

HIPE 2022

Participation Guidelines

Identifying **H**istorical **P**eople, Places and other **E**ntities

Shared Task on Named Entity Recognition and Linking
in Multilingual Historical Documents

v1.0 (February 2022)

[10.5281/zenodo.6045662](https://doi.org/10.5281/zenodo.6045662)

Maud Ehrmann, Matteo Romanello, Antoine Doucet, Simon Clematide.

This document is licensed under a [Creative Common Attribution-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. INTRODUCTION

1.1 Motivation and context

1.2 Overview

1.3 Information

1.4 Glossary

2. TASKS

2.1 Task 1 - Named Entity Recognition and Classification (NERC)

2.2 Task 2 - Entity linking (EL)

3. DATA

3.1 Primary Datasets

3.2 Entity Tag Sets

3.4 HIPE-2022 releases

3.4 HIPE format and tagging scheme

4. EVALUATION

4.1 Metrics

4.2 Task Bundles, Tracks and Challenges

4.3 System evaluation

5. SYSTEM RESPONSES

6. WORKSHOP and WORKING NOTE PAPERS

6. ACKNOWLEDGMENTS and ORGANIZATION

7. REFERENCES

8. APPENDICES

APPENDIX A - Overview of Mapping of Primary Dataset to HIPE-2022

APPENDIX B - Primary Dataset Entity Tag Sets

APPENDIX C - NERC System Annotation Guidelines.

Useful links:

- HIPE 2022 website: <https://hipe-eval.github.io/HIPE-2022/>
- HIPE 2022 data repository: <https://github.com/hipe-eval/HIPE-2022-data>
- CLEF 2022 website: <https://clef2022.clef-initiative.eu/>
- CLEF 2022 registration: <https://clef2022-labs-registration.dei.unipd.it/>
- HIPE 2022 scorer: <https://github.com/impresso/CLEF-HIPE-2020-scorer> (the repository might change, things will be documented)
- EasyChair: <https://easychair.org/conferences/?conf=clef2022>
- HIPE series website: <https://hipe-eval.github.io/>

1. INTRODUCTION

1.1 Motivation and context

HIPE-2022 is part of the ongoing efforts of the natural language processing and digital humanities communities to develop efficient approaches to retrieve and explore information from digitized historical texts. After years of massive digitization by cultural heritage institutions, semantic indexing of digitized historical documents is in high demand by humanities scholars and various interdisciplinary efforts are advancing the processing of facsimiles, as well as the extraction, linking and representation of the complex information contained in transcriptions thereof. In this regard, the recognition, classification and disambiguation of named entities (NE) can be considered among the most crucial processing steps.

Yet, NE processing in historical texts is not straightforward, and performances are not comparable to what is typically observed on contemporary, well-edited English news material. In particular, NE processing on historical documents faces the challenges of domain heterogeneity, input noisiness, dynamics of language, and lack of resources [1].

In this regard, the first [CLEF-HIPE-2020](#) edition proposed the tasks of NE recognition and classification (NERC) and entity linking (EL) in ca. 200 years of historical newspapers written in English, French and German [2]. The main conclusion of this edition was that neural-based approaches can achieve good performances on historical NERC when provided with enough training data, but that progress is still needed to further improve performances, adequately handle OCR noise and small-data settings, and better address entity linking. HIPE-2022 attempts to drive further progress on these points, and also confronts systems with new challenges.

1.2 Overview

[HIPE-2022](#) focuses on named entity processing in historical documents covering the period from the 18th to the 20th century and featuring several languages. HIPE-2022 is based on diverse NE-annotated datasets (see Section 3). Compared to the first edition, HIPE-2022 introduces several novelties, with:

- the addition of a **new type of document** alongside historical newspapers, namely classical commentaries;
- the consideration of a **broader language spectrum**, with 5 languages for historical newspapers and 3 for classical commentaries;
- the confrontation with the issue of the **heterogeneity of annotation tag sets and guidelines**.

HIPE-2022 will confront participants with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemes. The objective of the evaluation lab is to gain new insights on **how best to ensure the transferability of NE processing approaches across languages, time periods, document and annotation types**.

1.3 Information

For contact, registration, data download, important dates and updates, please refer to the [HIPE-2022 website](#), the [HIPE-2022-data](#) github repository, and subscribe to the [google group](#).

Initiated in the framework of the *impresso* project with the first CLEF-HIPE-2020 edition, HIPE-2022 is a [CLEF 2022](#) Evaluation Lab. In the medium term, HIPE is meant to become a series of shared tasks on NE processing in multilingual historical documents.

1.4 Glossary

- **Primary dataset:** a NE-annotated dataset produced by a third party that is part of the HIPE-2022 shared task. A primary dataset may or may not have been published prior to the task.
- **Document:** the textual unit present in a NE-annotated dataset (delimited by empty lines). HIPE-2022 inherits various document unit delimitations from the primary datasets and does not fully homogenize them. In practice, most document units correspond to newspaper articles (see dataset-specific READMEs in the [HIPE-2022-data](#) repository).
- **Annotation guidelines:** a document which defines a set of rules to follow when manually annotating a corpus (definition of typology or tag set; explanation of entity types and which linguistic units to consider). The primary datasets present in HIPE-2022 follow different annotation guidelines, some of which are compatible (see specific READMEs).
- **Entity tag set:** the set of entity types (e.g. person, organization, etc.) considered in a dataset (aka as typology). See [APPENDIX B](#) for an overview.
- **HIPE release:** a single package composed of neatly structured and homogeneously formatted primary datasets of diverse origins.
- **HIPE-2022 task:** a predefined type and set of predictions a system must provide. HIPE-2022 evaluates three tasks: NERC-Coarse, NERC-Fine and EL (see Section 2).
- **Annotation column:** specific columns in the annotated .tsv files carrying annotation information. A task may correspond to several annotation columns.
- **Task bundle:** a predefined set of tasks a team chose to run for. A task bundle is composed of all or a selection from: NERC-Coarse, NERC-Fine, end-to-end EL, EL-only (see [Table 5](#)).
- **Submission bundle:** a participant submission for a given triple [dataset-language-taskbundle].
- **Track:** a specific triple composed of the test set of [dataset-language-task].
- **Challenge:** a HIPE-2022 predefined set of tracks. A challenge can be seen as a kind of tournament with multiple tracks.

2. TASKS

HIPE-2022 focuses on the same task type as CLEF-HIPE-2020, namely:

2.1 Task 1 - Named Entity Recognition and Classification (NERC)

- *Subtask 1.1 - NERC Coarse-grained*: this task includes the recognition and classification of entity mentions according to coarse-grained types (cf. column 1 in Table 2).
- *Subtask 1.2 - NERC Fine-grained*: this task includes the recognition and classification of entity mentions according to fine-grained types (cf. column 2 in Table 2), plus the detection and classification of nested entities of depth 1. This subtask will be proposed for some datasets only, in English, French and German.

NERC system annotation guidelines:

The primary datasets on which HIPE-2022 datasets are based do not contain systematically all NERC annotation types (coarse, fine, nested) and feature different entity tag sets. The types of annotation that systems are expected to produce for each task are presented in Table 1.

	Task 1.1 NERC-Coarse	Task 1.2 NERC-Fine
NE mentions with coarse types	yes	no
NE mentions with fine types	no	yes
Nested entities of depth 1	no	yes

Table 1. Expected annotation types for Task 1.

Table 3 hereafter (Section 3.2) lists the entity types to consider for Task 1 (NERC) for each dataset. For more information about system annotation rules, please refer to dataset-specific READMEs and dataset annotation guidelines.

2.2 Task 2 - Entity linking (EL)

This task corresponds to the linking of named entity mentions to a unique item ID in [Wikidata](#), our knowledge base of choice, or to a NIL node if the mention does not have a corresponding item in the KB. We will allow submissions of both end-to-end systems (NERC and NEL) and of systems performing exclusively NEL on gold entity mentions provided by the organizers (EL-only).

EL system annotation guidelines:

1. Systems are required to link mentions of selected types by giving their corresponding wikidata id (the 'Q' id); the type of mentions to link are listed in Table 3.

2. In case the referred entity does not exist in the knowledge base, systems should indicate 'NIL'; please note that it is not allowed to annotate with wikipedia disambiguation pages.
3. Entity links must be set with respect to literal or metonymic (when present) mention annotations (in the NEL-LIT annotation column only).
4. Nested entities are excluded from linking.

EL task settings

The entity linking task includes two settings: with (EL only) and without (end-to-end EL) prior knowledge of mention boundaries. The evaluation period will consist of two consecutive rounds, where a first EL task without prior information on mentions will be evaluated during round 1 (bundles 1 and 2, see Table 5), and a second one with information on mention boundaries (but no NE type information) during the second round (bundle 5).

3. DATA

3.1 Primary Datasets

HIPE-2022 data consists of six primary NE-annotated datasets assembled and prepared for the shared task. Primary datasets originate from several European cultural heritage projects, from HIPE organizers' previous research project, and from the previous HIPE-2020 campaign. Some are already published, others are released for the first time for HIPE-2022.

Primary datasets are composed of historical newspapers and classical commentaries covering ca. 200 years; they feature several languages and were annotated with different entity tag sets and according to different annotation guidelines. See Table 2 for an overview.

Historical newspaper datasets. The historical newspaper data is composed of several datasets in English, Finnish, French, German and Swedish which originate from various projects and national libraries in Europe.

- *HIPE-2020* data: the datasets of the first HIPE-2020 campaign [2], composed of newspaper articles from Swiss, Luxembourgish and American newspapers in French, German and English (19C-20C). These datasets contain ca. 10,000 linked entities and are part of HIPE-2022 train and dev sets. For test sets, former HIPE-2020 test sets will be used for comparison purposes, as well as unpublished HIPE-2020 data (new test sets for some languages). No part of the existing test set can be used for training.
- *NewsEye* data: an NE-annotated dataset composed of newspaper articles from newspapers in French, German, Finnish and Swedish (19C-20C) and built in the context

Dataset alias	Readme	Document type	Languages	Suitable for	Project
hipe2020	link	historical newspapers	de, fr, en	NERC-Coarse, NERC-Fine, EL	CLEF-HIPE-2020
newseye	link	historical newspapers	de, fi, fr, sv	NERC-Coarse, NERC-Fine, EL	NewsEye
sonar	link	historical newspapers	de	NERC-Coarse, EL	SoNAR
letemps	link	historical newspapers	fr	NERC-Coarse, NERC-Fine	LeTemps
topres19th	link	historical newspapers	en	NERC-Coarse, EL	Living with Machines
ajmc	link	classical commentaries	de, fr, en	NERC-Coarse, NERC-Fine, EL	AjMC

Table 2. Overview of HIPE-2022 primary datasets.

of the NewsEye project¹ [3]. The already published part contains ca. 30,000 entities and are part of HIPE-2022 train and dev sets. The unpublished part (roughly 20% of the total) will be part of the test set.

- *SoNAR* data: an NE-annotated dataset composed of newspaper articles from the Berlin State library newspaper collections in German (19C-20C), produced in the context of the SoNAR project² [4]. The already published part of this dataset is part of HIPE-2022 dev sets, a part is set aside for HIPE-2022 test set. Only the dev set can be used for training.
- *Le Temps* data: an unpublished, NE-annotated diachronic dataset composed of historical newspaper articles from two Swiss newspapers in French (19C-20C) [5]. This dataset contains ca 10,000 entity mentions and is part of HIPE-2022 train, dev and test sets.
- *Living With Machines*³ data: the topRes19th annotated dataset composed of newspaper articles from the British Library newspapers in English (18C-19C), and annotated exclusively with geographical locations. The already published portion of the data [6] contains ca. 3,300 annotated toponyms and is included in HIPE-2022 train and dev sets. The unpublished portion will be part of the test set.

¹ <https://www.newseye.eu/>

² <https://sonar.fh-potsdam.de/>

³ <https://livingwithmachines.ac.uk/>

Historical commentaries⁴.

- The *AJMC classical commentaries* datasets originates from the *Ajax Multi-Commentary*⁵ project and are composed of 19C commentaries published in French, German and English [7]. These datasets are new.

The textual materials of the datasets come from different Optical Character Recognition (OCR) software products and are of varying quality.

3.2 Entity Tag Sets

Primary datasets have different entity tag sets and were built according to different annotation guidelines. In practice, datasets are converted to the HIPE format but entity tag sets are left untouched. APPENDIX B gives an overview of the entity types used in each dataset; also visit the dataset-specific READMEs in the repository [HIPE-2022-data](https://github.com/mromanello/hipe-2022-data).

Datasets are suitable for certain HIPE-2022 tasks according to their annotation types (NERC-Coarse, NERC-Fine, EL); see column “Suitable For” in [Table 2](#) and the Challenge overview in [Table 6](#).

3.4 HIPE-2022 releases

A HIPE-2022 release corresponds to a single package composed of neatly structured and homogeneously formatted primary datasets, which undergo the following preparation steps:

- conversion to the tabulator-separated HIPE format (with correction of data inconsistencies and metadata consolidation);
- rearrangement or composition of train and dev splits.

⁴ Classical commentaries are scholarly publications dedicated to the in-depth analysis and explanation of ancient literary works. As such, they aim to facilitate the reading and understanding of a given literary text.

⁵ <https://mromanello.github.io/ajax-multi-commentary/>

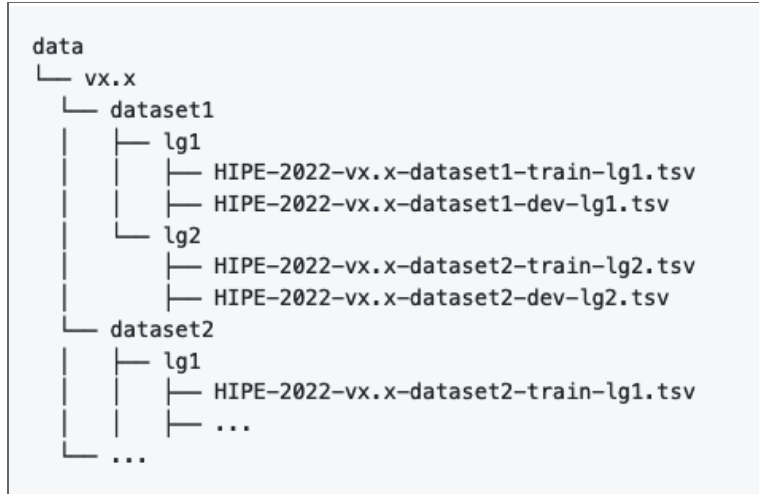


Figure 1. HIPE-2022 release directory structure.

Directory structure and naming conventions:

[HIPE-2022-data](#) “data” directory is organized per HIPE release version, dataset and language, as shown in Fig. 1.

- Training and development datasets consist of UTF-8, tab-separated files.
- There is one .tsv file per dataset, language and split.
- Files contain information needed for all tasks (NERC-Coarse, NERC-Fine, EL).
- Files are named according to this schema:

HIPE-2022-<hipeversion>-<dataset-alias>-<split>-<language>.tsv, where the value of split can be sample, train, dev, dev2, test.

For example, the file HIPE-2022-v1.0-newseye-dev-sv.tsv contains NE-annotated documents of the Swedish part of the newseye corpus which are meant as development set, in HIPE format and from HIPE-2022 release v1.0.

Versioning:

- HIPE-2022 releases are versioned with a two-part version number (Major.Minor) which is present in 1) the data directory structure and 2) the filename of each file.
- Each HIPE-2022 release has an equivalent git repository release, with release notes.
- The version of a primary dataset is mentioned in its document metadata (see below).

3.4 HIPE format and tagging scheme

HIPE format is a simple tab-separated column textual format using an IOB⁶ tagging scheme (inside-outside-beginning format), similarly to that of the CoNLL-U format⁷.

3.4.1 File structure

Files encode annotations needed for all tasks (NERC-Coarse, NERC-Fine and NEL) and contain the following lines:

- empty lines, which mark the boundaries between documents;
- comment lines, which give further information and start with the character '#' followed by a space;
- annotated lines, which contain a token followed by its tab-separated annotations.

A file contains all the documents of one dataset/language/split. Documents are separated with empty lines and are preceded with several metadata comment lines. The notion of document varies from one dataset to another, please refer to dataset-specific READMEs.

3.4.2 Document metadata

Primary datasets provide different document metadata, with different granularity. This information is kept in HIPE-2022 files in the form of "metadata blocks". HIPE-2022 metadata blocks encode as much information as necessary to ensure that each document is self-contained with respect to HIPE-2022 settings. Metadata blocks use namespacing to distinguish between mandatory HIPE-2022 metadata and dataset-specific metadata:

# hipec2022:document_id	identifier for the document inside a dataset
# hipec2022:date	original document publication date (YYYY-MM-DD, with YYYY-01-01 if month or date are not available)
# hipec2022:language	iso two-letter language code
# hipec2022:dataset	dataset alias as in file name
# hipec2022:document_type	newspaper or commentary
# hipec2022:original_source	path to source file in original dataset release
# hipec2022:applicable_columns	columns originally present in the primary dataset; non-applicable columns have _ values everywhere; applicable columns do not equal to evaluated columns, check evaluation settings.
# DATASET:doi	DOI url of primary dataset release (if available)
# DATASET:version	version of the primary dataset used in the HIPE-2022 release
# DATASET: xxx	any other metadata about the dataset (e.g. license)

⁶ [https://en.wikipedia.org/wiki/Inside-outside-beginning_\(tagging\)](https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging))

⁷ <https://universaldependencies.org/format.html>

3.4.3 File contents

Each line consists of 10 columns:

1. TOKEN: the annotated token.
2. NE-COARSE-LIT: the coarse type (IOB-type) of the entity mention token, according to the literal sense.
3. NE-COARSE-METO: the coarse type (IOB-type) of the entity mention token, according to the metonymic sense.
4. NE-FINE-LIT: the fine-grained type (IOB-type.subtype.subtype) of the entity mention token, according to the literal sense.
5. NE-FINE-METO: the fine-grained type (IOB-type.subtype.subtype) of the entity mention token, according to the metonymic sense.
6. NE-FINE-COMP: the component type of the entity mention token.
7. NE-NESTED: the coarse type of the nested entity (if any).
8. NEL-LIT: the Wikidata Q id of the literal sense, or `NIL` if an entity cannot be linked. Rows without link annotations have value `_`.
9. NEL-METO: the Wikidata Q id of the metonymic sense, or `NIL`.
10. MISC: a flag which can take the following values:
 - NoSpaceAfter: to indicate the absence of white space after the token.
 - EndOfLine: to indicate the end of a layout line.
 - EndOfSentence: to indicate the end of a sentence.
 - Partial-START:STOP: to indicate the zero-based character on-/offsets of mentions that do not cover the full token (esp. for German compounds). START and STOP follow Python's slicing semantics: "abcd"[1:3] means "bc".

Non-specified values are marked by the underscore character “_”.

3.4.4 HIPE-2022 NE annotation types

HIPE-2022 annotation scheme originates from the CLEF-HIPE-2020 shared task and contains detailed named entity annotation columns (reflected in the IOB file columns presented above).

△ However, given its wide scope in terms of languages and datasets, **HIPE-2022 evaluation only focuses on a selection of annotation columns**, as shown in Table 3 and Fig. 2.

HIPE-2022 Tasks	HIPE-2022 relevant annotation columns
NERC-Coarse	NE-COARSE-LIT
NERC-Fine	NE-FINE-LIT, NE-NESTED
NEL	NEL-LIT

Table 3: HIPE-2022 tasks and relevant annotation columns.

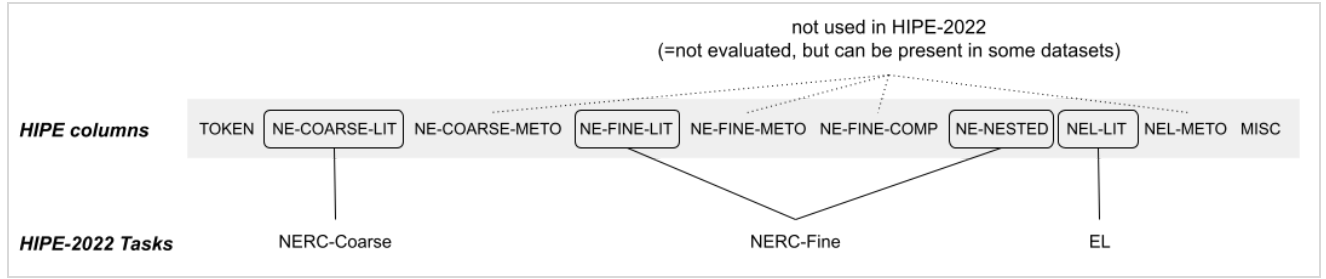


Fig. 2. Used annotation columns in HIPE-2022.

The annotation types NE-COARSE-METO, NE-FINE-METO, NE-FINE-COMP are not considered in HIPE-2022 tasks and evaluation scenarios but are left in the IOB files when present with a dataset, for systems to use this information if beneficial.

Since they were created according to different annotation schemes, datasets do not systematically include all columns. Applicable columns for a dataset are specified in each document metadata and an overview is given in Table 4. When a column does not apply, all its values are " _".

NE annotation type	ajmc	hipe2020	letemps	topres19th	newseye	sonar
NE-COARSE-LIT	x	x	x	x	x	x
NE-COARSE-METO	-	x	-	-	-	-
NE-FINE-LIT	x	x	x	-	x	-
NE-FINE-METO	-	x	-	-	-	-
NE-FINE-COMP	-	x	-	-	-	-
NE-NESTED	x	x	x	-	x	-
NEL-LIT	x	x	-	x	x	x
NEL-METO	-	x	-	-	-	-

Table 4: Applicable annotation columns for each dataset.
Annotation columns taken into account for evaluation are highlighted in bold.

About sentence splitting and tokenization:

Primary datasets have sentence information (EndOfSentence flags in the MISC column). Sentence splitting quality varies according to each dataset, please refer to the dataset-specific READMEs. When possible, HIPE-2022 files provide all necessary information to rebuild the OCR text (EndOfLine and NoSpaceAfter flags).

About nested entities:

Annotations step from the outer to the innermost entity. NE-COARSE and NE-FINE correspond to the outermost entity mentions, and NE-NESTED corresponds to the first nested level. When nesting applies, only one level of nested entities is annotated.

About metonymic annotations:

Both NERC and EL annotations can be set according to the literal or the metonymic sense of a given entity mention. Primary datasets usually annotate the metonymic sense of an entity (either for NERC or EL), however annotation is implemented in different ways:

- a unique annotation, with either the literal or the metonymic sense according to the mention annotated. In a .tsv file this corresponds to having one column with either the metonymic sense when present, the literal otherwise, with no way to know whether there is metonymy or not.
- a double annotation, with both literal and metonymic annotations. In a .tsv file this corresponds to having two distinct columns.

The *hipe2020* datasets and the HIPE format (used in HIPE-2022) follow the two-column encoding, i.e.: NE-COARSE-LIT/NE-COARSE-METO, NE-FINE-LIT/NE-FINE-METO and NEL-LIT/NEL-METO. In practice, *hipe2020* metonymic annotations are present in the NE-COARSE-METO, NE-FINE-METO and NEL-METO columns, while NE-COARSE-LIT, NE-FINE-LIT and NEL-LIT always contain annotations according to the literal sense.

The datasets *newseye*, *sonar*, *topres19th* and *ajmc* do consider metonymic senses but follow the one-column approach. In practice, these datasets have values for the columns NE-COARSE-LIT and NEL-LIT only, but in case of metonymy, the metonymic type or link is put in the x-LIT column.

In the HIPE-2022 releases, datasets stay with their original setting, i.e. *hipe2020* have annotations for both columns, and other datasets only for the x-LIT columns.

△ **For the evaluation**, only the NE-COARSE-LIT, NE-FINE-LIT and NEL-LIT columns will be considered, however for the *hipe2020* data both values (as stated in the distinct LIT and METO columns of the test data) will be accepted as valid answers.

D. Additional resources

The HIPE Evaluation lab will provide additional lexical resources, please refer to the [HIPE website](#).

4. EVALUATION

4.1 Metrics

NERC is evaluated in terms of macro and micro Precision, Recall, F1-measure. Two evaluation settings are considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching). Each column is evaluated independently, according to the following metrics:

- **Micro average P, R, F1** at entity level (not at token level), i.e. consideration of all true positives, false positives, true negatives and false negatives over all documents.
 - strict (exact boundary matching) and fuzzy (at least 1 token overlap).
 - separately per type and cumulative for all types.
- **Document-level macro average P, R, F1** at entity level (not on token level). i.e. average of separate micro evaluation on each individual document.
 - strict and fuzzy
 - separately per type and cumulative for all types

Our definition of macro differs from the usual one, and macro measures are computed as aggregates on document-level instead of entity-type level. Specifically, macro measures average the corresponding micro scores across all the documents, accounting for (historical) variance in document length and not for class imbalances.

Note that in the strict scenario, predicting wrong boundaries leads to severe punishment of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (predicted entity by the system but not present in the gold standard). Although this may be severe, we keep this metric in line with CoNLL and refer to the fuzzy scenario if the boundaries of an entity are considered as less important.

The evaluation for NEL works similarly as for NERC. The link of an entity is interpreted as a label. As there is currently no IOB encoding, a consecutive row of identical links is considered as a single entity. As for boundaries, NEL is only evaluated according to the fuzzy scenario. Thus, to get counted as correct, the system response needs only one overlapping link label with the gold standard.

HIPE-2022 evaluation will use the [CLEF-HIPE-2020-scorer](#) (repository might change).

4.2 Task Bundles, Tracks and Challenges

To accommodate the different dimensions that characterize the HIPE-2022 Evaluation Lab (tasks, languages, document types, entity tag sets) and foster research on transferability, the evaluation lab is organized around **challenges** and **tracks**. Challenges help guide participation towards the development of approaches that work across settings, e.g. with documents in at least two different languages or annotated according to two different tag sets or guidelines, and ensure a clear and defined evaluation frame.

Bundle id	Associated tasks	Relevant columns in the IOB response file
bundle1	NERC-Coarse, NERC-Fine, NEL	TOKEN, NE-COARSE-LIT, NE-FINE-LIT, NE-NESTED, NEL-LIT
bundle2	NERC-Coarse, NEL	TOKEN, NE-COARSE-LIT, NEL-LIT
bundle3	NERC-Coarse, NERC-Fine	TOKEN, NE-COARSE-LIT, NE-FINE-LIT, NE-NESTED
bundle4	NERC-Coarse	TOKEN, NE-COARSE-LIT
bundle5	NEL-only	TOKEN, NEL-LIT

Table 5. List of task bundles that a system can participate in.

To manage the full combinatory (dataset, language, document type, task), we define the following:

- **task bundle:** as for CLEF-HIPE-2020, a task bundle is a predefined set of tasks, as presented in Table 5.
- **submission bundle:** a submission bundle corresponds to a triple composed of [dataset-language-taskbundle].
- **track:** a track corresponds to a triple composed of [dataset-language-task].
- **challenge:** a challenge corresponds to a predefined set of tracks. A challenge can be seen as a kind of tournament composed of tracks.

HIPE-2022 specifically evaluates 3 challenges:

1. Multilingual Newspaper Challenge (MNC):

The multilingual newspaper challenge aims at fostering the development of multilingual NE processing approaches on historical newspapers. The characteristics and participation requirements of this challenge are:

- submission bundles are composed of datasets of type “**newspaper**” only;
- submission bundles may be for different newspaper datasets (thus it is possible to run for one dataset only);
- submission bundles are for **at least two languages** for the same task (thus teams should submit a minimum of two submission bundles for this challenge);
- submission bundles are for two task bundles at most, among: bundle 2, 4, 5.

	Multilingual Newspaper Challenge		Multilingual Classical Commentary Challenge		Global Adaptation Challenge		
Evaluated tasks	NERC-Coarse	end-to-end EL, EL only	NERC-Coarse	end-to-end EL, EL only	NERC-Coarse	NERC-Fine	end-to-end EL, EL only
Possible datasets	all newspaper datasets	hipe2020, newseye, topres19th, sonar	ajmc		any (but at least one newspaper and one commentary doc type)		
Minimum number of languages	2		3		2		
Required doc type	newspaper		commentary		newspaper and commentary		
Possible languages	de, en, fr, fi, sv		de, en, fr		any		
Possible task bundles	2, 4	2, 5	2, 4	2, 5	1, 2, 3, 4	1, 3	1, 2, 5

Table 6. Overview of challenge possible settings.

2. Multilingual Classical Commentary Challenge (MCC):

The multilingual classical commentary challenge aims at adapting NE solutions to domain-specific entities in a specific Digital Humanities (DH) text type. The characteristics and participation requirements of this challenge are:

- submission bundles are composed of the dataset “**ajmc**” only;
- submission bundles are for at least **three** languages for the same task;
- submission bundles are for two task bundles at most, among: bundle 2, 4, 5.

3. Global Adaptation Challenge (GAC):

The global adaptation challenge aims at appreciating how efficiently systems can be retargeted to any language, document type and guidelines. Submitted bundles for this challenge may be the same as for Challenge 1 and 2.

The characteristics and participation requirements of this challenge are:

- submission bundles are composed of datasets of **both types** “ajmc” and “newspaper”;
- submission bundles are for at least **two** languages for the same task;
- submission bundles are for two task bundles at most, among: bundle 1, 2, 3, 4, 5.

4.3 System evaluation

4.3.1 Track Evaluation

Participant teams are invited to submit system responses on a submission bundle basis (dataset-language-taskbundle). System performances will be computed, reported and published in terms of micro and macro P, R, and F1 for each track (dataset-language-task, e.g. system results for the NERC-Coarse task on the English hipec2020 dataset, for the EL tasks on the German sonar dataset, etc.). For each track, systems will be ranked according to their F1 scores.

4.3.2 Challenge evaluation

At submission time, participant teams must declare at least one challenge to which their submitted bundles belong. Challenges correspond to an aggregation of tracks and are meant to assess the capacities of (a) system(s) to perform well across settings.

Given a specific challenge and the tracks submitted by a team for this challenge:

- for each track, the submitted responses/systems are rewarded points according to their F1-based rank (**considering only the best of the submitted team runs for a given track**), as presented in Table 7.
 - if two systems score equally, they obtain the same number of points.
 - for the NERC-Fine tracks, the F1 scores of the columns NE-FINE-LIT and NE-NESTED are summed.
- the points obtained are summed over all submitted tracks;
- systems/teams are ranked according to their points.

F1-score rank	Points awarded
1	50
2	40
3	30
4	20
5	10
6 and after	0

Table 7. Track scoring system.

This means that the more (good) tracks a team submits for a given challenge, the more chances they have to win the challenge. We encourage teams to develop approaches that can adapt well across settings with minimal efforts, but teams are free to tackle the tracks and challenge as they wish.

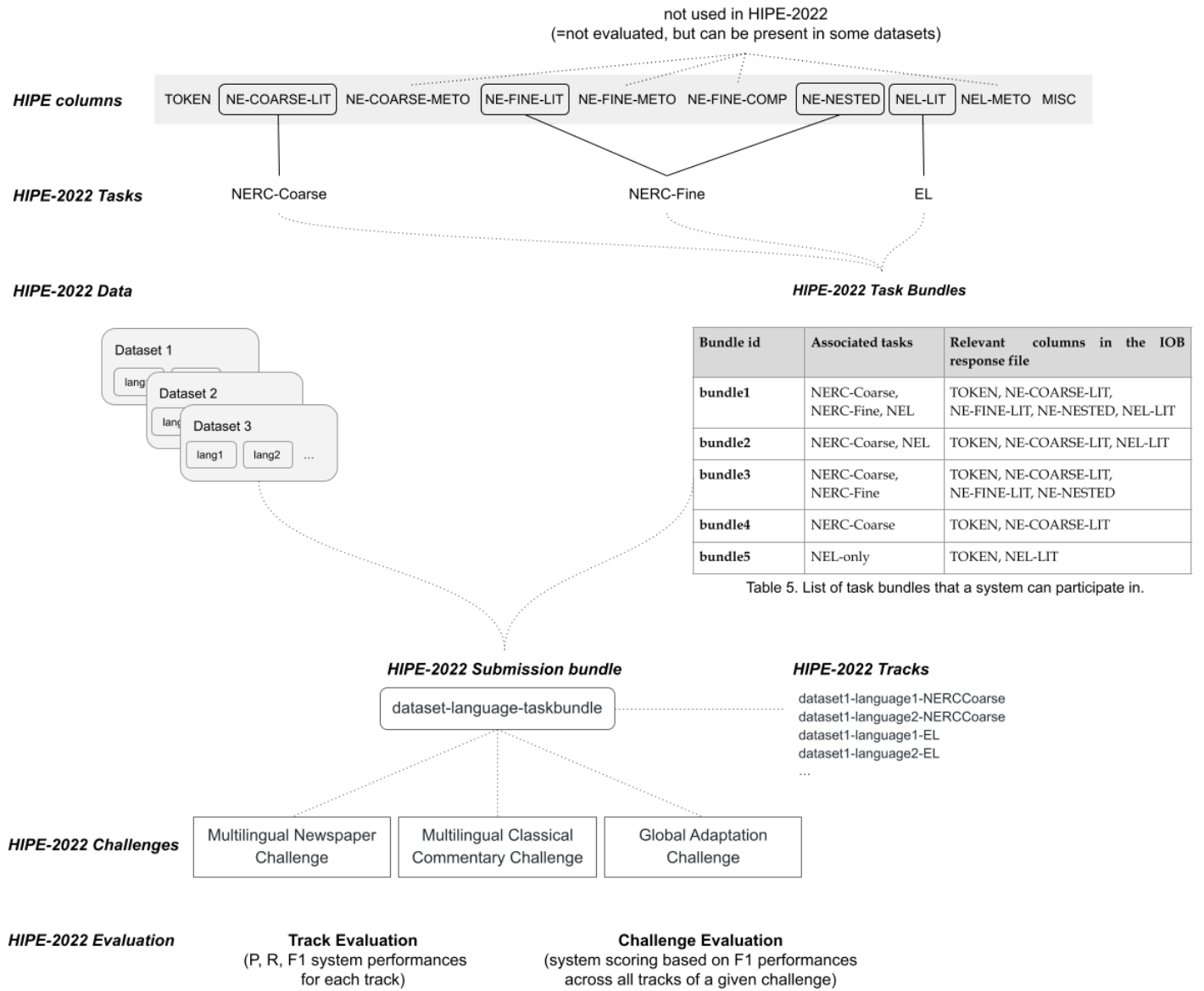


Fig. 3 HIPE-2022 Evaluation Setting Overview.

HIPE-2022 Challenges & Tracks						
Tracks (flat list of dataset-task-lang)		Multilingual Newspaper Challenge (MNC)		Multilingual Classical Commentary Challenge (MCC)		Global Adaptation Challenge (GAC)
1	hipe2020-coarse-en	NERC-Coarse	Multilingual Newspaper			NERC-Coarse
2	hipe2020-coarse-de					
3	hipe2020-coarse-fr					
4	newseye-coarse-de					
5	newseye-coarse-fi					
6	newseye-coarse-fr					
7	newseye-coarse-sv					
8	letemps-coarse-fr					
9	topres19th-coarse-en					
10	sonar-coarse-de					
11	hipe2020-el-en	EL				EL
12	hipe2020-el-de					
13	hipe2020-el-fr					
14	newseye-el-de					
15	newseye-el-fi					
16	newseye-el-fr					
17	newseye-el-sv					
18	topres19th-el-en					
19	sonar-el-de*					
20	ajmc-coarse-de**	NERC-Coarse		Multilingual Classical Commentary		NERC-Coarse
21	ajmc-coarse-en					
22	ajmc-coarse-fr					
23	ajmc-el-de	EL				EL
24	ajmc-el-en					
25	ajmc-el-fr					
26	hipe2020-fine-de					NERC-Fine
27	hipe2020-fine-fr					
28	newseye-fine-fi					
29	newseye-fine-fr					
30	newseye-fine-sv					
31	newseye-fine-de					
32	letemps-fine-fr					
33	ajmc-fine-de					NERC-Fine
34	ajmc-fine-en					
35	ajmc-fine-fr					

* Will be part of HIPE-2022 release v2.0,

** Only a sample of ajmc datasets is part of HIPE-2022 release v1.0, full datasets will come in 2.0.

Table 8: Overview of possible track per challenge.

5. SYSTEM RESPONSES

5.1 General rules

- Registration is open until 22 April 2022. Please refer to the [HIPE website](#) for more information.
- Teams must participate in at least one challenge; the number of submission bundles to submit for a challenge depends on the minimum number of languages required for a challenge (2 for MNC, 3 for MCC, free for GAC).
- Teams can participate in one task bundle per language.
- Teams can submit up to **two** runs per submission track.
- Teams can use any external resources (e.g. additional language resources provided by HIPE, available elsewhere or homemade, and other annotated data).
- ⚠ Due to the wide dataset setting of HIPE-2022, some test data is already available publicly. **Teams cannot use any additional data from the primary data projects than the material available via HIPE-2022 train/sample/dev sets and released in the HIPE-2022-data repository.** But they can use annotated data from any other project. The principles of trust and academic integrity apply.
- Teams are highly encouraged to share the additional resources they use, either during or after the evaluation.

5.2 Evaluation period

Please check important dates on the HIPE-2022 website. At the end of each evaluation period, participants will send their system responses via email to the task organizers, which will be evaluated using the scorer. Gold standard data will be distributed after the publication of the evaluation results.

5.3 System response submission guidelines

Input test data will consist of historical documents of each dataset and language. Data will be encoded in the same way as the train and dev data, but without the annotations: one token per line, with each document separated with a blank line.

A system submission consists of:

- several system response files;
- a manifest file

Rules for system response files:

- Files must be in UTF-8, tsv encoded (.tsv extension), with annotations in the same format as in train/dev sets.
- Files need to contain all document lines and empty lines in the order of the original input file. Metadata lines (#) may be included, but this is not mandatory.

- Files must comply with the following naming convention:
TEAMNAME_TASKBUNDLEID_DATASETALIAS_LANG_RUNNUMBER.tsv
where:
TEAMNAME: is the name of the team such as registered via the CLEF portal
TASKBUNDLEID: is one of the bundle ids as indicated in Table 4.
DATASETALIAS: is the alias of the dataset as indicated in Table 2 (and in all files).
LANG: is de, fi, fr, en, sv
RUNNUMBER: is 1 or 2.
Example: dreamteam_bundle1_topres19th_en_2.tsv
- Files must include all columns and instantiate the unspecified values in the required columns according to the chosen task bundle and dataset.

Rule for manifest file:

Submissions must include a manifest file declaring which track is submitted to which challenge:

- The manifest file must comply with the following naming convention:
TEAMNAME_hipe2022_manifest.txt
- The manifest file must list the submitted files and the challenge in which they participate, in this format:
SUBMISSION_FILE_NAME:CHALLENGE_ACRONYM

Example:

dreamteam_bundle1_topres19th_en_2.tsv:MNC

dreamteam_bundle1_topres19th_en_2.tsv:GAC

If a file “participates” in more than one challenge, there must be a line for each challenge.

System submission:

System response files and manifest must be:

- archived with the ZIP format (.zip);
- named as TEAMNAME.zip ;
- sent via email to [maud.ehrmann @ epfl.ch] by the submission deadline indicated on the [HIPE website](#). An acknowledgement of receipt will be sent upon reception.

6. WORKSHOP and WORKING NOTE PAPERS

Participants will submit a Working Note paper to be presented during the final workshop co-located with the CLEF conference (September 2022) and to be published online via the CEUR Workshop Proceeding open access publication service. Please check submission instructions and important dates on the [HIPE website](#).

Previous HIPE-2020 proceedings are available [here](#), and presentation recordings [here](#).

6. ACKNOWLEDGMENTS and ORGANIZATION

The HIPE 2022 team expresses her greatest appreciation to the CLEF-2022 Lab Organising Committee for the overall organization, to the members of the HIPE-2022 advisory board, namely Sally Chambers, Frédéric Kaplan and Clemens Neudecker, for their support, and to the partnering projects, namely [AJMC](#), *impresso*-HIPE-2020, [Living with Machines](#), [NewsEye](#), and [SoNAR](#), for contributing (and hiding) their NE-annotated datasets.

HIPE-2022 is organized by Maud Ehrmann (EPFL-DHLAB), Simon Clematide (University of Zurich), Matteo Romanello (University of Lausanne) and Antoine Doucet (University of La Rochelle).

7. REFERENCES

- [1] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, and A. Doucet, “Named Entity Recognition and Classification on Historical Documents: A Survey,” *ArXiv210911406 Cs*, Sep. 2021, <http://arxiv.org/abs/2109.11406>.
- [2] M. Ehrmann, M. Romanello, A. Flückiger, and S. Clematide, “Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers,” in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece, 2020, vol. 2696, p. 38. doi: 10.5281/zenodo.4117566, <https://infoscience.epfl.ch/record/281054>.
- [3] A. Hamdi *et al.*, “A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers,” in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, Jul. 2021, pp. 2328–2334, <https://doi.org/10.1145/3404835.3463255>.
- [4] S. Menzel *et al.*, “Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten,” in *Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten*, De Gruyter Saur, 2021, pp. 229–258, doi: [10.1515/9783110691597-012](https://doi.org/10.1515/9783110691597-012).
- [5] M. Ehrmann, G. Colavizza, Y. Rochat, and F. Kaplan, “Diachronic evaluation of NER systems on old newspapers,” in *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, 2016, pp. 97–107, <https://infoscience.epfl.ch/record/221391>.
- [6] M. C. Ardanuy *et al.*, “A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers,” *J. Open Humanit. Data*, vol. 8, no. 0, Art. no. 0, Jan. 2022, doi: 10.5334/johd.56. <http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.56/>
- [7] M. Romanello, N.-M. Sven, and R. Bruce, “Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs,” in *The 6th International Workshop on Historical Document Imaging and Processing*, New York, NY, USA, Sep. 2021, pp. 1–6. <https://doi.org/10.1145/3476887.3476911>

APPENDIX A - Overview of Mapping of Primary Dataset to HIPE-2022

	already published before HIPE-2022
	newly published for HIPE-2022
	unpublished until end of HIPE-2022 evaluation

Last edited: 2022-02-12

From primary datasets to HIPE-2022 release(s):
comprehensive mapping overview.

	ORIGINAL				HIPE-2022	
	Lang.	Format / Structure	Split	Filename	Split	Filename
hipe2020	en	HIPE IOB tsv / 1 file per split containing sev.docs	dev	HIPE-data-v1.3-dev-en.tsv	train	HIPE-2022-vx.x-hipe2020-train-en.tsv
		old and new test data			test	HIPE-2022-vx.x-hipe2020-test-en.tsv
	fr	id.	train	HIPE-data-v1.3-train-fr.tsv	train	HIPE-2022-vx.x-hipe2020-train-fr.tsv
			dev	HIPE-data-v1.3-dev-fr.tsv	dev	HIPE-2022-vx.x-hipe2020-dev-fr.tsv
		old and new test data			test	HIPE-2022-vx.x-hipe2020-test-fr.tsv
	de	id.	train	HIPE-data-v1.3-train-de.tsv	train	HIPE-2022-vx.x-hipe2020-train-de.tsv
			dev	HIPE-data-v1.3-dev-de.tsv	dev	HIPE-2022-vx.x-hipe2020-dev-de.tsv
		old and new test data			test	HIPE-2022-vx.x-hipe2020-test-de.tsv
letemps	fr	Brat / 1 file per document	no split	1 file per article, for two newspaper titles, sampled per decade.	train	HIPE-2022-vx.x-letemps-train-fr.tsv
					dev	HIPE-2022-vx.x-letemps-dev-fr.tsv
					test	HIPE-2022-vx.x-letemps-test-fr.tsv
topres19th	en	WebAnno TSV 3.2	no split	1 file per article, for several newspaper titles, sampled per decade.	train	HIPE-2022-vx.x-topRes19th-train-fr.tsv
					dev	HIPE-2022-vx.x-topRes19th-dev-fr.tsv
			no split	id.	test	HIPE-2022-vx.x-topRes19th-test-fr.tsv
newseye	de	IOB .tsv / 1 file per split containing several docs	train	NewsEye-German/train.tsv	train	HIPE-2022-v1.x-newseye-train-de.tsv
			dev	NewsEye-German/dev.tsv	dev	HIPE-2022-v1.x-newseye-dev-de.tsv
			test	NewsEye-German/test.tsv	dev2	HIPE-2022-v1.x-newseye-dev2-de.tsv
			test	NewsEye-German/test.tsv	test	HIPE-2022-v1.x-newseye-test-de.tsv
	fi	IOB .tsv / 1 file per split containing several docs	train	NewsEye-Finnish/train.tsv	train	HIPE-2022-v1.x-newseye-train-fi.tsv
			dev	NewsEye-Finnish/dev.tsv	dev	HIPE-2022-v1.x-newseye-dev-fi.tsv
			test	NewsEye-Finnish/test.tsv	dev2	HIPE-2022-v1.x-newseye-dev2-fi.tsv
			test	NewsEye-Finnish/test.tsv	test	HIPE-2022-v1.x-newseye-test-fi.tsv
	fr	IOB .tsv / 1 file per split containing several docs	train	NewsEye-French/train.tsv	train	HIPE-2022-v1.x-newseye-train-fr.tsv
			dev	NewsEye-French/dev.tsv	dev	HIPE-2022-v1.x-newseye-dev-fr.tsv
			test	NewsEye-French/test.tsv	dev2	HIPE-2022-v1.x-newseye-dev2-fr.tsv
			test	NewsEye-French/test.tsv	test	HIPE-2022-v1.x-newseye-test-fr.tsv
	sv	IOB .tsv / 1 file per split containing several docs	train	NewsEye-Swedish/train.tsv	train	HIPE-2022-v1.x-newseye-train-sv.tsv
			dev	NewsEye-Swedish/dev.tsv	dev	HIPE-2022-v1.x-newseye-dev-sv.tsv
			test	NewsEye-Swedish/test.tsv	dev2	HIPE-2022-v1.x-newseye-dev2-sv.tsv
			test	NewsEye-Swedish/test.tsv	test	HIPE-2022-v1.x-newseye-test-sv.tsv
ajmc	de	HIPE IOB tsv / 1 file per split containing sev.docs	train	n/a	train	HIPE-2022-v1.x-ajmc-train-de.tsv
		id.	dev	n/a	dev	HIPE-2022-v1.x-ajmc-dev-de.tsv
		id.	test	n/a	test	HIPE-2022-v1.x-ajmc-test-de.tsv
	en	HIPE IOB tsv / 1 file per split containing sev.docs	train	n/a	train	HIPE-2022-v1.x-ajmc-train-en.tsv
		id.	dev	n/a	dev	HIPE-2022-v1.x-ajmc-dev-en.tsv
		id.	test	n/a	test	HIPE-2022-v1.x-ajmc-test-en.tsv
	fr	HIPE IOB tsv / 1 file per split containing sev.docs	train	n/a	train	HIPE-2022-v1.x-ajmc-train-fr.tsv
		id.	dev	n/a	dev	HIPE-2022-v1.x-ajmc-dev-fr.tsv
		id.	test	n/a	test	HIPE-2022-v1.x-ajmc-test-fr.tsv
sonar	de	IOB .tsv file / 1 file per doc	no split	information will be present in post-evaluation release.	dev	HIPE-2022-v1.x-sonar-dev-de.tsv
					test	HIPE-2022-v1.x-sonar-test-de.tsv

APPENDIX B - Primary Dataset Entity Tag Sets

	Coarse-grained tag set	Fine-grained tag set	Nesting applies	Linking applies
<i>hipe2020</i>	pers	pers.ind pers.coll pers.ind.articleauthor	yes	yes
	org	org.adm org.ent org.ent.pressagency	yes	yes
	prod	prod.media prod.doctr	no	yes
	time	time.date.abs	no	no
	loc	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup	yes	yes
		loc.phys.geo loc.phys.hydro loc.phys.astro	yes	yes
		loc.oro	yes	yes
		loc.fac	yes	yes
		loc.add.phys loc.add.elec	yes	yes
		loc.unk	no	no
<i>newseye</i>	pers	pers.articleauthor	yes	yes
	org		yes	yes
	humanprod			
	loc		no	yes
<i>letemps</i>	pers	pers.ind pers.coll	yes	no
	loc	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup	yes	no
		loc.phys.geo loc.phys.hydro loc.phys.astro	yes	no
		loc.oro	yes	no
		loc.fac	yes	no
		loc.add.phys loc.add.elec	yes	no
		loc.unk	no	no

	Coarse-grained tag set	Fine-grained tag set	Nesting applies	Linking applies
<i>topres19th</i>	loc	-	no	yes
	building	-	no	yes
	street	-	no	yes
	alien	-	no	yes
	other	-	no	yes
	unknown	-	no	yes
	fiction	-	no	yes
<i>ajmc</i>	pers	pers.author pers.editor pers.myth pers.other	yes	yes*
	work	work.primlit work.seclit work.fragm	yes	yes*
	loc	-	yes	yes*
	object	object.manuscr object.museum	yes	no
	date	-	yes	no
	scope	-	yes	no
sonar	pers			
	loc			
	org			

* yes, unless token flagged as InSecondaryReference.

Table 3. Overview of entity types to annotate per dataset.

APPENDIX C - HIPE2020 dataset NERC System Annotation Guidelines.

We give hereafter the main annotation rules to consider while designing a NERC system for the hipec2020 dataset. Please note that this is a summary of the annotation guide. For more information one should refer to the [Impresso Annotation Guidelines](#).

B.1 Entity types and components

We reproduce Table 2 about the entity types to consider. An exact definition of each type is given in the annotator guidelines, and a brief one is given in Table 5 hereafter.

Coarse-grained tag set	Fine-grained tag set	Metonymy applies	Entity nesting applies	Linking applies
pers	pers.ind pers.coll pers.ind.articleauthor	yes	yes	yes
org	org.adm org.ent org.ent.pressagency	yes	yes	yes
prod	prod.media prod.doctr	yes	no	yes
time	time.date.abs		no	no
loc	loc.adm.town loc.adm.reg loc.adm.nat loc.adm.sup	yes	yes	yes
	loc.phys.geo loc.phys.hydro loc.phys.astro	yes	yes	yes
	loc.oro	yes	yes	yes
	loc.fac	yes	yes	yes
	loc.add.phys loc.add.elec	yes	yes	yes
	loc.unk	no	no	no

B.2 Lexical characteristics

Linguistic units considered as named entities must include a proper name, or a definite description having the status of a proper name, i.e. definite descriptions with a nominative function and a certain referential stability (see section 2.2.A on p. 3 of annotation guidelines).

Phrases such as

- *Die präkolumbianische Zivilisation, la civilisation précolombienne*

- *l'armée bavaroise*
- *les forces tchadiennes*
- *le gouvernement français*

are *not* annotated because they do not contain proper names.

Phrases such as *le gouvernement Franco* are annotated:

```
le <org.adm> gouvernement
    <comp.name> <pers.ind> Franco </pers.ind> </comp.name>
</org.adm>
```

B.3 Named entity boundaries

Each token is either completely part of a named entity or not at all. Named entity mentions exclude subordinate clauses, incidental clauses and determiners. They include pre- and post-modifiers (see section 2.2.B of annotation guidelines).

B.4 About very noisy OCR entities

Such entities were annotated including the garbage characters which the annotator – while looking at the article facsimile – thought they should be part of the mention.

B.5 About nested entities

System should annotate nested entities of depth one only.

B.6 About metonymy

NERC annotation: When it applies, entities of type PERS, ORG, LOC and PROD are annotated according to their metonymic sense in both coarse and fine NERC settings.

NE Linking: Please refer to Section 4 of annotation guidelines.

B.7 About coordinated entities

Refer to Section 2.4.B of the annotation guidelines p6.

B.8 About components

Components are to be annotated for Task 1.2 (fine-grained) are the following:

For the type PERS:

- comp.func
- comp.title
- comp.name
- comp.qualifier
- comp.demonym

For all other types, except DATE

- name, used to mark the name of the entity.

The component name is optional when the mention contains only one name.

B.9 Quick guide (also present in the annotation guidelines)

<i>Entity types and subtypes</i>	
pers.ind	A single person (<i>Roger Federer</i>)
pers.ind.articleauthor	A single person who is the author of an article.
pers.coll	A named group of people including musical groups (<i>die Beatles, La Mano Negra</i>). (note: <i>die Schweizer, Les français</i> are not annotated.)
org.ent	Organization that markets products or provides services (<i>Die Peugeot Gesellschaft, Die Waid; La société Peugeot, la Pitié-Salpêtrière</i>). (note: <i>Die schweizer Polizei; la police française</i> is not annotated)
org.ent.pressagency	Special type related to newspaper to spot press agencies.
org.adm	Organization that plays a mainly administrative role (<i>Die Stadtverwaltung Bern; la mairie de Paris</i>). (note: <i>Das Département für auswärtige Angelegenheiten; Le Ministère des Affaires Étrangères</i> is not annotated)
loc.adm.town	District, locality, hamlet, village, city, etc. (<i>Paris, Val de Crüye</i>).
loc.adm.reg	Cantons, communities of municipalities, departments, regions, etc. (<i>Autonome Gemeinschaft Baskenland; les Bouches du Rhône, Le Pays-Basque espagnol</i>).
loc.adm.nat	Countries (<i>Schweiz; France</i>).
loc.adm.sup	World regions, continent (<i>Maghreb; Pays-Basque</i>).
loc.phys.geo	Mountains, plains, plateaus, caves, volcanoes, canyons (<i>Die Alpen, Der Vesuv; gouffre de Padirac, Le mont Ventoux</i>).
loc.phys.hydro	Oceans, seas, rivers, streams, ponds, marshes (<i>Der Atlantik, Der Golfstrom; La Seine, Le Lac Paladru</i>).
loc.phys.astro	Planets, stars, galaxies and their parts (<i>Der Mond, Die Milchstrasse; La terre, la mer de la Tranquillité</i>).
loc.oro	Refers to roads, highways, streets, avenues, squares, etc. (<i>Die Autobahn A6; L'autoroute A6</i>).

loc.fac	Refers to the buildings (<i>Der Prime Tower; Le Palais de l'Élysée</i>).
loc.add.phys	Refers to physical addresses (<i>LIMSI-CNRS, Bâtiment 508, BP133, 91403 Orsay Cedex</i>).
loc.add.elec	Refers to electronic contact information (telephone and fax numbers, URL, e-mail address, identification of social network or Internet communication tools, etc., <i>http://www.limsi.fr/, 01-69-85-80-00</i>)
Loc.unk	Type used when it is not possible to choose among other location types.
prod.media	Newspapers, magazines, broadcasts, sales catalogues, etc. (<i>Die Zeit; Le Figaro, Le sept à huit, La ferme célébrités</i>).
prod.doctr	Political, philosophical, religious, sectarian doctrines. (<i>Der Sozialismus, Theravada Buddhismus; Zeugen Jehovas; Le socialism, le bouddhisme theravâda, le structuralism, la scientology</i>).
time.date.abs	An absolute date (<i>Sonntag der 13. November 2016; lundi 25 janvier 2010</i>)
Component	
name	is the only transversal component and is applied to any class except time. (<i>Die Peugeot Gesellschaft; la société Peugeot</i>)
comp.name	The component includes first, middle and last names as well as nickname and initials of individuals (<i>Samuel L. Jackson, S.L.J.</i>)
comp.title	Title or designator of a person. (<i>Herr Chirac, Ihre Hoheit Rainier; M. Chirac, Son Altesse le prince Rainier</i>).
comp.qualifier	A qualifier specifies a person in the form of a qualifying adjective. (<i>Der konservative Christoph Blocher; le socialiste Bertrand Delanoë</i>)
comp.function	A function or job of a named person. (<i>Bürgermeister Ann Hidalgo von Paris; maire de Paris Anne Hidalgo</i>).
comp.demonym	The geographical origin of a person (<i>Le français Alain Vigneron</i>).