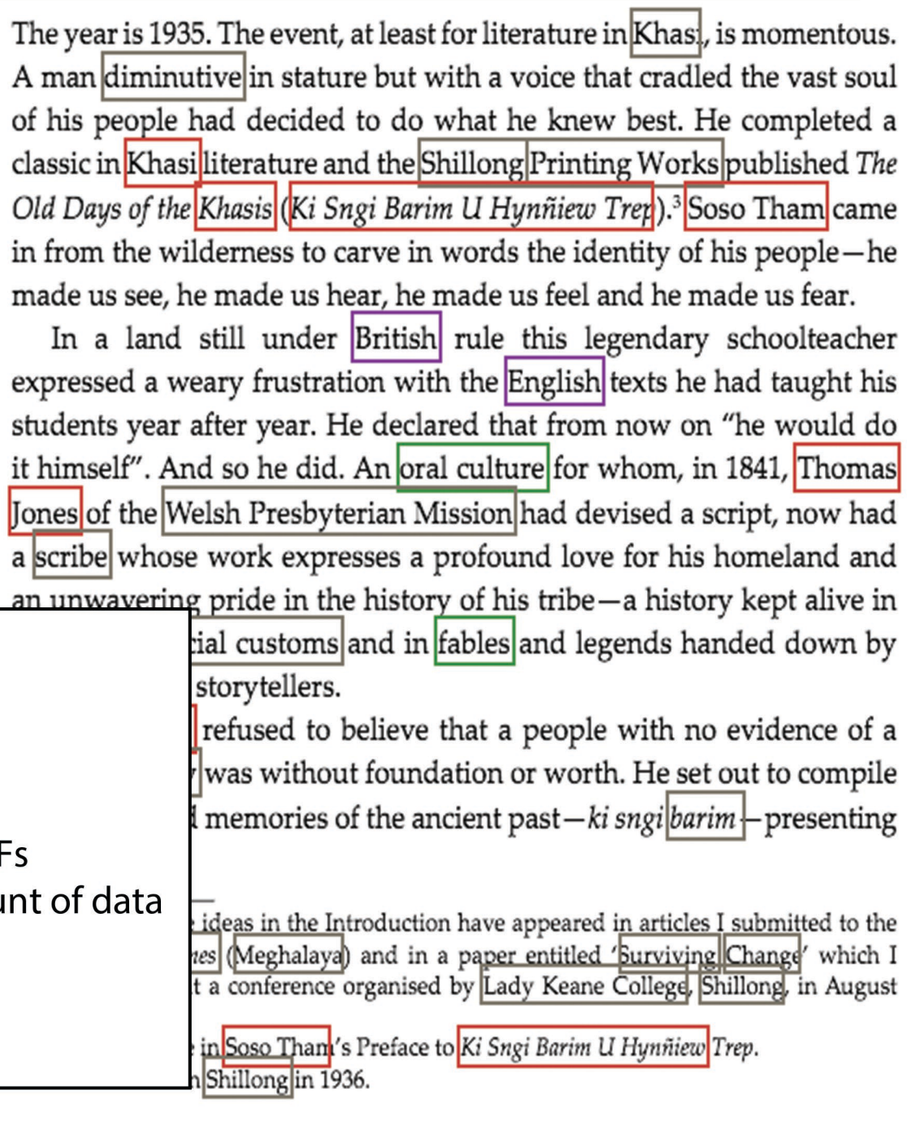
1. Available as a service in the whole DARIAH infrastructure
2. Flexible infrastructure
3. Integrated in OPERAS Open Access publishing platforms (financed by HIRMEOS, H2020 European project)
4. Ready for supporting more DH projects
5. Access to free Open Data via standardized REST API

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 ial 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

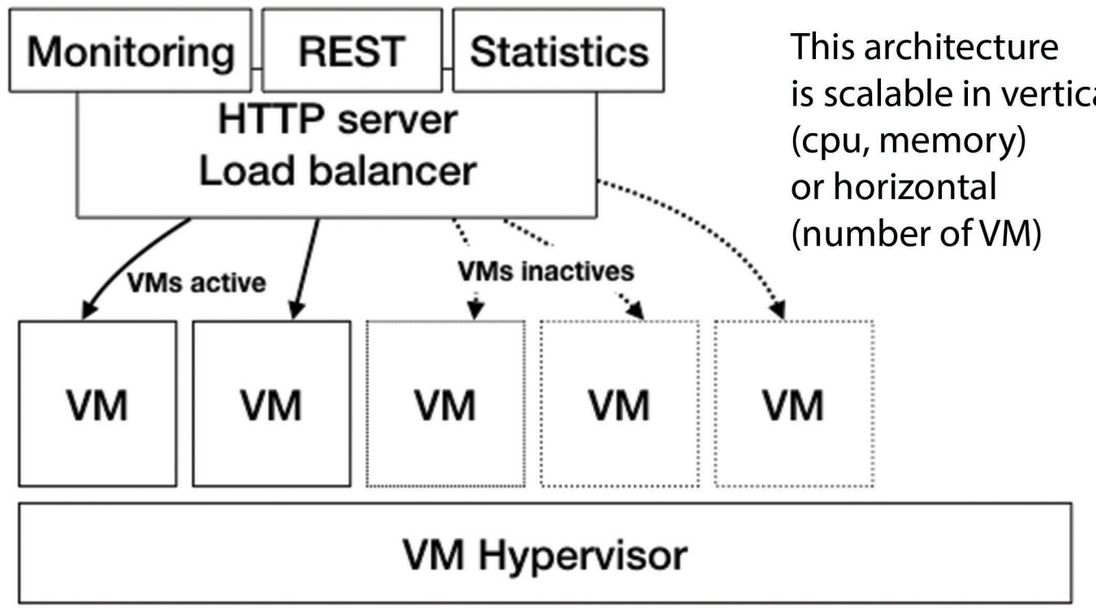
ideas in the Introduction have appeared in articles I submitted to the nes (Meghalaya) and in a paper entitled 'Surviving Change' which I t a conference organised by Lady Keane College, Shillong, in August

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

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.