



FAIRer transcriptions: HTR-United and the possibility of a common for training data

Alix Chagué

► To cite this version:

Alix Chagué. FAIRer transcriptions: HTR-United and the possibility of a common for training data. Horizons of digital philology, Università degli Studi di Napoli Federico II; Université de Montréal, Apr 2024, Naples, Italy. hal-04697566

HAL Id: hal-04697566

<https://inria.hal.science/hal-04697566v1>

Submitted on 13 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



FAIRer transcriptions: HTR-United and the possibility of a common for training data

Alix Chagué - alix.chague@inria.fr
(Inria; UdeM; EPHE)

Naples - April 16, 2024
Horizons of digital philology

Foreword

This presentation was partly inspired
from Thibault Clérice's introduction to
HTR with eScriptorium.



A decorative background featuring a faint, stylized map of the world in light blue. Overlaid on the map are several orange and brown industrial elements: three interlocking gears in the top left (one large orange, one small orange, one dark brown), a network of pipes in the top right (including a prominent vertical pipe and a horizontal pipe with a U-bend), and a single yellow gear at the bottom right.

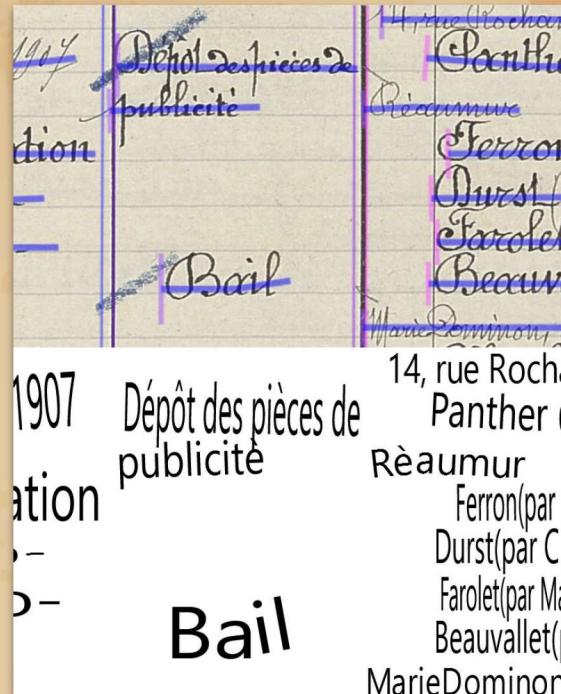
**WHAT ARE WE
TALKING ABOUT?
SOME
DEFINITIONS**

What is text recognition?

- HTR : Handwritten Text Recognition
- OCR : Optical Character Recognition

Key points to understand how it works:

1. Predict a textual content
2. From the image of a document (page)
3. Thanks to IA trained by humans
4. In an iterative process involving
 - a. human interventions
 - b. automatic/computation steps



Processing manuscript vs. prints

OCR (for prints)

- <2% character error
- Commercial software (FineReader, Google Cloud Vision, etc.) or open source software (Tesseract 4)
- Language-specific generic models ready to be used, rely on fonts

HTR (for manuscripts)

- 5-10% of character error
- Commercial software (Transkribus, ...) or open source software (Kraken / eScriotorium, ...)
- Require creating data to build a corpus-specific model

Processing manuscript vs. prints

OCR (for prints)

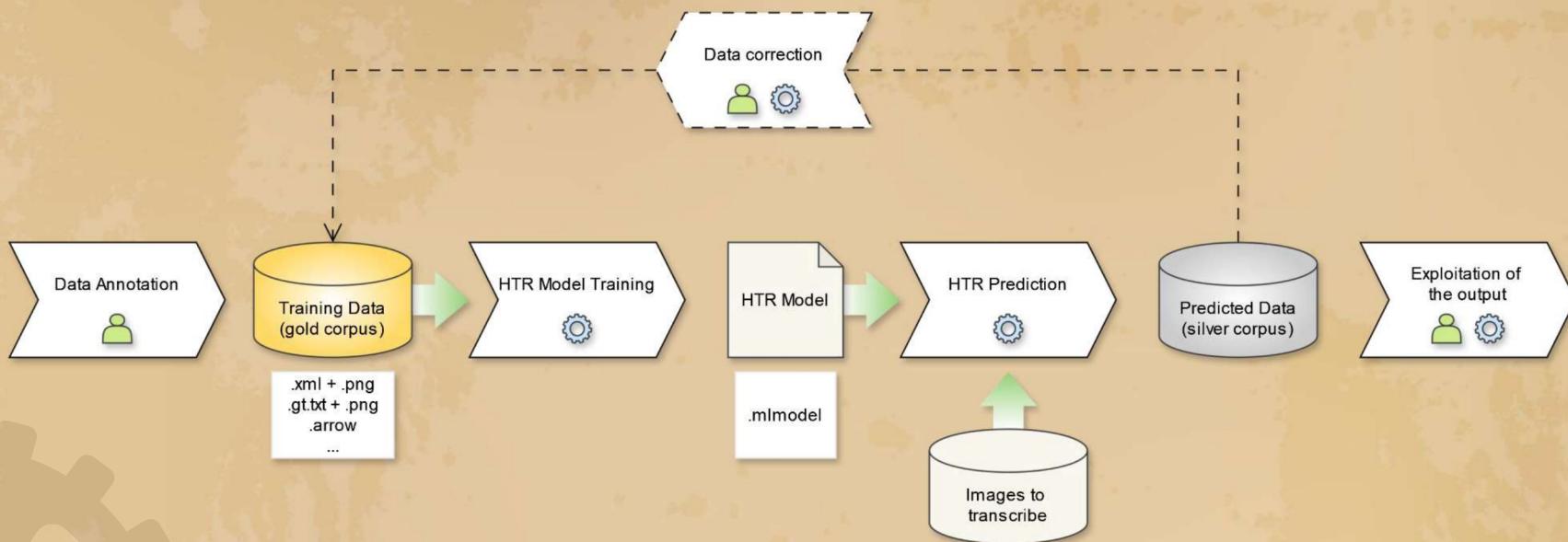
- <2% character error
- Commercial software (FineReader, Google Cloud Vision, etc.) or open source software (Tesseract 4)
- Language-specific generic models ready to be used, rely on fonts

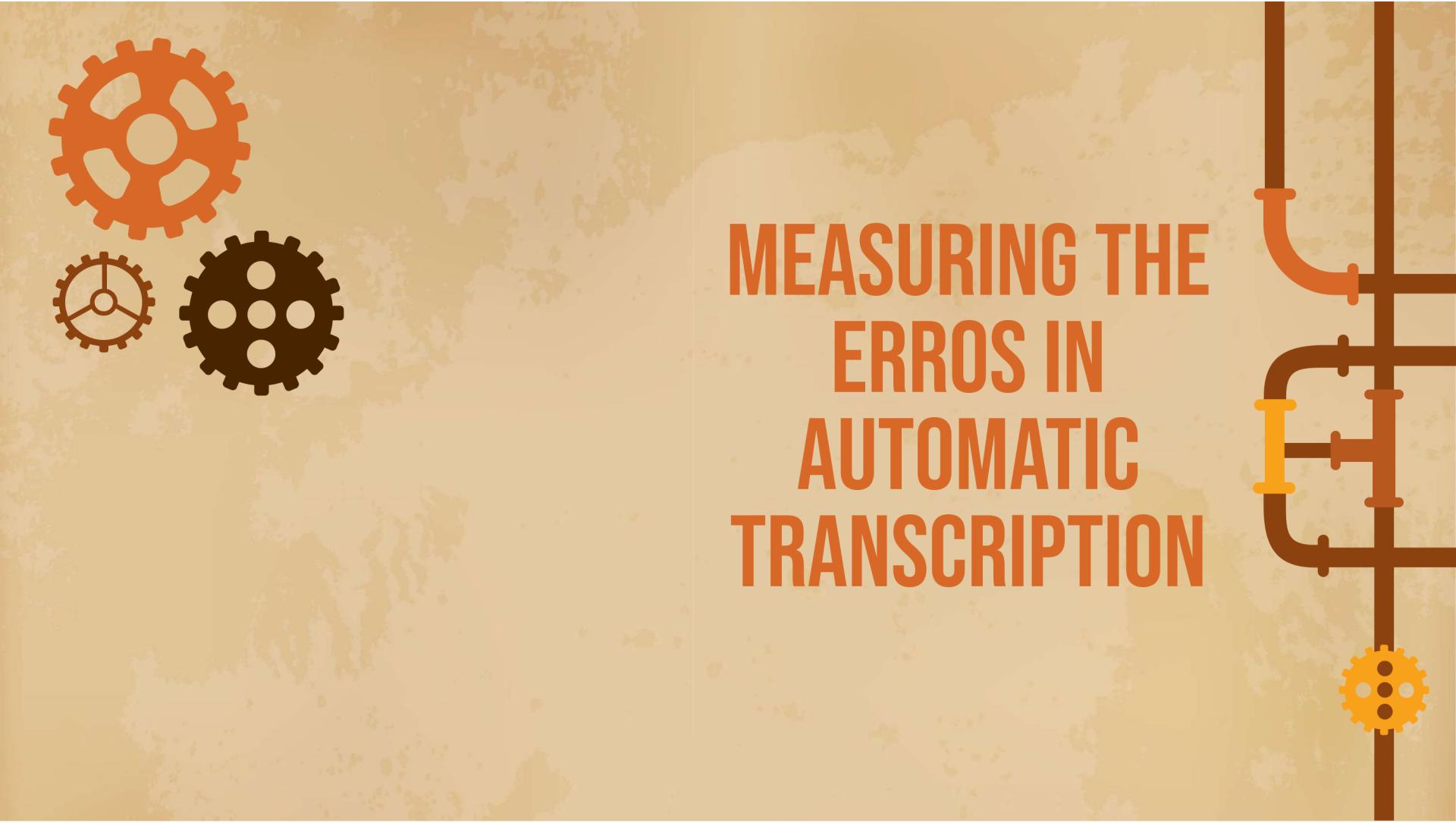
HTR (for manuscripts)

- 5-10% of character error
- Commercial software (Transkribus, ...) or open source software (Kraken / eScriptorium, ...)
- Require creating data to build a corpus-specific model

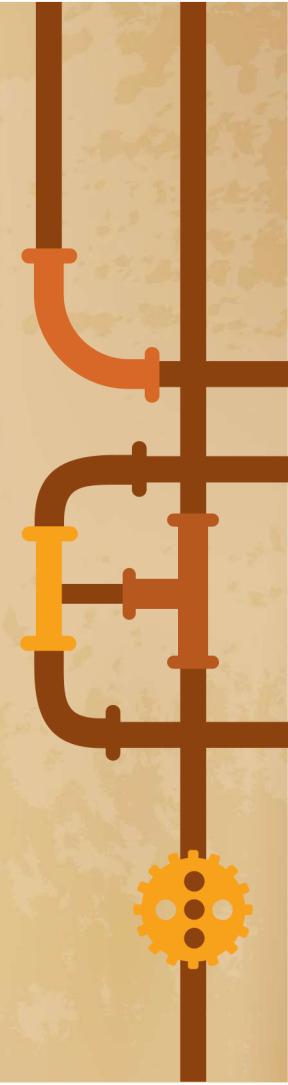
demands more skills from users of HTR!

Creating data and training a model





MEASURING THE ERROS IN AUTOMATIC TRANSCRIPTION



Measuring the success of a Text Recognition model

- Metrics are a useful (but imperfect) way to assess the success of HTR
- Understanding how we evaluate an HTR models helps understanding how it operates and what to expect from using HTR
- When assessing the performance of an HTR model:
 - **Character Error Rate**
 - **Word Error Rate**
- Computed by comparing a **reference text** and the model's **prediction** (ideally using the same segmentation)
- A model should not be tested on the data used to train it!

Types of errors

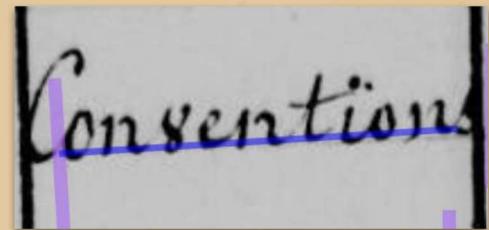
Categories to compute metrics:

- Substitution : STEAM → STEAL
- Deletion : STEAM → TEAM
- Insertion : STEAM → STREAM

$$CER = \frac{S + D + I}{N}$$

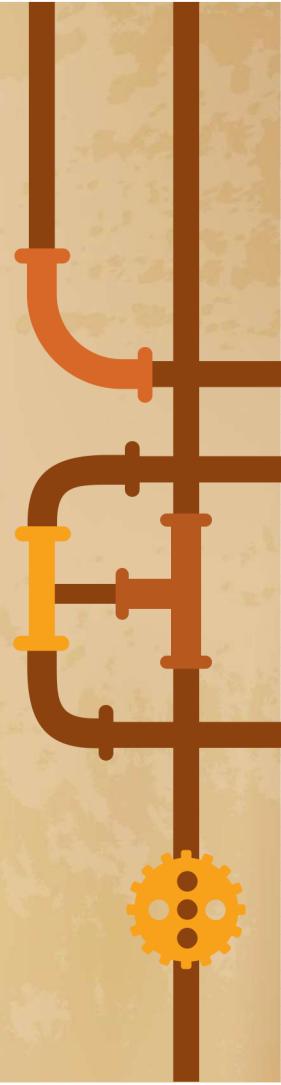
But there's more subtlety to it!

- Non-word errors (STEAN) vs. real-word errors
- Case-errors (StEAM) ; diacritic errors (STÉAM) ;
- Punctuation errors ; whitespace errors ; digits
- Many errors are caused by an imperfect segmentation (missing end of line, etc.)





**TRANSCRIBING IS
NOT SELF
EVIDENT**



Text Recognition is a Computer Vision task

- Relies on visual clues
- Doesn't *understand* the text, only sees it
- Produces what it is trained to produce

What is textual content?

15

Sur le port Mirabeau.

Sur le port Mirabeau court la Seine.
Et nos amours, faut-il qu'il n'en souvienne ?
Tua joie venant toujours après la peine.

| Viens le matin, comme l'heure,
| tes jours s'en vont, je demande.

| tes mains dans les mains, restas face à face
Tandis que sur le port de nos bras, passe
Des éternels regards l'anhe n'lace.

| Viens le matin, comme l'heure
| tes jours s'en vont, je demande.

| t'amour s'en va comme cette eau courante,
| t'amour s'en va; comme la vie est courte
Et comme l'Espérance est violente !

| Viens le matin, comme l'heure
| tes jours s'en vont, je demande

| Passeut les jours et passeut le remains,
| Ni temps, passé, ni les amours reviennent ;
Sur le port Mirabeau court la Seine

BNF
MS

| Viens le matin, comme l'heure,
| tes jours s'en vont, je demande !

C. 100 1. 111.

What is textual content?

- the main body of text (different paragraphs, title)

15

Sur le port Mirabeau.

Sur le port Mirabeau court la Seine.
Et nos amours, faut-il qu'il n'en souvienne ?
Tua joie court toujours après la peine.

Venue le matin, comme l'heure,
Les jours s'en vont, je demeure.

Tes mains dans les mains, restas face à face
Tandis que sur le port de nos bras, passe
Des éternels regards l'anhe n'lace.

Venue le matin, comme l'heure
Les jours s'en vont, je demeure.

T'amour s'en va comme cette eau courante,
T'amour s'en va; comme la vie est courte
Et comme l'Espérance est violente !

Venue le matin, comme l'heure
Les jours s'en vont, je demeure.

Panent les jours et panent le remains,
Un temps passé, un les amours reviennent;

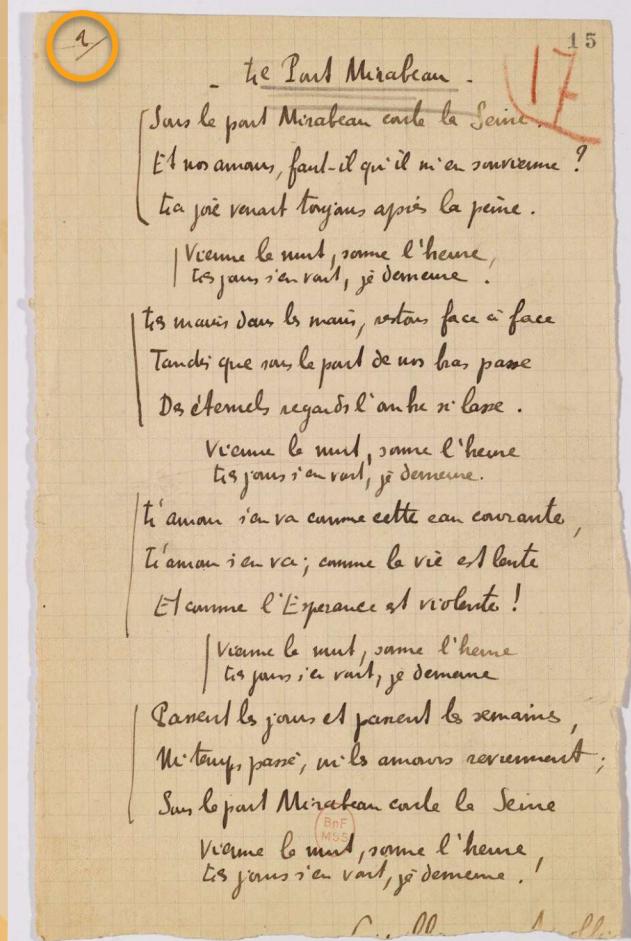
Sur le port Mirabeau court la Seine

BNF MS

Venue le matin, comme l'heure
Les jours s'en vont, je demeure !

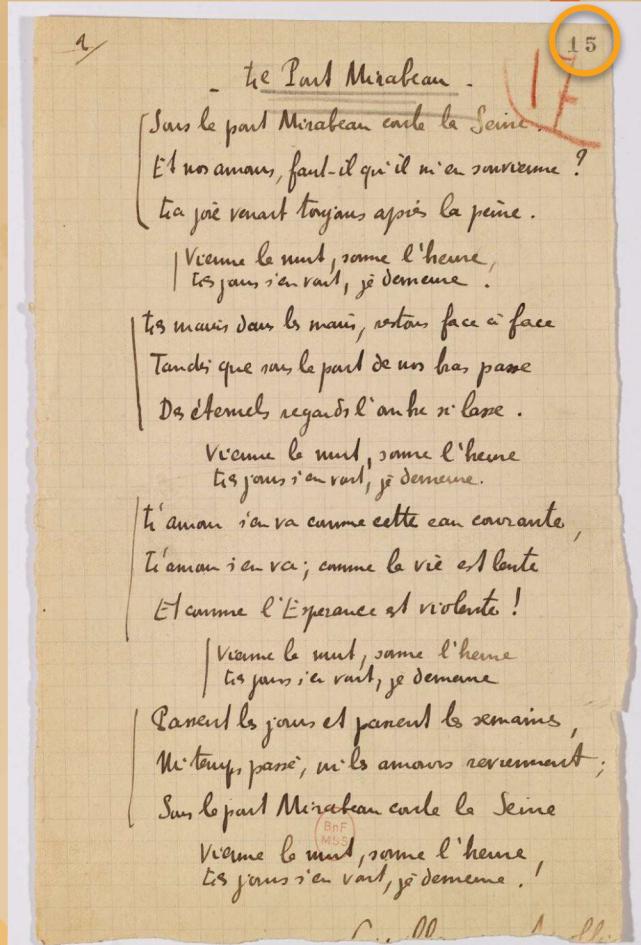
What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?



What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?



What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?
- the manual page/inventory number?

2
- Sur le port Mirabeau -
Sur le port Mirabeau court la Seine.
Et nos amours, faut-il qu'il n'en souffre ?
ta joie venant toujours après la peine.
Vient le nuit, come l'heure,
les jours s'en vont, je demeure.
les mains dans les mains, restons face à face
Tandis que sur le port de nos bras, passe
Des éternels regards l'anhe n'lace.
Vient le nuit, come l'heure
les jours s'en vont, je demeure.
L'amour s'en va comme cette eau courante,
L'amour s'en va; comme la vie est courte
Et comme l'Espérance est violente !
Vient le nuit, come l'heure
les jours s'en vont, je demeure.
Passent les jours et passent le remains,
Mé temps, passé, où les amours reviennent;
Sur le port Mirabeau court la Seine
Vient le nuit, come l'heure
les jours s'en vont, je demeure !
BNF MS

What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?
- the manual page/inventory number?
- the text in the library stamp?

15

Sur le port Mirabeau.
Sur le port Mirabeau court la Seine.
Et nos amours, faut-il qu'il n'en souvienne ?
Tua joie venant toujours après la peine.

| Vierme le matin, comme l'heure,
| tes jours s'en vont, je demande.
| tes mains dans les mains, restas face à face
| Tandis que sur le port de nos bras, passe
| Des éternels regards l'anhe n'lace.
| Vierme le matin, comme l'heure
| tes jours s'en vont, je demande.
| t'amour s'en va comme cette eau courante,
| t'amour s'en va; comme la vie est courte
| Et comme l'Espérance est violente !

| Vierme le matin, comme l'heure
| tes jours s'en vont, je demande

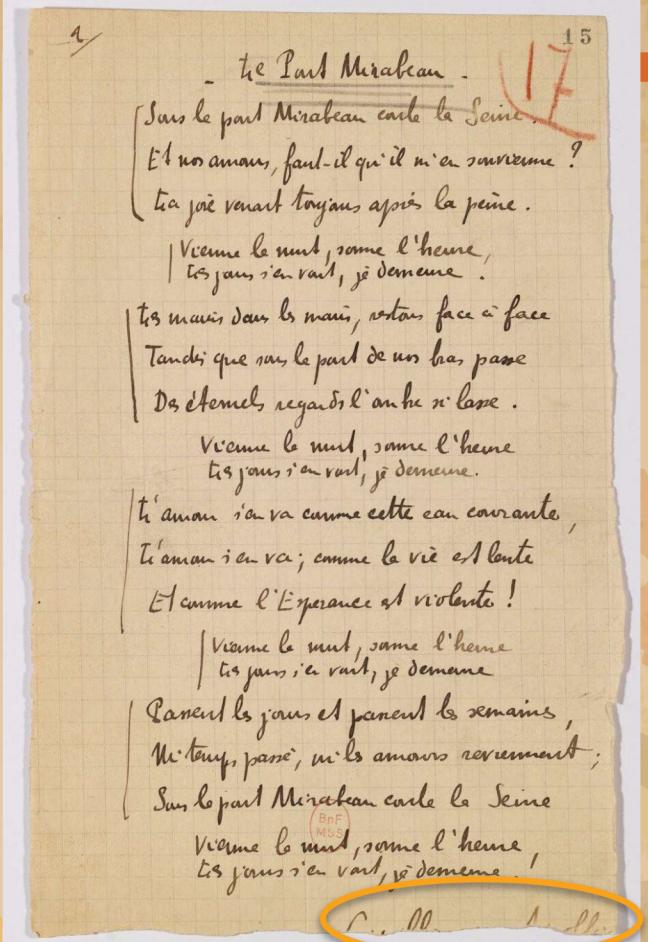
Panent les jours et panent le remains,
Mi temps, passé, mi les amours reviennent;
Sur le port Mirabeau court la Seine

Vierme le matin, comme l'heure,
Les jours s'en vont, je demande !

BNF
MS

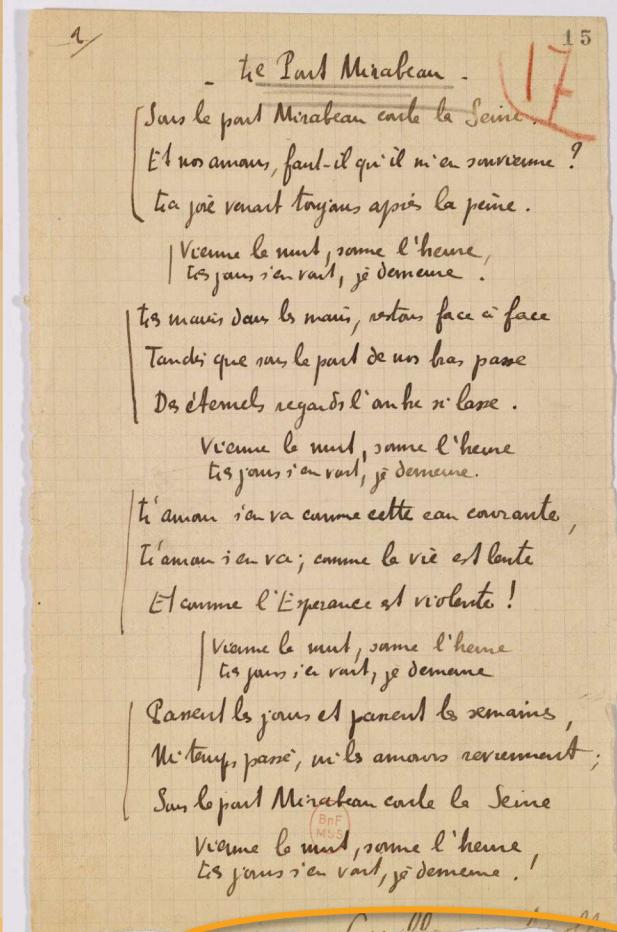
What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?
- the manual page/inventory number?
- the text in the library stamp?
- the partially illegible text?



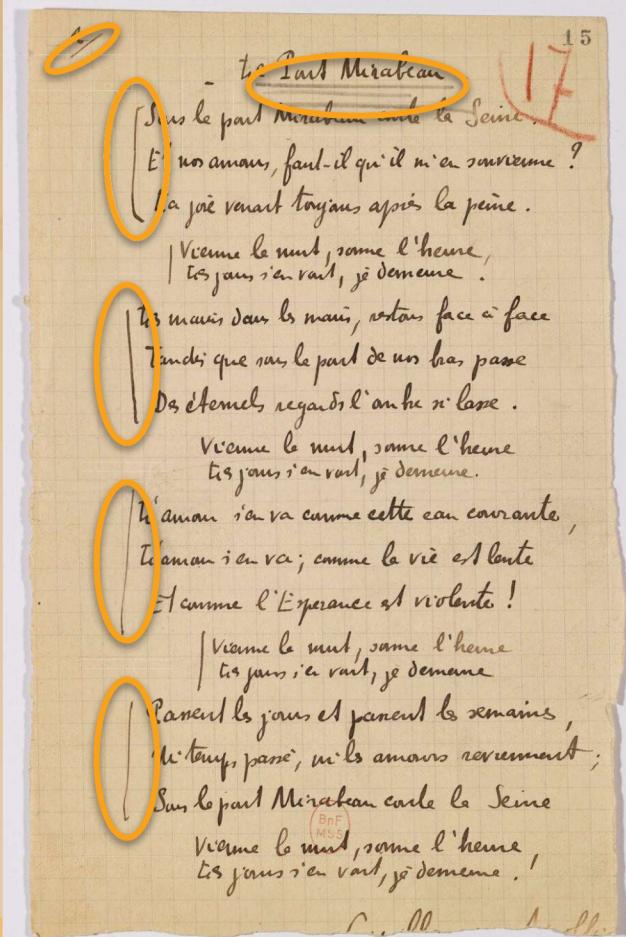
What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?
- the manual page/inventory number?
- the text in the library stamp?
- the partially illegible text?
- the credit line from the online library?



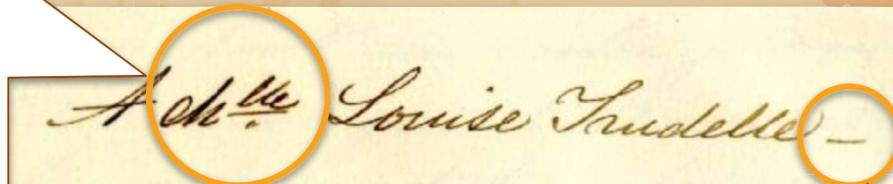
What is textual content?

- the main body of text (different paragraphs, title)
- the handwritten page number?
- the pre-printed page number?
- the manual page/inventory number?
- the text in the library stamp?
- the partially illegible text?
- the credit line from the online library?
- underlines? paragraph brackets?

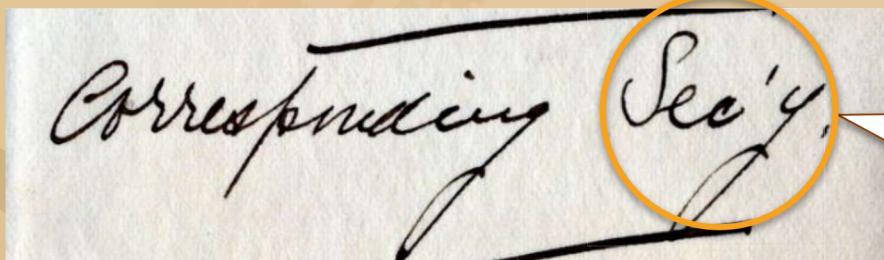


How do we transcribe?

- A **Mademoiselle** Louise Trudelle
- A **M^{lle}** Louise Trudelle
- A **M^{lle}** Louise Trudelle
- A **M^{lle}** Louise Trudelle
- A **M^{alle}** Louise Trudelle



A M^{lle} Louise Trudelle



Corresponding Sec'y.

- Corresponding **Sec'y.**
- Corresponding **Secretary.**

How do we transcribe?

was left please sent
it and I will paid
the **changed/s** charges

- the **changed/s** charges
- the charges
- the **changed/s** charges
- the **[[changed/s]]** charges
- the **[[]]** charges
- the **XXX** charges

Key takeaways

Transcribing is not self-evident!

- Transcription paradigm (regularized/graphemic/graphetic)
- Choice of characters to render different phenomena
- Transcription software usually don't include text formatting

In any case: document your choices, even more so if you are a team of 1 transcribers!!



TWO EXAMPLES OF ATTEMPTS TO HOMOGENIZE PRACTICES

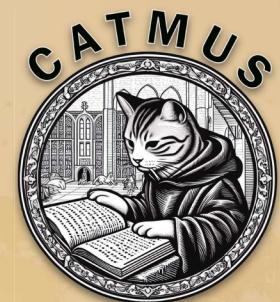


Community: transcription rules

Harmonization of data allows to exchange data and models!

How to ensure that?

- Define transcription methods adapted to a research problem AND to machine learning capacities
- Define the targeted precision/accuracy
- Define a set of characters to render the source and document choices
- Use recommendations from existing projects
 - **CATMuS** : <https://catmus-guidelines.github.io>

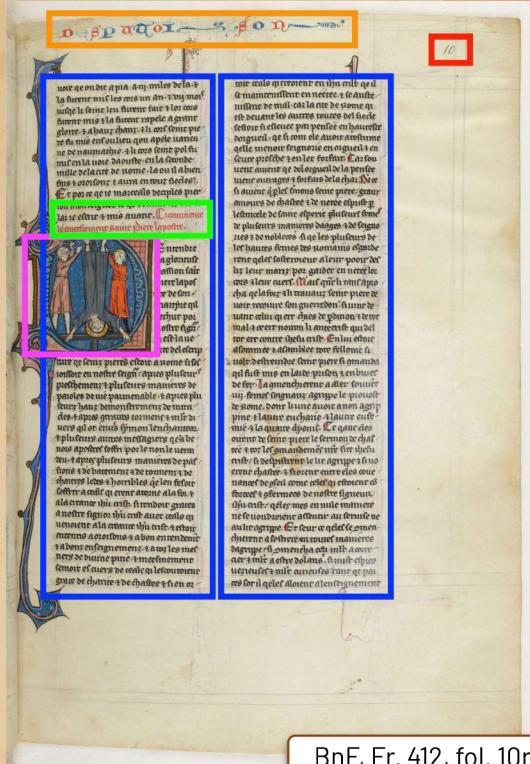


Community: layout description

SegmOnto provides guidelines and a controlled vocabulary to describe zones and lines on a page!

- **Lines:** DefaultLine, InterlinearLine
- **Pages:** MainZone, MarginZone, DropCapitalZone
- **Customization syntax:** Name:subtype#number
- **Documentation:** <https://segmonto.github.io/>

S



BnF, Fr. 412, fol. 10r



The background features a light beige color with a faint, textured map of the world. In the upper left corner, there are three stylized gears: one large orange gear at the top, one small grey gear below it, and one medium black gear to its right. In the lower right corner, there is a network of industrial pipes in various shades of brown and orange, with a single yellow pipe segment.

**WHAT SOLUTION
OFFERS
HTR-UNITED?**

Why is HTR-United important?

- Need to rely on pre-existing models and pre-existing data
- There is no centralized data warehouse for (gold) HTR datasets
- They are rarely “FAIR”:
 - Hard to **Find** and not always **Accessible**
 - uncertain formats and varying annotations rules
 - unclear **Reuse** conditions



[https://www.go-fair.org/
fair-principles/](https://www.go-fair.org/fair-principles/)

Make your data findable!

- Emphasize key aspect of sharing data:
 - Data quality and use of standards
 - Documentation
 - Citation
 - Reusability



The logo consists of two overlapping green speech bubble shapes. The top shape contains the letters "HTU" in a bold, black, sans-serif font. The bottom shape contains the word "United" in a similar black font. Both shapes have a thin blue outline and a slightly irregular, torn-paper-like edge.

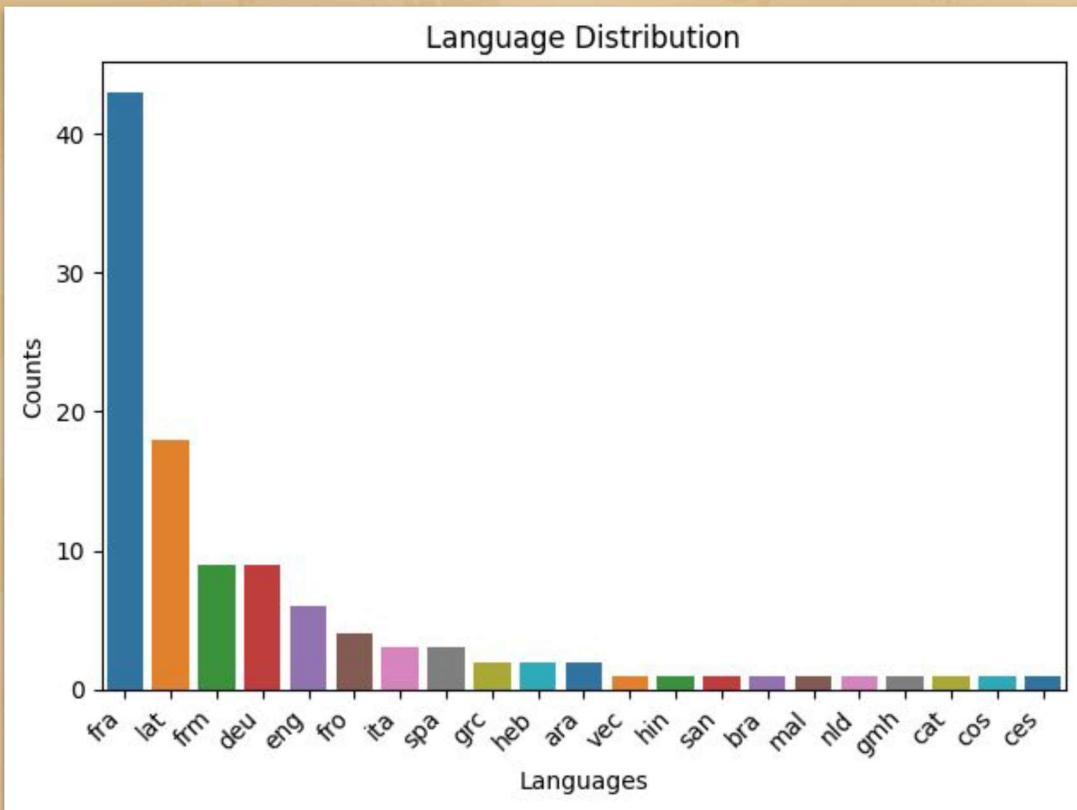
HTR-United in details

- Editors: Alix Chagué & Thibault Clérice
- A catalog browsable by humans and by machines
- Low tech environment for easy maintenance
- An ecosystem of recommendations and tools to ensure data quality
- A catalog fed by data creators
 - Data is published wherever they want
 - A form is filled online to create a structured description
 - The description is reviewed by editors, then added to the catalog
- HTR-United focuses on datasets, not on transcription models

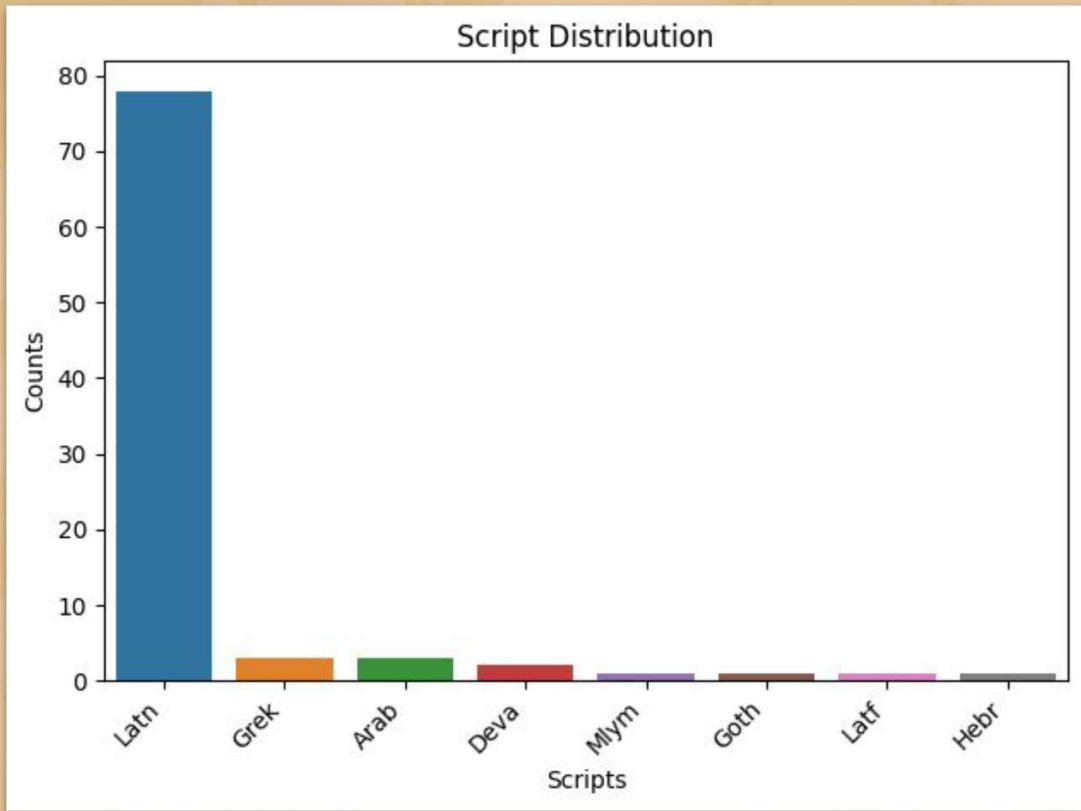
Content of HTR-United (as of April 2024)

- 83 datasets created by at least 38 different projects
- 21 languages (a lot of French (43) and Latin (18)) for 8 scripts (mostly Latin)
- Handwritten and printed documents, mixed or not
- A handwriting-period coverage spanning from 800 to 2023
- Created with at least 6 different software (eScriptorium, Kraken, Transkribus ...)
- more than 43M characters, 1M lines or 11K images

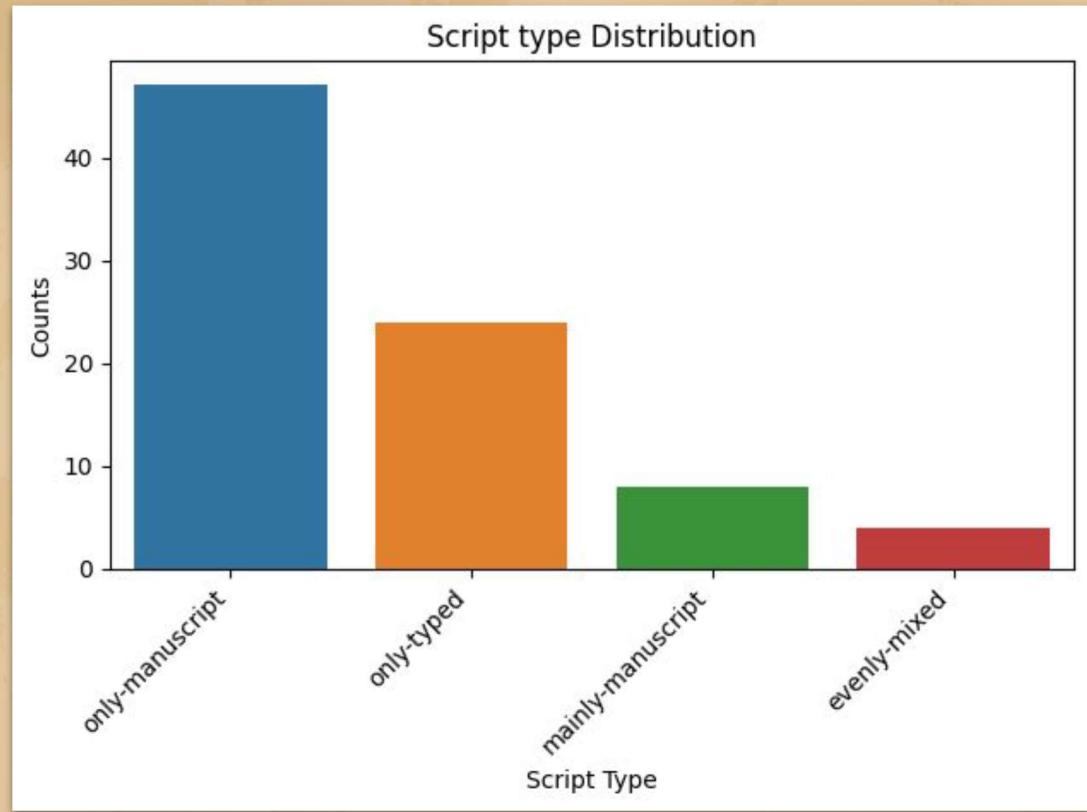
Content of HTR-United (as of April 2024)



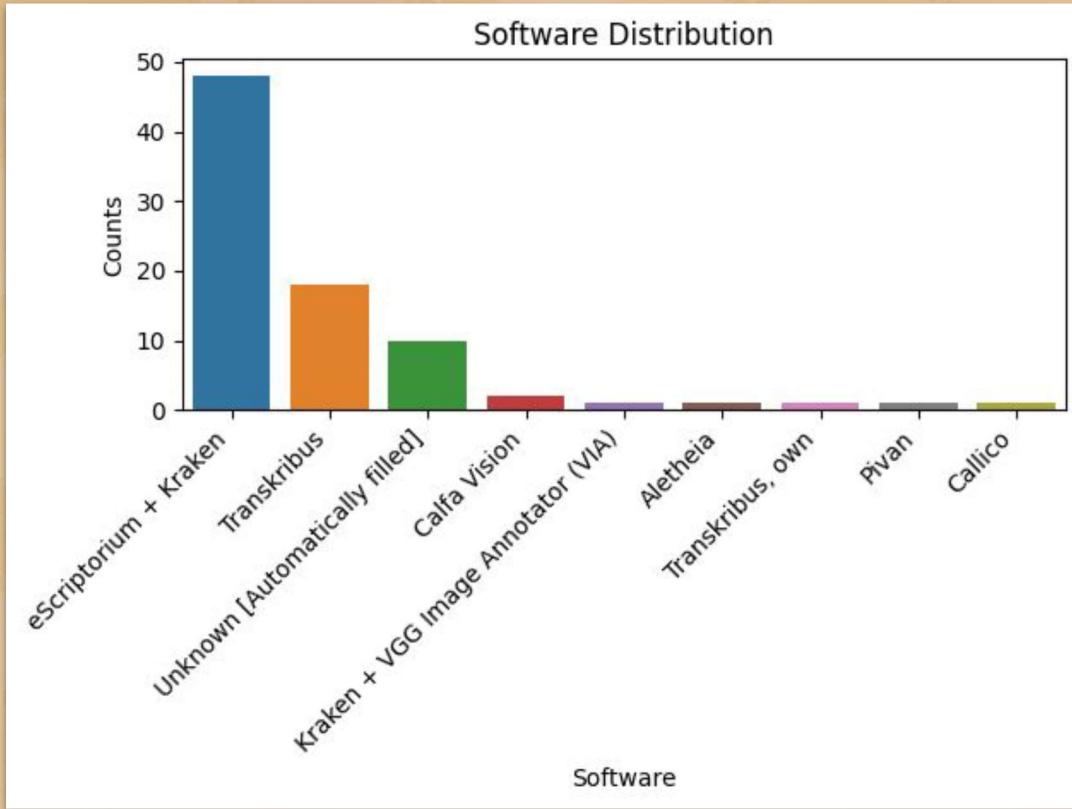
Content of HTR-United (as of April 2024)

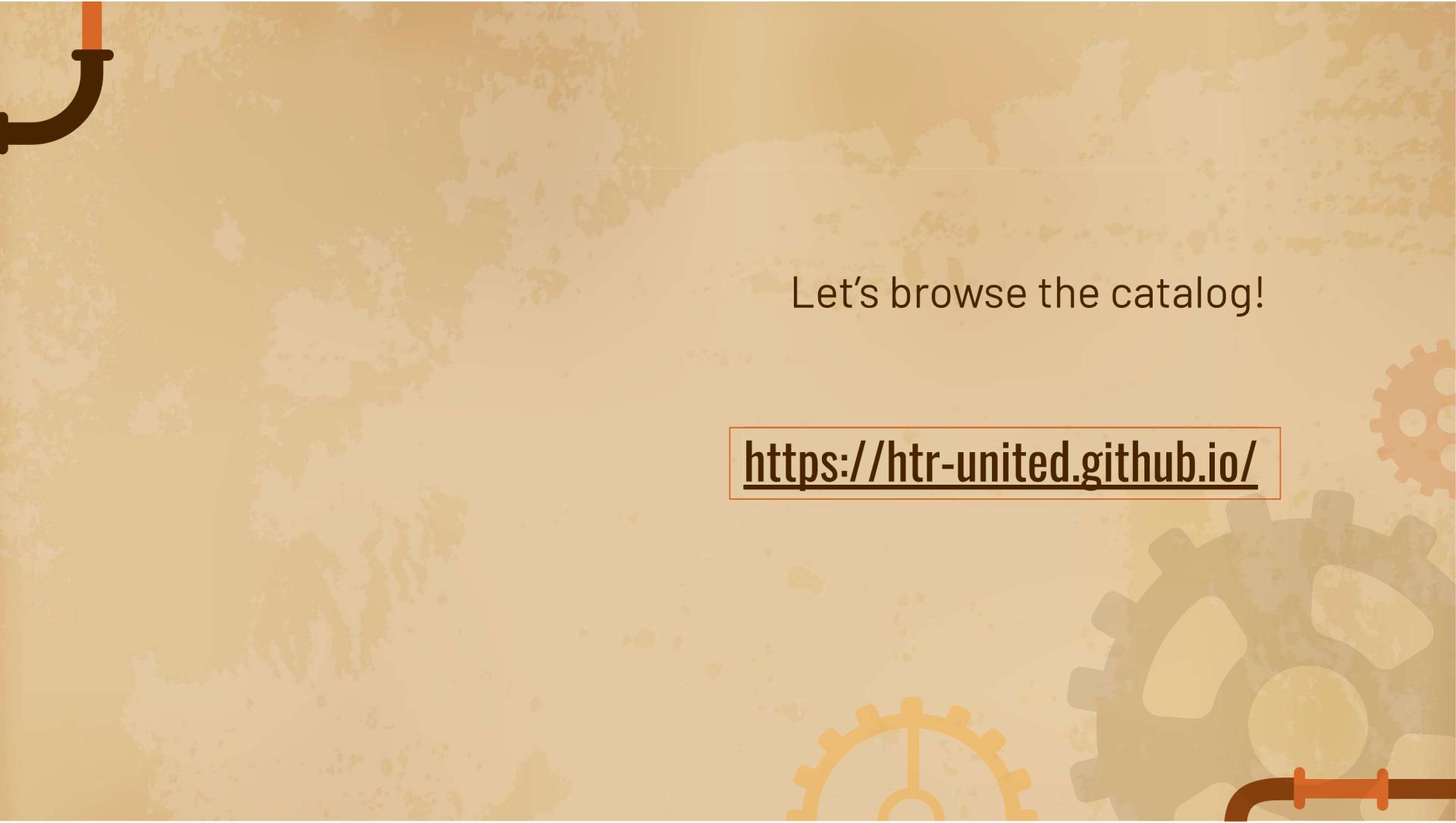


Content of HTR-United (as of April 2024)



Content of HTR-United (as of April 2024)



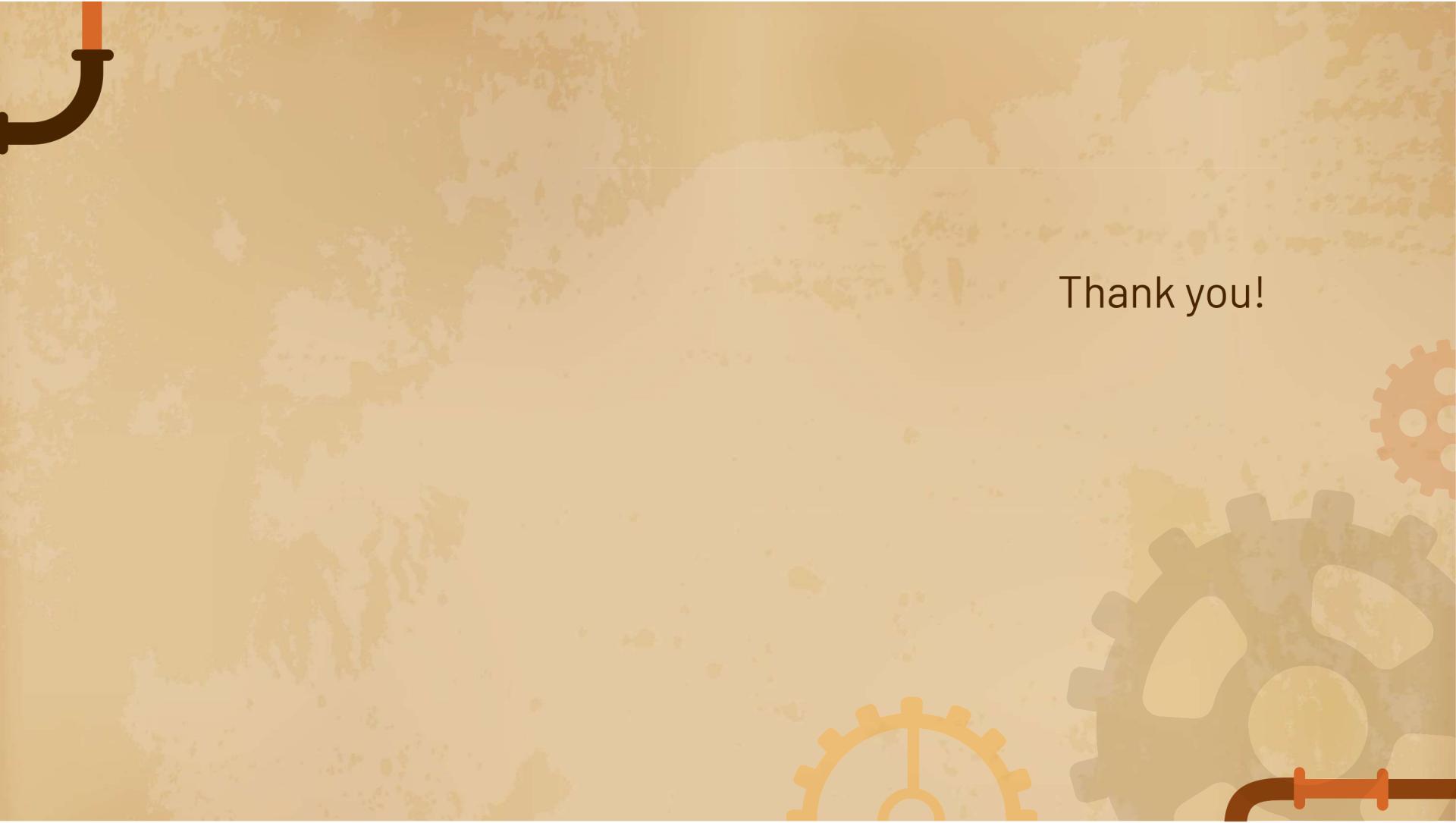
The background of the slide features a textured, light brown or beige color with faint, dark brown industrial motifs. These include a thick vertical pipe on the left, several horizontal pipes at the bottom right, and several large, interlocking gears of different sizes scattered across the lower half of the slide.

Let's browse the catalog!

<https://htr-united.github.io/>

Key takeaway

- HTR technology gains from users forming a community
 - learn faster
 - avoid mistakes
 - homogenize practices
- HTR-United intends to facilitate the creation of this community
 - reliable data
 - community-organized tool
 - transparent process



Thank you!