CONTACT PERSONS
Maud Ehrmann, EPFL
Matteo Romanello, EPFL
Simon Clematide, UZH

CLEF HIPE 2020 Evaluation Lab

Identifying Historical People, Places and other Entities

STAY CONNECTED
https://impresso-project.ch
TWITTER @ImpressoProject

Nº 0002 LUGANO (CH) / THURSDAY, SEPTEMBER 12TH, 2019 FREE / OPEN-SOURCE

Shared Task on Named Entity Recognition and Linking on Historical Newspapers

GOAL & DESCRIPTION

» HIPE is a CLEF 2020 Evaluation Lab aimed at carrying out a named entity processing evaluation campaign on **historical newspapers** in French, German and English.

» The objective is to **assess and advance the development of robust named entity processing systems** able to deal with challenging, multilingual, diachronic historical material, thereby supporting information extraction and text understanding of cultural heritage data.

» Challenges:

» Multilingual corpus (English, French, German)

» Noisy Optical Character Recognition

» Partial coverage of knowledge bases with respect to historical entities

TASKS

» **Task 1: Named Entity Recognition and Classification (NERC)**

» **Subtask 1.1** - NERC Essentials: recognition and classification of high-level entity types: Person, Organisation, Location and Product.

» **Subtask 1.2** - NERC fine-grained: NERC Essentials, plus the detection and classification of sub-types (e.g. Person-individual vs. Person-collective) and the detection of NE components (e.g. function, title, name).

» **Task 2 : Named Entity Linking (EL)**

» **Subtask 2.1** - Entity coreference resolution: given a set of mentions, cluster them into coreferent sets and give them a unique identifier.

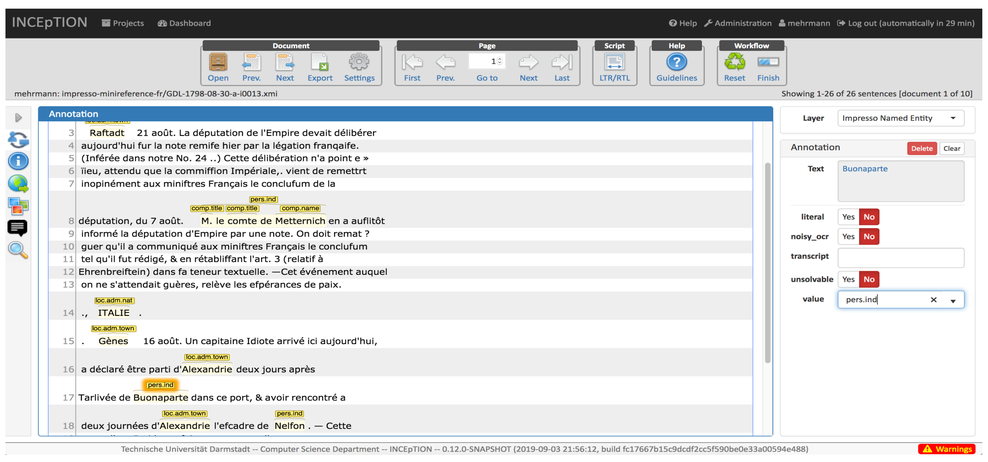
» **Subtask 2.2** - Entity Linking: linking of named entity mentions to a unique referent in Wikidata (or to a NIL entity).

CORPORA & RESOURCES

» Evaluation corpora consist of articles diachronically sampled from several Swiss, Luxembourgish and British/American historical newspapers. Articles and accompanying metadata will be publicly released as part of the shared task.

» HIPE will provide in-domain word-level and character-level embeddings.

» Participants are encouraged to share any external resources they might use, during and/or after the evaluation campaign.

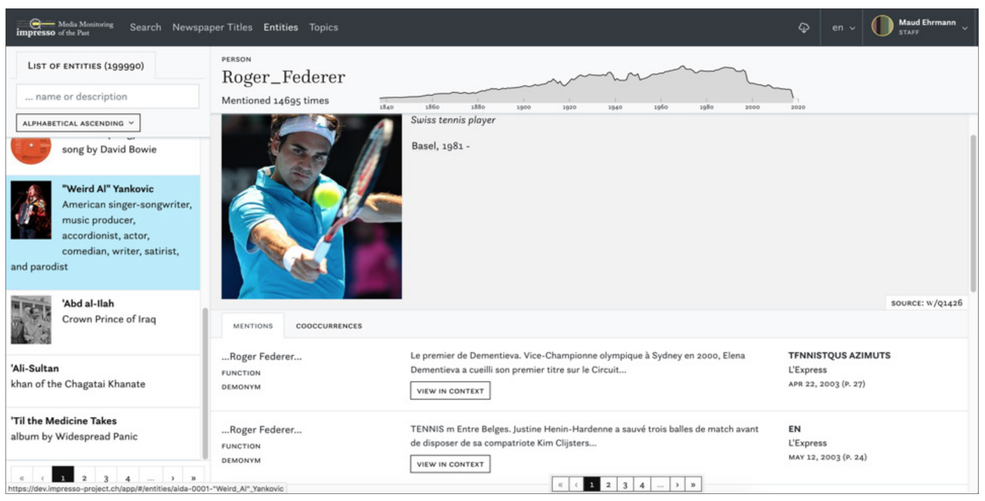


Named entity annotation in the INCEpTION environment.

» **OCR quality** varies according to digitization time and archival material. Newspaper publication date and digitization date are included in the metadata, and original images will be released (if possible).

SEMANTIC ENRICHMENT OF HISTORICAL TEXTS

» Impresso, Media monitoring of the past is a project aimed at developing NLP tools and visualization for active and goal-oriented exploration and critical analysis of newspaper corpora.



Screenshot of the named entity exploration page of the impresso interface.

» Text mining can be used to search, extract, process, link, and explore data from historical texts.

» For example, named entity information in the impresso interface enables historians to track actors (politicians, journalists, etc.) in a large, multilingual and diachronic corpus of digitized newspapers.

TIMETABLE

» **Early November 2019**: registration opens.

» **End of November 2019**: release of sample data for all tasks and languages.

» **End of January 2020**: release of full data (training and dev).

» **Early May 2020**: test phase (one week).

» **Early June 2020**: release of results to participants.

» **June-July 2020**: submission of papers (participant and lab- overview papers).

PARTICIPATION & REGISTRATION

» Teams should register via the CLEF 2020 registration portal.

» It will be possible to participate in one, some or all of the subtasks.

» Up to 3 runs per task and per team will be accepted.

https://impresso.github.io/CLEF-HIPE-2020/

https://groups.google.com/d/forum/clef-hipe-2020

ORGANIZERS Maud Ehrmann, EPFL - Matteo Romanello, EPFL -Simon Clematide, UZH
ADVISORY BOARD Richard Eckart de Castilho, TU Darmstadt - Clemens Neudecker, Berlin State Library - Sophie Rosset, LIMSI-CNRS - David Smith, Northeastern University
CONTRIBUTORS Camille Watter, UZH - Alex Flückiger, UZH

FNSNF Fonds national suisse Schweizerischer Nationalfonds Fondo nazionale svizzero Swiss National Science Foundation
Grant n° CR- SII5_173719
EPFL École polytechnique fédérale de Lausanne
University of Zurich[UZH]
impresso Media Monitoring of the Past