

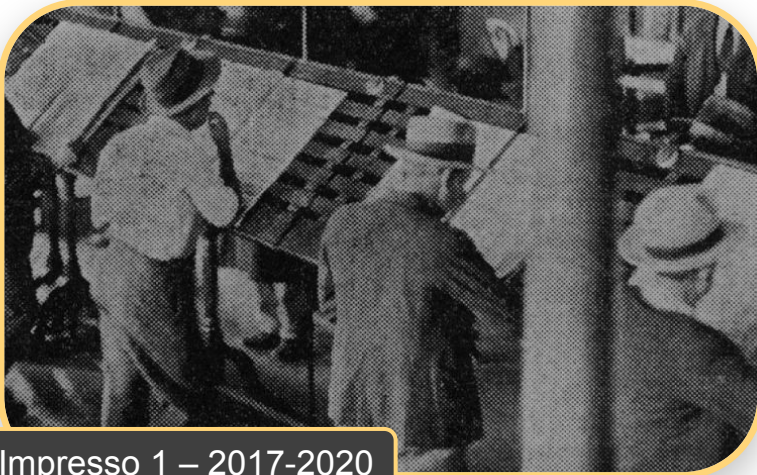
Media Monitoring of the Past – *Beyond Borders*

Marten Düring, C2DH
Maud Ehrmann, EPFL-DHLAB
& Impresso Team

ONB Labs Symposium

25.11.2024 - Vienna

CC-by Marten Düring, Maud Ehrmann, Impresso Team



Impresso 1 – 2017-2020



Impresso 2 – 2023-2027

Developing new approaches and interfaces
for the exploration and critical analysis of
historical media archives.

NLP / History / Design



Impresso 1 – 2017-2020

Where we come from

***How to enable semantic indexing
and exploration of large collections
of historical newspapers?***



Impresso 2 – 2023-2027

Our present focus

The Impresso Web App

IMPRESSO DATA RUNDOWN

76 newspapers, 2 countries
600,919 issues,
5,429,656 pages,
47,798,468 content items,
3,462,799 images,
12,493,358,703 tokens.



High-level objectives for historians

Help me search and discover relevant content → Fine-grained search and rich, faceted information retrieval

Help me explore large collections → Scalable reading

Help me understand what I am working with → Data and tool criticism

Help me contextualise what I find → Comparative perspectives

The Impresso Web App

Language identification
OCR QA
Word embeddings
Named Entities
Topic Modeling
Article classification
Ngrams
Text reuse
Image similarity



The screenshot shows the Impresso web application interface. At the top, there is a navigation bar with links for 'Search', 'Newspapers', 'Inspect & Compare', 'Text reuse', 'Help', 'LOGIN', and 'REGISTER'. Below the navigation bar, there are three main sections: 'SEARCH ARTICLES', 'SEARCH IMAGES', and 'NGRAMS'. The 'SEARCH ARTICLES' section has a search bar and a dropdown menu. The 'SEARCH IMAGES' section has a search bar and a dropdown menu. The 'NGRAMS' section has a search bar and a dropdown menu. In the center, there is a large heading 'Media Monitoring of the Past' with a subtitle 'Mining 200 years of historical newspapers, and radio.' Below this, there is a section 'Learn Impresso with the Impresso Challenges' with a 'DOWNLOAD CHALLENGES PDF' button. To the right, there is a section 'How can newspapers help understand the past? How to explore them?' with a 'Just a few examples to get you started!' text. At the bottom, there are two search suggestions: 'Search for the composer Robert Schumann' and 'Find performances of Robert Schumann's music'.

Content search and discovery based on semantic enrichments.

Comparative and critical perspectives on the data.

GROUP BY | ARTICLE ▾

ORDER BY RELEVANCE ▾

≡ "AUTOROUTE OR AUTOURROUTE(▼

SWITZERLAND ARTICLE

CONSEIL • INITIATIVE • LOI • PEL

add keyword to search

PERSON (278 OPTIONS) ⓘ

LOCATION (369 OPTIONS) ⓘ

TOPIC (100 OPTIONS) ⓘ

results are filtered when:

RESET

CONTAINING

FR conseil • initiative • loi • peuple • projet (894 results)

☐ FR construction • route • place • projet • ville (522 results topic)

☐ FR problème · fait · question · exemple · monde (303 results topic)

☐ FR budget · impôt · conseil · déficit · taxe (265 results topic)

☐ FR conseil • commission • projet • loi • rapport (208 results topic)

☐ FR fer · ligne · chemin · voie · trafic (181 results topic)

(175 results topic)

894 articles found containing **autoroute** or **autouroute** or **autoroule** or **contournement** or **autoroutes** appearing on the front page - tagged as article; with topic CONSEIL · INITIATIVE · LOI · PEUPLE · PROJET; printed in **Switzerland**

LESS ...

COMPARE ...

SAVE / EXPORT

BULLETIN SUISSE
Affiner la
démocratie

Personal use — provided by Swiss National Library

LOCATIONS Suisse, Moselle, Zurich Open

de vitesse autorisées sur les routes et les **autoroutes** ; ensuite, toute une série de domaines se situant

SAVE TO COLLECTION ...

VOTATIONS

Journal de Genève — Monday, April 2, 1990 — p.1

LOCATIONS Yverdon-les-Bains, Biel/Bienne, Solothurn, Neuchâtel Xamax, Jura Mountains, Suisse, Moselle, Zurich
 Airport, Gare de Cornavin

PEOPLE Jean-Pascal Delamuraz

« pour une région sans autoroute entre Morat et Yverdon » (67,3 % de non), « pour un district du Knonau

sans autoroute » (68.6 % de non), et « contre la construction d'une autoroute entre Bienne et Soleure / Zuchwil »

s'achever le réseau d'autoroutes. Fluidité du trafic oblique ! En ce qui concerne la NI entre Morat

SAVE TO COLLECTION

FILTERS

"AUTOROUTE OR AUTOUROUTE"

article text

☒ Contains

☐ NOT contains

☐ all of the following

☒ at least one of the following

☒

autoroute

☒

autouroute

☒

autoroule

☒

contournement

☒

autoroutes

ADD NEW ...

ADD SIMILAR ...

RREMOVE

KW SUGGESTION

FR budget • impôt • conseil • déficit • taxe (265 results topic)

FR conseil • commission • projet • loi • rapport (208 results topic)

FR fer • ligne • chemin • voie • trafic (181 results topic)

FR canton • projet • développement • recherche • région (175 results topic)

UP BY **ARTICLE** ▾

articles found containing **autoroute** or **autouroute** or **autoroule** or **autournement** or **autoroutes** appearing on the front page - tagged as article; with **CONSEIL · INITIATIVE · LOI · PEUPLE · PROJET**; printed in **Switzerland**

TOPIC ✕

FR conseil · initiative · loi · peuple · projet

25368

☐ Apply current search filters (6)

1,412,710 articles with topic

All results fall between 1740 and 2018

TOP WORDS IN TOPIC

conseil · initiative · loi · peuple · projet · canton · vote · référendum · droit · oui (FRENCH)

MORE DETAILS...

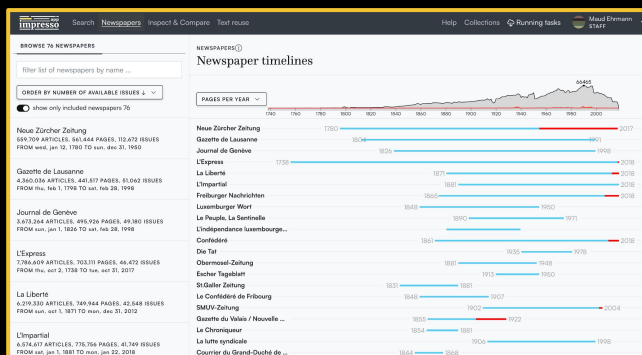
APPLY FILTER

CLOSE

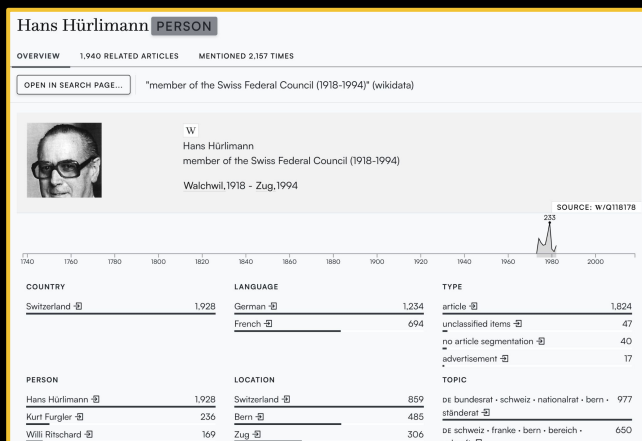
[illegible]

INSPECT SEM. ENRICHMENT

SCALABLE READING



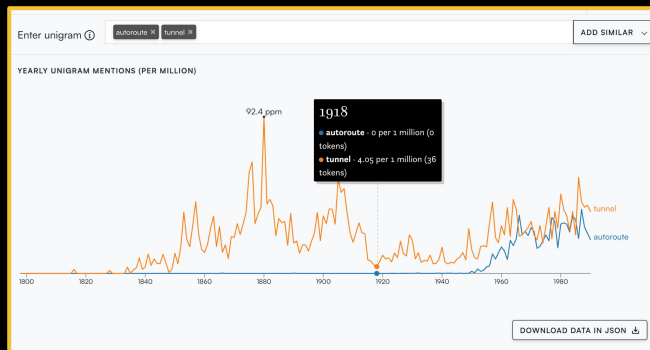
CORPUS



NAMED ENTITY PROFILES



TEXT REUSE



N-GRAMS

DATA AND TOOL CRITICISM

DATA PROCESSING AND TEXT ANALYSIS

What are OCR, OLR, METS/ALTO and ASR?

OCR and OLR describe the automated recognition of text and article layouts. METS and ALTO are xml standards used to store this information together with other metadata. ASR is the process and related technology for converting spoken language into text.

1. **OCR (Optical Character Recognition):** OCR is the automated process of identifying and converting printed text from scanned images into machine-readable text. In addition to recognizing letters and characters, OCR tools can detect elements like images, tables, and preserve the coordinates of each character within the scanned image. These coordinates allow features such as keyword highlighting in facsimile view. (MORE ...)

How did you identify content types?

The identification of article types happens during the Optical Layout Recognition (OLR) phase of the newspaper digitization process.

Specific object types such as articles, obituaries, and advertisements are tagged based on visual features present in the scanned newspapers. This process was supervised by the partners of Impresso and executed by external service providers for most cases. Each partner defined the categories of newspapers articles that the external service provider should tag. As a result, Impresso inherits these predefined categories, which vary from (MORE ...)

What is named entity processing?

Named entities correspond to elements of interest that appear in texts, usually of type Person, Location or Organisation.

They are referential units which underlie the meaning of texts. Their definition is quite vague, but as a rule of thumb, they correspond more or less to proper names. The important criteria is here the name, which acts as a 'rigid designator': to differentiate for instance a common definite description such as 'the pope' referring to any pope, from the mention of a particular named entity such as 'Pope Francis'. (MORE ...)

EXTENSIVE FAQ (just revised!)

rist, Basel

ute • place • p

z • électricité •

available ⓘ

AG) à S

ronçon

What is text reuse and what can I do with it?

Text reuse detection refers to the automated identification of repeated text passages within a corpus, such as copies of the same article published across different newspapers.

MORE INFO →



INFORMATION ON DATA LEAKAGES

INFORMATION ON TOOLS

CONTRASTIVE VIEWS

impresso Search Newspapers Inspect & Compare Text reuse Help Collections Running tasks Maud Ehrmann STAFF

QUERY

frontpage **AUTOROUTE OR AUTOURROUTE**

SWITZERLAND

PARTI - CONSEIL - VOIX - ÉLECT

search for ...

OPEN IN SEARCH PAGE... (315 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

L'impartial	95
La Liberté	60
Confédéré	39
La Peuple, La Sentinelle	38
L'Express	36
Gazette de Lausanne	28
Journal de Genève	18
La lutte syndicale	1

297 results in common

Lists of newspapers, named entities and topics for results for (A), (B) and in both (A) and (B)

OPEN IN SEARCH PAGE... (297 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

L'impartial	91
La Liberté	48
Confédéré	39
La Peuple, La Sentinelle	38
L'Express	36
Gazette de Lausanne	27
Journal de Genève	18
La lutte syndicale	1

93,655 RESULTS

OPEN IN SEARCH PAGE... (93,655 RESULTS)

YEAR OF PUBLICATION

Total number of articles per year

NEWSPAPER

La Liberté	27,549
L'Express	21,156
L'impartial	14,520
Gazette de Lausanne	12,868
Journal de Genève	12,317
Confédéré	2,437
Le Peuple, La Sentinelle	1,276
Solidarité	618
L'Enteiburger Land	530
Freilivene Nachrichten	298

INSPECT AND COMPARE QUERY RESULTS



Impresso 1 – 2017-2020

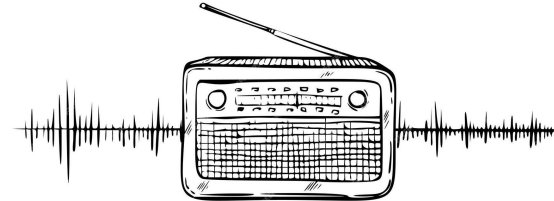


Key motivations for Impresso 2

Diversifying Media Types: Broadcast archives are also digitized but remain harder to access.

Overcoming Collection Silos: Digital archives are often limited by institution, media type, and language.

Supporting Computational Historical Research: Static exploration tools fall short for versatile, comparative research.



impresso

Media
Monitoring of the Past



Beyond Borders: Connecting Historical Newspapers and Radio

EPFL



University of
Zurich ^{UNZH}

Unil

UNIL | Université de Lausanne



Swiss National
Science Foundation



Luxembourg
National
Research Fund

- Enrichment and integration of newspaper and radio sources in a single semantic space;
- Collections from 20 European partners;
- Interfaces for exploratory and computational research;
- Case studies in (media) history

13 collaborators in NLP, History and Design | Sept 2023 - Feb 2027

Impresso 'doppio' Associated Partners

National or state libraries

- Bibliothèque Nationale Suisse, BN
- Bibliothèque Nationale du Luxembourg, BNL
- Bibliothèque Cantonale et Universitaire Lausanne, BCUL
- Österreichische Nationalbibliothek, ONB
- Staatsbibliothek zu Berlin, SBB
- The British Library (BL)
- Bibliothèque nationale de France, BnF
- Staats- und Universitätsbibliothek Hamburg, HUB
- Koninklijke Bibliotheek van België, KBR
- Koninklijke Bibliotheek, KB

Newspapers

- Le Temps
- Neue Zürcher Zeitung

Audiovisual heritage institutions and archives

- Radio Television Suisse (RTS)
- Österreichischer Rundfunk, ORF
- British Broadcasting Corporation (BBC)
- DeutschlandRadio
- Institut National de l'Audiovisuel, INA
- Nederlands Instituut voor Beeld en Geluid, NISV

Research Networks

- Entangled Media Histories Research Network for European media historians (EMHIS)
- Memoriav (Swiss network for audiovisual CH)
- infoclio.ch

Processing & NLP



Maud Ehrmann, EPFL

(Digital) History



Arthur Michelet, UNIL

Processing & NLP



Pauline Conti, EPFL

Processing & NLP



Juri Opitz, UZH

(Digital) History



Marten Düring, C2DH

UX & Design



Daniele Guido, C2DH

Processing & NLP



Emanuela Boros, EPFL

Processing & NLP



Simon Clematide, UZH

(Digital) History



Raphaëlle Rupen Coutaz, UNIL

Processing & NLP



Andrianos Michail, UZH

(Digital) History



Estelle Bunout, C2DH

UX & Design



Roman Kalyakin, C2DH

(Digital) History



Kaspar Beelen, C2DH (associate)

(Digital) History



Martin Grandjean, UNIL

(Digital) History



Caio Mello, C2DH

(Digital) History

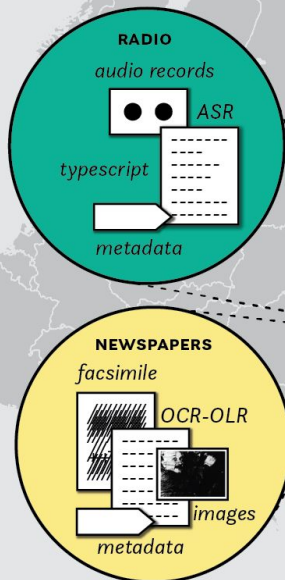


Ferdaous Affan, C2DH

CO-by Marten Düring, Maud Ehrmann, Impresso Team

1 Source collection

European media archives

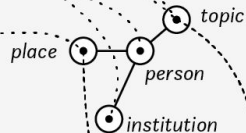
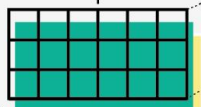


2 Media processing

Enriching & connecting

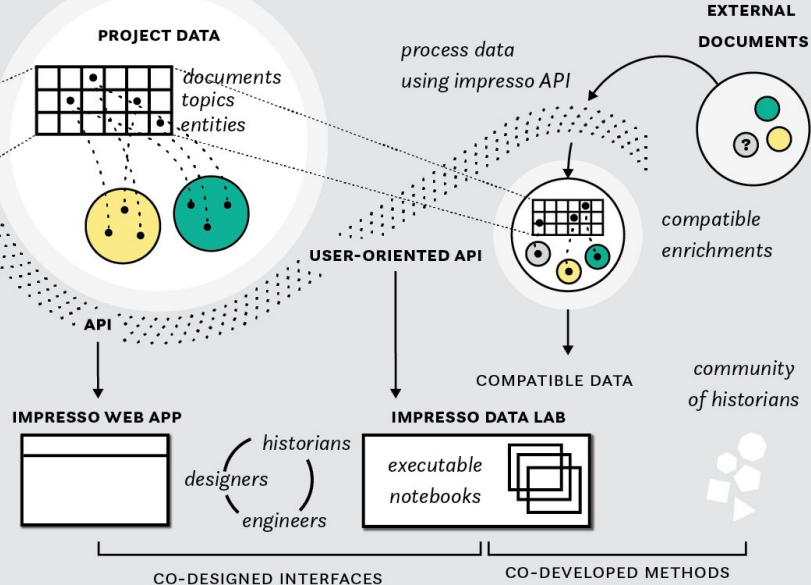
SEMANTIC ENRICHMENT
ACROSS LANGUAGES
ACROSS MEDIA

dense vector
representations



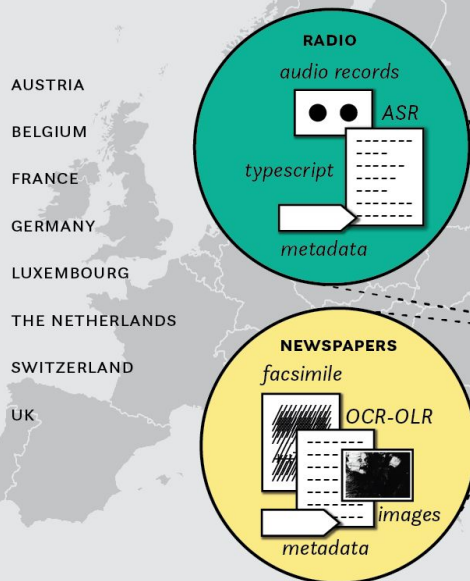
3 Media exploration

Connected and comparable enriched media sources



1 Source collection

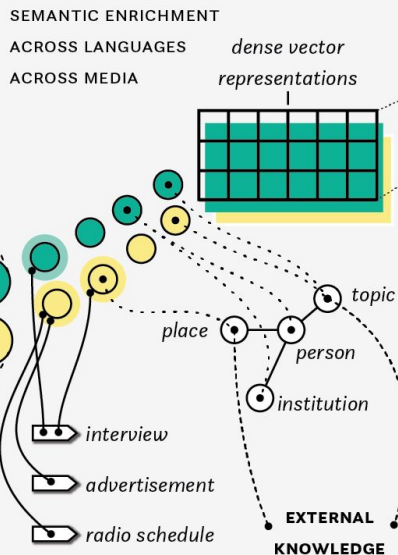
European media archives



Transnational and transmedia corpus

2 Media processing

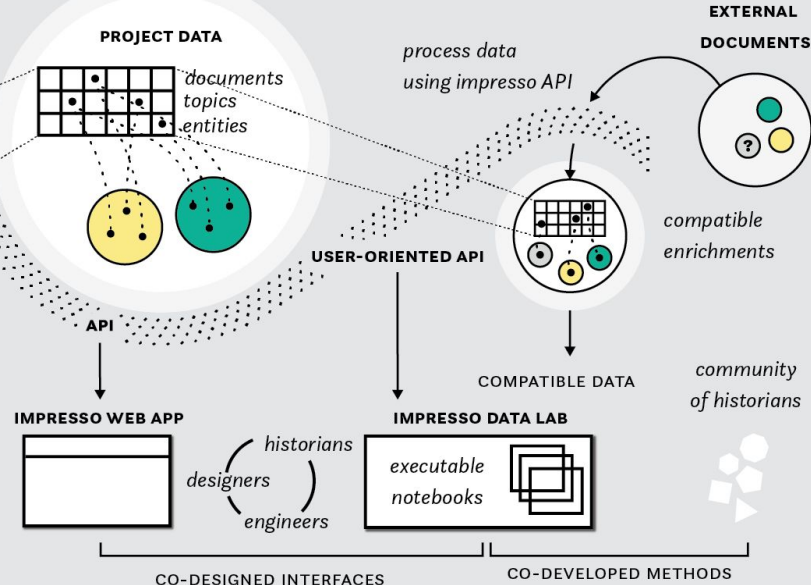
Enriching & connecting



Semantic connectivity across language, time, modality

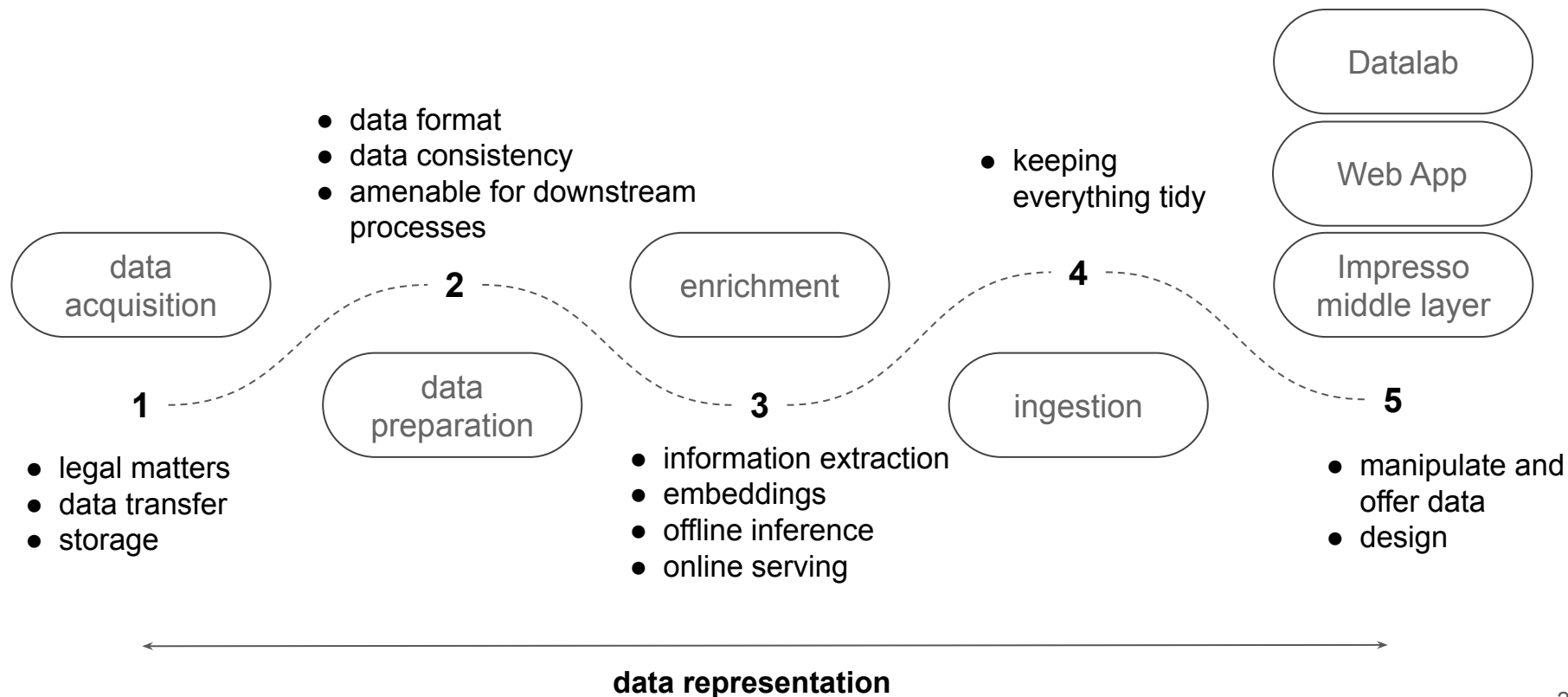
3 Media exploration

Connected and comparable enriched media sources

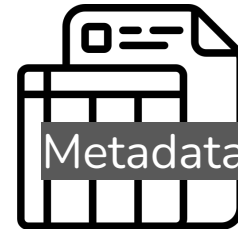
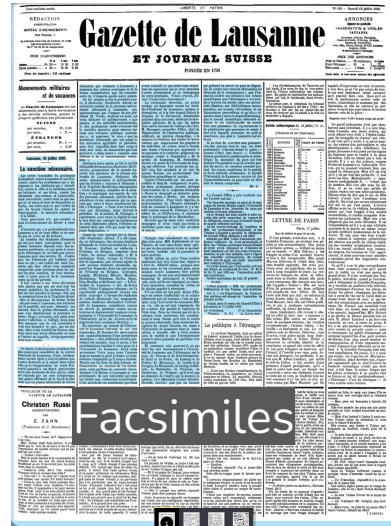
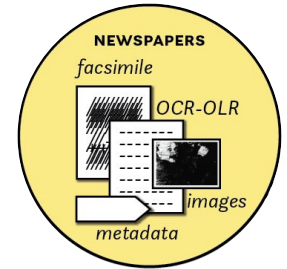


Data and tools closer to users

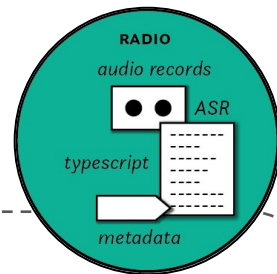
Data journey & Processes (in a nutshell)



Source material - The world of Newspapers



Source material - The world of Radio



RADIO BROADCASTS



Audio recordings

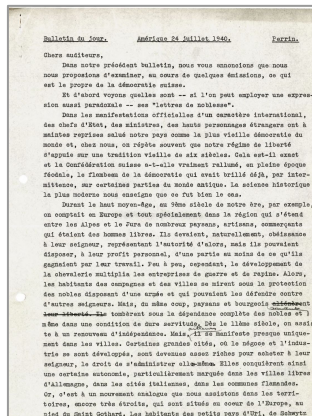


ASR



Metadata

RADIO TYPESCRIPTS



Facsimiles, OCR
& Metadata

RADIO SCHEDULES

...by radio channels

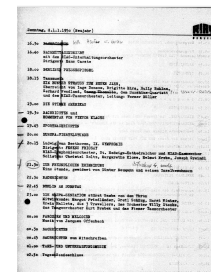


...in newspapers

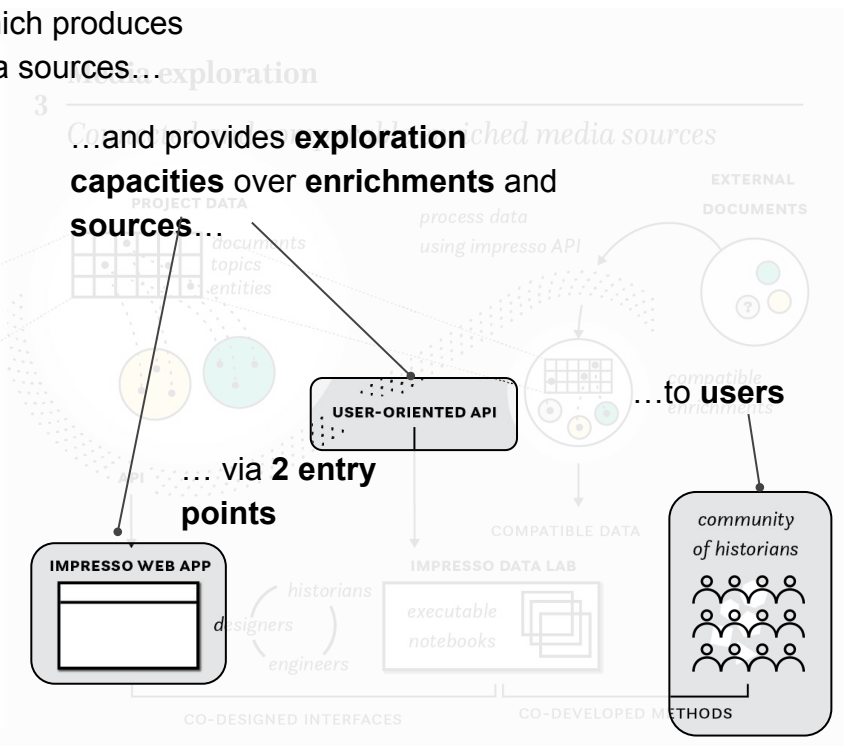
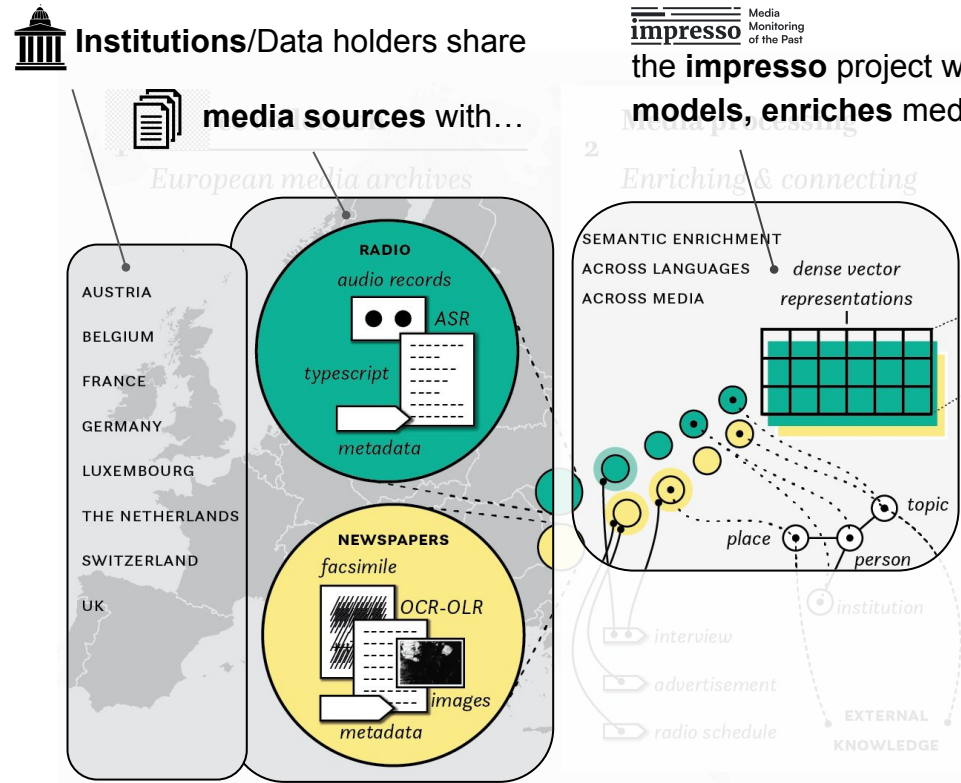


Facsimiles, OCR
& Metadata

...for internal use



Acquisition of a Historical Media Corpus





Institutions

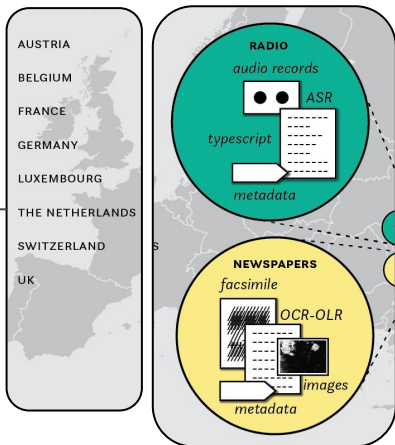
**Data Sharing
Agreement**

impresso

Media
Monitoring
of the Past

Terms of Use

users



• **different jurisdictions**

different constraints

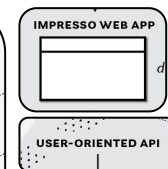
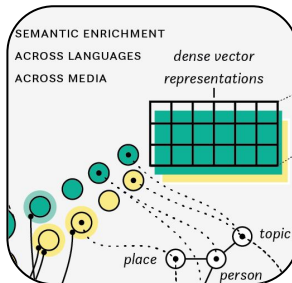
according to **data owner** and
institutional policies

• **different data elements**

- metadata
- facsimiles
- audio records
- transcripts (OCR/ASR)

different copyright statuses

- public domain
- under copyright
- grey zone

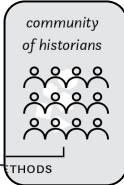


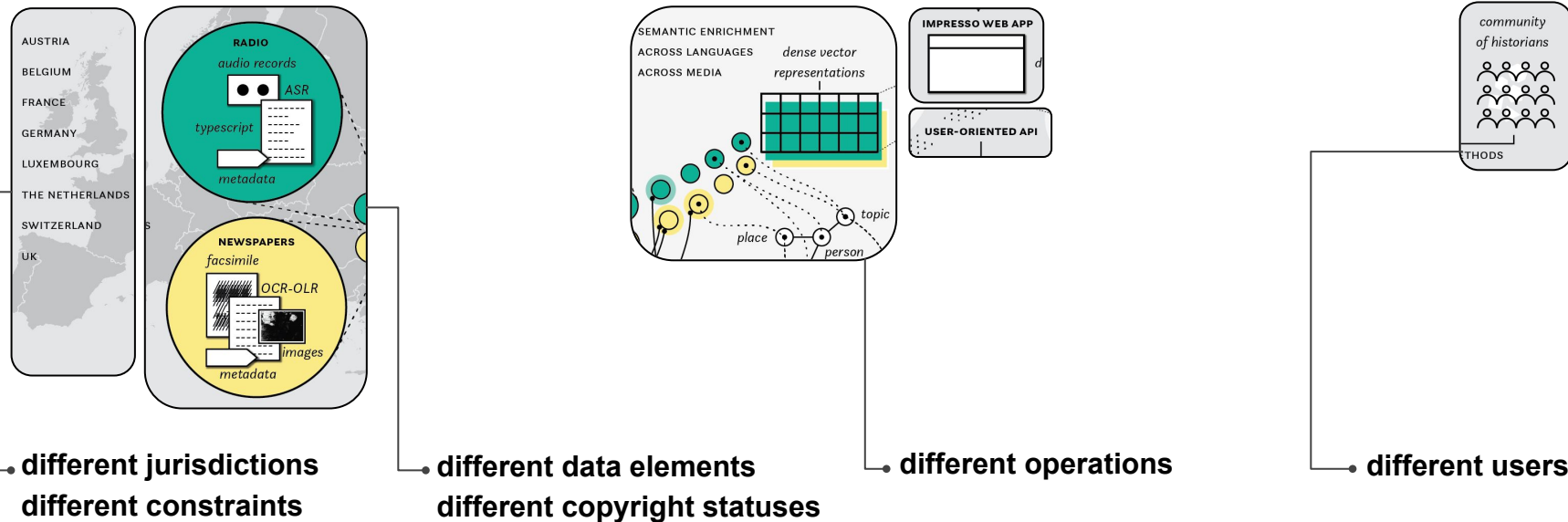
• **different operations**

- view, read, explore
- export via web app & API

• **different users**

- researchers
- students
- journalists
- members of GLAM
- citizens





Objective

Find a way to share and use data that:

- 1) respects copyright and institutional constraints,
- 2) provides researchers with maximum research opportunities.

Approach

Modular **data sharing and access framework** through **differentiated access** based on user status, data type and authorised operations.

Implementation

Several dimensions:

- **Social:** extensive exchanges
- **Legal:** modular data access policy

	Data domain		Data elements		Operations		User status			
Plan	Guest		Basic		Student		Researcher		Special Archive Membership	
Conditions	User must agree to the Terms of Use. No account.		User must agree to the Terms of Use. User must register an account.		User must agree to the Terms of Use. User must register an account. User must be affiliated as a student of a university or high school.		User must agree to the Terms of Use. User must register an account. User must have an academic affiliation.		User must agree to the Terms of Use. User must register an account. Account request must receive approval of institution.	
User status	Public User		Impresso User		Education Impresso User		Academic Impresso User		Impresso User Special Archive X	
Operations	Explore	Get	Explore	Get	Explore	Get	Explore	Get	Explore	Get
PUBLIC DOMAIN DATA										
Metadata	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Facsimiles	yes	no	yes	no	yes	no	yes	no	yes	no
Audio	yes	no	yes	no	yes	no	yes	no	yes	no
Transcripts	yes	no	yes	yes	yes	yes	yes	yes	yes	yes
Images	yes	no	yes	yes	yes	yes	yes	yes	yes	yes
Derived data	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
IMPRESSO PROTECTED DOMAIN DATA										
Metadata	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Facsimiles	no	no	no	no	no	no	no	no	no	no
Audio	no	no	Determined by the Disclosing Party in the Media List.	no	Determined by the Disclosing Party in the Media List.	no	Determined by the Disclosing Party in the Media List.	no	Determined by the Disclosing Party in the Media List.	no
Transcripts	no	no	Determined by the Disclosing Party	no	Determined by the Disclosing Party	no	Determined by the Disclosing Party	no	Determined by the Disclosing Party	no
Images	no	no								
Derived data	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

Several dimensions:

- **Social:** extensive exchanges
- **Legal:** modular data access policy



Implementation

Several dimensions:

- **Social:** extensive exchanges
- **Legal:** modular data access policy
- **Technical:** efficient solution for managing access rights

We use **bitmaps** to represent each dimension of data domain and user status – fast, safe, scalable.

Implementation

Several dimensions:

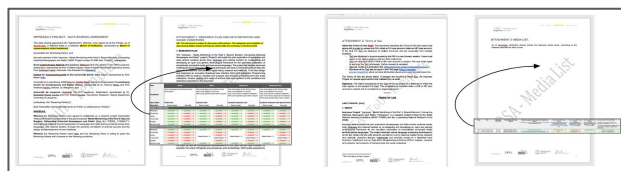
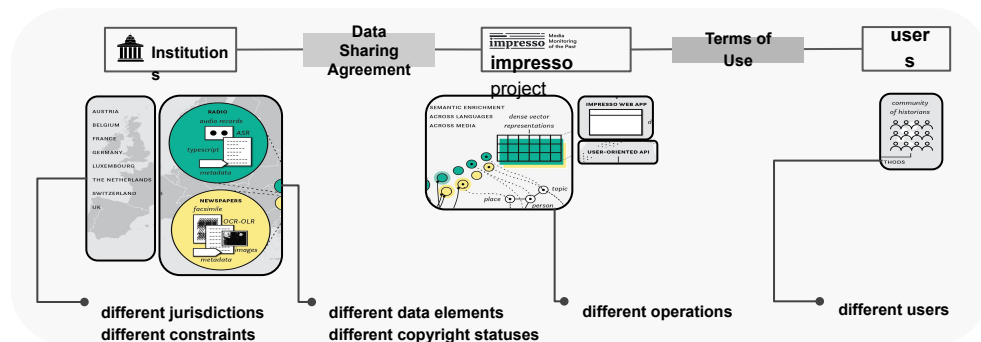
- **Social:** extensive exchanges
- **Legal:** modular data access policy
- **Technical:** efficient solution for managing access rights
- **Communication:** clear information for users on permissions and provenance

Plans for Impresso Databab

	Quest CURRENT PLAN	Impresso User	Coming soon: Student User	Academic User	Coming soon: Academic User
Try the Impresso Web app and the Impresso Databab with access to public domain data.		Adds the ability to work with personal collections in the Impresso Web App and to access our API via the Databab.	Adds access to copyright-protected data available to students in higher education.	Adds access to copyright-protected data available for general research purposes.	Adds ability to request access to data which requires individual permission.
Explore all features of the Impresso Web App and Databab.	X	✓	✓	✓	✓
Create, share and export personal collections.	X	✓	✓	✓	✓
Generate API keys to access parts of our corpus in the Impresso Databab.	X	✓	✓	✓	✓
Requirements					
Agreement to Terms of Use	○	○	○	○	○
Creation of an Impresso Account	—	○	○	○	○
Proof of current student enrolment in higher education	—	—	○	—	—
Proof of academic affiliation	—	—	—	○	—
Data access requires approval by content provider	—	—	—	—	○
Data availability					
Bibliographic Metadata	✓	✓	✓	✓	✓
Bibliographic Metadata public domain	X	✓	✓	✓	✓
Facsimiles - images of documents created during scanning	X	🔒	🔒	🔒	🔒
Facsimiles - images of documents created during scanning public domain	X	👁	👁	👁	👁
Audio - mostly spoken word radio, mostly no	X	🔒	🔒	🔒	🔒

1. **What can I access:** Overview of access rights and available features through **Impresso User Plans**.
2. **What am I allowed to do:** Right statements and permitted uses at title and content item levels.
3. **Source Information:** Indication of content provider for each title and content item.
4. **Citation Guidelines:** Instructions for proper citation.

From obstacles to (viable) solutions?



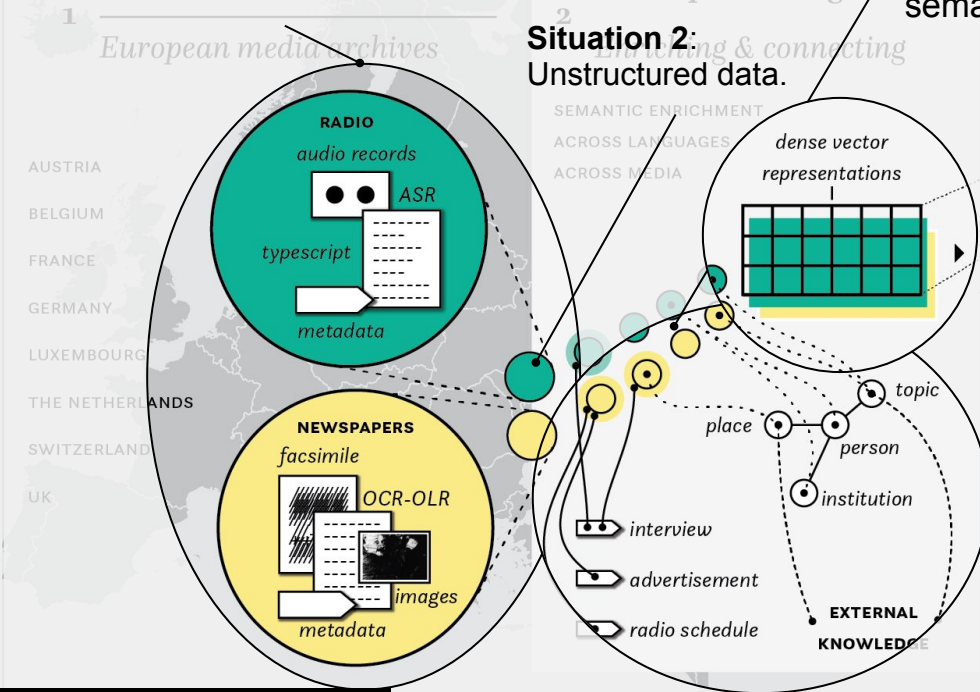
How the Impresso Data	General (Public)	Impresso User	Academic User	Community of historians
General (Public)	Yes	Yes	Yes	Yes
Impresso User	Yes	Yes	Yes	Yes
Academic User	Yes	Yes	Yes	Yes
Community of historians	Yes	Yes	Yes	Yes

Towards a Data Sharing and Access Framework where:

- data is more shareable
- institutions retain control
- users understand their permissions and can use data responsibly
- transnational and transmedia historical research can progress

Enriching and connecting historical media sources.

Situation 1: Very heterogeneous digitised sources in terms of refinement **quality** and **granularity**.



1. Source consolidation

Elevate the corpus to a unified and higher quality level.

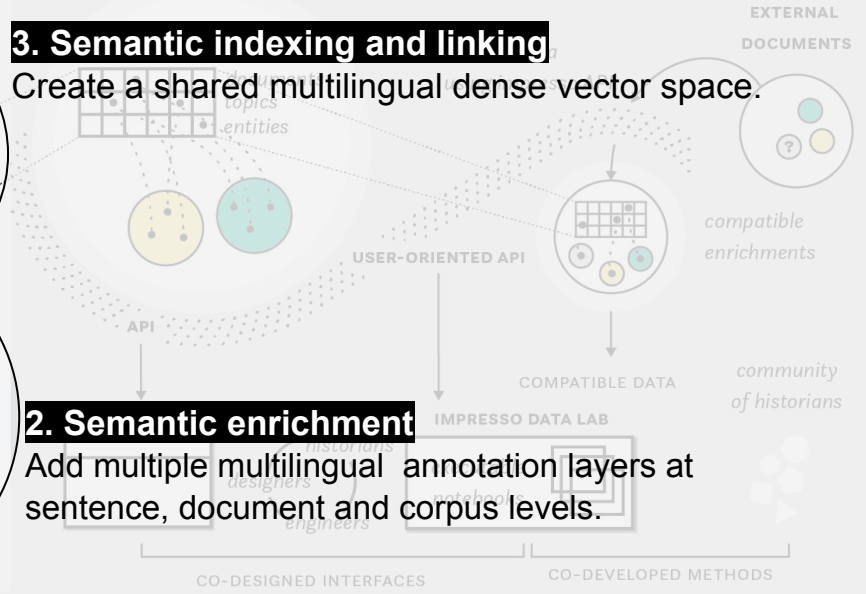
Situation 3: Text structured with multiple layers of semantic information, but no semantic connectivity.

Situation 2: Unstructured data.

3. Semantic indexing and linking.
Create a shared multilingual dense vector space.

2. Semantic enrichment

Add multiple multilingual annotation layers at sentence, document and corpus levels.

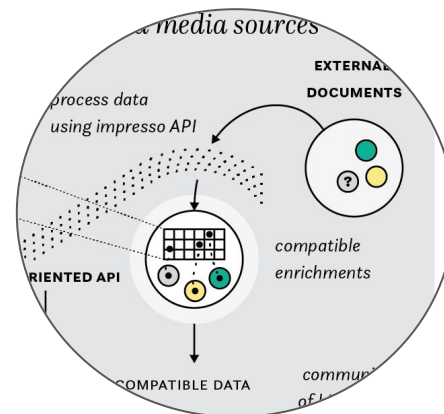
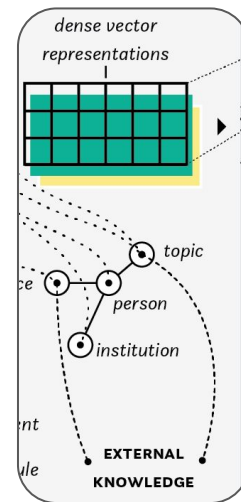


Semantic Connectedness

“**Linking within**” – monolingual and multilingual vector representation for several units.

“**Linking out**” – linking Impresso content to external knowledge bases (entities and articles).

“**Linking in**” – enable user documents to be connected to Impresso content.



Overview of Semantic Enrichments

Operations:

- at the phrase and document levels
- monolingual and cross lingual
- primarily on text, but also on images
- embed all elements

For almost each processing:

- model development
- inference on the whole corpus
- model release on Hugging Face
- annotation service via API

Source consolidation

- experiments with OCR post-correction
- language identification
- basic segmentation (if missing)
- item type classification

Semantic enrichment

- named entity processing
- news agency detection
- (cross-lingual) text reuse detection
- topic modeling
- image type classification
- image captioning

Shared multilingual dense vector space

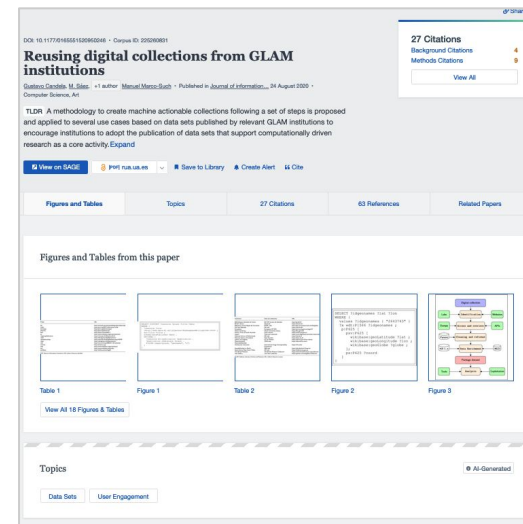
- multilingual and crosslingual word, paragraph and document embeddings
- entity embeddings
- image embeddings

GLAM Perspectives: Almost Already Computational?

*“As technologies have evolved over the years, GLAM organisations **need to adapt to remain relevant** in this new context.*

***New skills** are required regarding digital innovation ranging from service design, data science, digital research and artificial intelligence.*

*Moreover, the research community has highlighted the need for **reproducible research** by providing articles, as well as, data and code”*



<!-- ONB . Labs -->

The ONB Labs are the Austrian National Library's platform for digital collections as data. We offer open digital collections and metadata to promote and inspire research, active experimentation as well as artistic and creative usage.

Among our datasets we provide access to images, texts and metadata that are fully open for you to do whatever you like. We also offer a selection of tools that may support your engagement. The ONB Labs are a dynamically expanding service, which will continuously be revised and extended in correspondence to our users' requirements.

Whether you might need support in accessing the data, whether you have an idea for a useful tool or want to contribute your project to our forum, please contact us. We invite you to share your project or codebase with the Labs community.

In the following topic gallery you can explore projects related to our data sets or tools that may inspire you:

Data lab survey: Overview

Survey

1. MediaSuite
2. NISV Open Data Lab
3. Data Foundry
4. GLAM Workbench
5. Trove API
6. BNF Labs
7. KB Labs
8. LexisNexis
9. HTRC
10. Constellate
11. HASS data lab
12. ProQuest TDM Studio
13. NLN Norway
14. DNB

Interviews

1. MediaSuite
2. Constellate
3. NLN Norway
4. HTRC
5. GLAM Workbench

Survey of CH Data labs, ongoing, with currently 14 data labs, 5 expert interviews, and secondary literature.

Data lab strategies

Shared objective: Enable computational access to institutional collections whilst respecting legal restrictions.

Closed

- Access restricted computing environment for legally protected data
- Large investment in infrastructure, training and access management

Examples: [HTRC](#), [Constellate](#), [Nexis Datalab](#), [ProQuest](#)

Open

- Limited to publicly shareable (meta-) data
- Variable offers of APIs and exploratory tools

Examples: [NLS Data Foundry](#), [NLN DH Lab](#), [BNF Data Lab](#)

Current State of Cultural Heritage Data Labs

Ongoing, with currently 14 data labs, 5 expert interviews, and secondary literature.

Strengths


- Data access
- Education
- Text mining is relevant for a new generation of SSH researchers
- Jupyter notebooks as media
- CPU access is easy

Challenges

- Education
- Legal restrictions
- Pressure to demonstrate impact
- User numbers and community building
- Investment in / Dependencies on large infrastructures
- GPU access is less easy


Opportunities?

- Generative AI for education and analysis
- Models
- Usage success stories




National Library of Scotland
Leabhrairean Nàiseanta na h-Alba

HOMEABOUT • DATA • TOOLSPROJECTSCONTACTQ

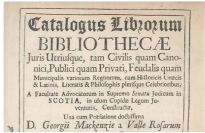


Our data collections are updated on a regular basis.




Explore collections which have been digitised: all of our digitised collections datasets include METS/ALTO files, image files and plain text files.

Digitised collections




Analyse metadata for the Library's collections: available as MARCXML and Dublin Core.

Metadata collections



Discover data extracted from the Library's map collections.


Map and spatial data



Find out more about the daily working of the National Library of Scotland through our organisational data, available as CSV files.

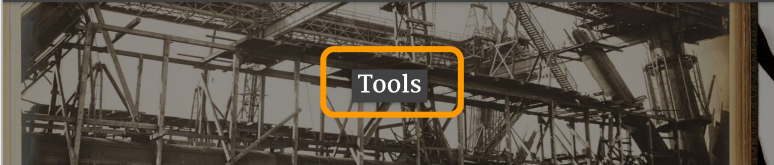
Organisational data

HOMEABOUT • DATA • TOOLSPROJECTSCONTACT




National Library of Scotland
Leabhrairean Nàiseanta na h-Alba

HOMEABOUT • DATA • TOOLSPROJECTSCONTACTQ




Explore the tools available through our Maps website, as well as other options for analysing collections.




Explore our range of Jupyter Notebooks analysing Data Foundry collections.

Jupyter Notebooks




The Library has over 220,000 freely available zoomable maps online on our Map Images website, as well as tools for using georeferenced maps and tools for viewing geospatial data.

Maps and tools



Find tutorials providing introductions to data analysis tools, using Data Foundry datasets.

Tutorials




There are a number of freely-available tools which can be used to analyse library collections data.

External tools

HOMEABOUT • DATA • TOOLSPROJECTSCONTACT

CC-by Marten Düring, Maud Ehrmann, Impresso Team

Ithaka: Constellate


CONSTELLATE

Dataset BuilderClasses & EventsGet ConstellateHelpDashboard

Teach, learn, and perform text analysis with scholarly and primary source content

Constellate helps you get insights out of text data


Everything you need to get started:

- Integrate no-code text and data techniques into your research
- Uncover patterns and trends in 38+ million books and journals
- Join live classes with text and data analysis experts

Help us make this better!

	A	B
1	phrase	documents
2	womens	982
3	constellate	870

<https://constellate.org/>


CONSTELLATE

Dataset BuilderClasses & EventsGet ConstellateHelpDashboard

Dashboard

Basic access

You are not authenticated to an institution. To [check for upgraded access](#), connect to your institution's VPN or proxy and log in.

[Open my lab](#)
[Build a dataset](#)

[Overview](#)
[My datasets](#)
[Tutorials](#) **NEW**
[Lab](#)
[My snapshots](#) **BETA**
[Import from GitHub](#)
[Classes & events](#)

Join Constellate's Skill-build 2024

September 13 - December 6

Advance your text and data analysis skills for research and teaching

Join our 10-week program to dive into Large Language Models (LLMs) and machine learning. Take classes to earn digital certifications, connect via Slack, and more.

[Browse the classes](#)

Resources for getting started

Constellate lab

Use the cloud computing environment Constellate lab to open Constellate notebooks and other Jupyter notebooks, execute workable code, or share your own notebooks with other Constellate users.

[Open my lab](#)

Constellate notebooks are created using Jupyter notebooks, documents that contain executable code and are supported in web-browsers. Learn more about [Jupyter notebooks](#).

Datasets

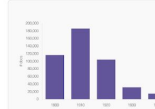
Build a new dataset with the Dataset Builder or view specific datasets to open in Constellate lab.

[Build a dataset](#) [View my datasets](#)

Learn more about building datasets with the [Dataset Builder](#).

Your dataset results

total: 1,912,370 total docs



Constellate empowers educators and students with text analysis and data skills.

[SKILL-BUILD 2024](#) [GET CONSTELLATE](#) [NEWS](#) [LIBGUIDE](#) [CONTACT US](#) [HELP](#)

[Terms and Conditions](#) [Accessibility](#) [Privacy Policy](#) [Use of Cookies](#) [Cookie settings](#)

© 2024 ITHAKA. All Rights Reserved. Constellate® and ITHAKA® are trademarks of ITHAKA.

CC-by Marten Düring, Maud Ehrmann, Impresso Team

Research Perspectives

1. Researchers deal with heterogeneous collections and wish to analyse their data using their own code.
2. Users want support for data discovery and curation.
3. Users require custom and up-to-date tools.
4. Users are interested in data and novel methods rather than computing infrastructure.

User demand for supporting advanced analysis of historical text collections

Max Kemman (Dialogic / [@MaxKemman](#) / 0000-0002-7707-3756),
Steven Claeysens (KB, National Library of the Netherlands / [@sclaeysens](#))

Digital research environments are confronted with a gap between simple search interfaces and advanced functionalities for (text) analysis. Researchers can usually choose between either a user-friendly search interface to read individual sources with few or no options for analysis, or API-access or data export which necessitates coding skills for advanced analyses of (big) datasets (Edmond & Garnett 2015). GLAM-institutions, researchers and even publishers are in search of the silver bullet for user-friendly text analysis, usable for scholars (and others) without coding skills, to take advantage of recent developments in artificial intelligence, natural language processing and other (computational) fields (e.g. Gale Digital Scholar Lab, ProQuest TDM Studio, Constellate and the recent plans for a Trove researcher platform). One development has been the increasing uptake of Jupyter Notebooks as interactive tools for data analysis (Candela et al. 2020), but these too depend on the researcher's capacity to write code.

User studies and literature on comparable platforms for textual collections provide little to no indications on how to bridge the gap. User research of the KB's existing services (user-friendly systems for searching and reading individual sources) show that available advanced functionalities (e.g. n-gram viewers) are used only to a limited extent. Advanced methods for research are also not mentioned by users, or only to a limited extent, as possibilities for improving these services. While other researchers have attempted to bridge the gap, we find no indications that this is achieved successfully. For example, one evaluation of Nederlab (a Dutch platform for diachronic research of text collections, see <https://www.nederlab.nl>) concluded that it could support plenty of research questions, albeit questions that are mainly related to qualitative analyses (Struk 2015). The developers of the Media Suite (a Dutch platform for mixed methods research on media collections, see <https://mediasuite.clariah.nl>) have organized workshops with users, but an extensive user evaluation has not yet been published (or possibly conducted) (Ordeman et al. 2019).

In the present study we explored whether the KB, the national library of the Netherlands, could develop a digital research environment for historical text collections that bridges this gap and offers advanced analysis tools that are sufficiently usable for scholars (and other users) without programming skills, a so-called 'textsuite'. The study employed a user-centric approach to understand how this gap affects research practices (Kemman & Kleppe 2015; Thoden et al. 2017; Warwick 2012), including interviews with employees of the KB (5), developers of comparable research environments (6) and potential users (15).

To determine the building blocks of such a textsuite we started from the scholarly primitives introduced by Unsworth (2000). In later research these primitives have been grouped into distinct research phases to be supported (Blanke & Hedges 2013). We furthermore extended

Historian's wishlist for an Impresso Datalab

1. Datalab (programmatic access) shall complement the Impresso Web App (discovery, exploration)
2. Reflect changing research practices:
 - a. Support question-specific data analysis
 - b. Enable user data enrichment
 - c. Link diverse user data to Impresso
 - d. Encourage research process documentation and publication
3. Support Teaching



2019
IFLA WLIC

Submitted on: 02/09/2019

Historical Newspaper User Interfaces: A Review

Maud Ehrmann

Digital Humanities Laboratory (DHLAB), EPFL, Lausanne, Switzerland.
maud.ehrmann@epfl.ch

Estelle Bunout

Centre for Contemporary and Digital History (C2DH), Luxembourg University,
Luxembourg.
estelle.bunout@uni.lu

Marten Düring

Centre for Contemporary and Digital History (C2DH), Luxembourg University,
Luxembourg.
marten.during@uni.lu



Copyright © 2019 by Maud Ehrmann, Estelle Bunout and Marten Düring. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0>

Abstract:

After decades of large-scale digitization, many historical newspaper collections are just one click away via online portals developed and supported by various public or private stakeholders. Initially offering access to full text search and facsimile visualization only, historic newspaper user interfaces are increasingly integrating advanced exploration features based on the application of text mining tools to digitized sources. As gateways to enriched material, such interfaces are however not neutral and they fundamentally role in how users perceive historical sources, understand potential biases of upstream processes and benefit from the opportunities of digitalization. What features can be found in current interfaces, and to what degree do interfaces adopt novel technologies? This paper presents a survey of interfaces for digitized historical newspapers with the aim of mapping the current state of the art and identifying recent trends with regard to content presentation, enrichment and user interaction. We devised 6 interface assessment criteria and reviewed twenty-four interfaces based on ca. 140 predefined features.

Keywords: digitized historical newspapers, user interfaces, digital scholarship

1. Introduction

Historical newspapers are mirrors of past societies. They reflect the political, moral, and economic environments in which they were produced and they hold dense, continuous, and multi-level information which can help us understand how contemporaries experienced their

Ehrmann, Maud, Estelle Bunout, and Marten Düring. "Historical Newspaper User Interfaces: A Review." Athens, Greece: IFLA, 2019. <http://library.ifla.org/2578/>.

9 preliminary insights

1. **Experimental:** No such thing as a data lab but different models.
2. **Open vs closed systems:** Copyright determines all.
3. **Access & hardware:** Determined by national/corporate research infrastructures.
4. **Data:** Mostly image/metadata/text-focused.
5. **Education:** Apps, tutorials, collaboration.
6. **Pressure:** Small user numbers vs. perceived importance vs. need to demonstrate impact.
7. **Gen. AI:** Desired but not (yet) needed?
8. **Instability:** (Closed) data lab infrastructure and support is volatile.
9. **Sustainability:** Unclear what needs to be sustained, by whom until when.

Media exploration

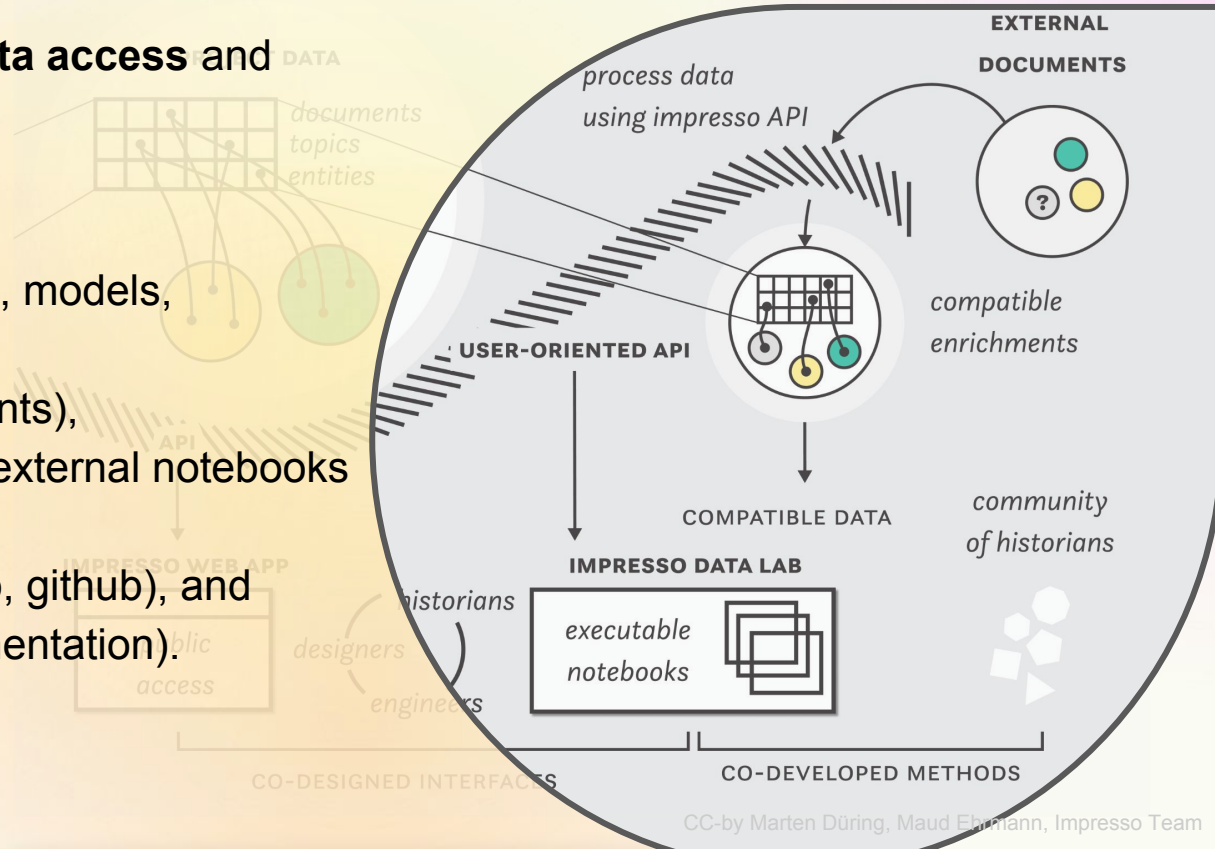
3

Connected and comparable enriched media sources

A platform for **programmatic data access** and **annotation services**.

A space that gives access to

- **Tools** (Impresso Public API, models, annotation services),
- **Material** (corpus, enrichments),
- **Spaces for experiments** (external notebooks notably),
- Basic **infrastructure** (colab, github), and
- **Guidance** (tutorials, documentation).



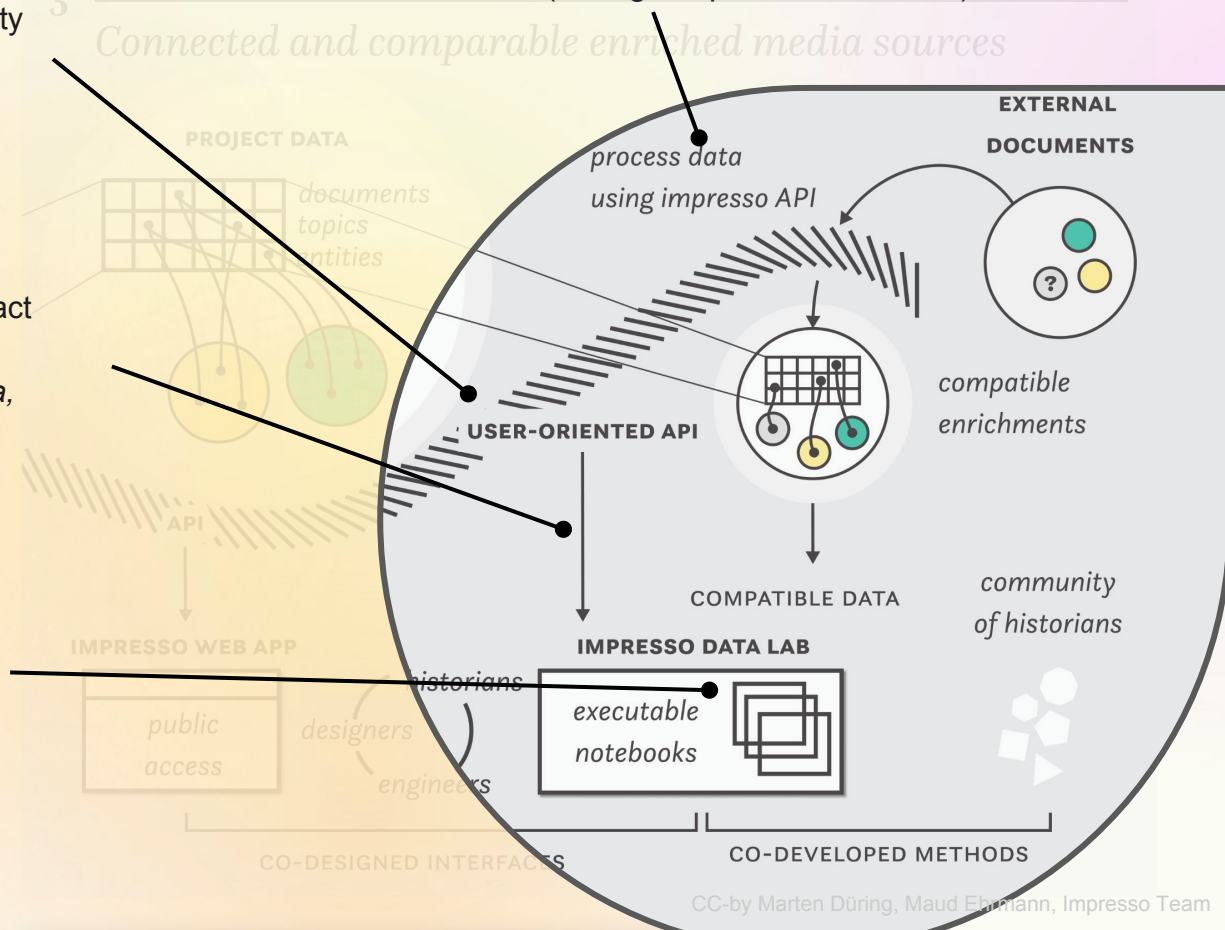
The piece of software that enables third party access to Impresso backend.
Derived from the Impresso Middle Layer.
Exposed to users via tokens.

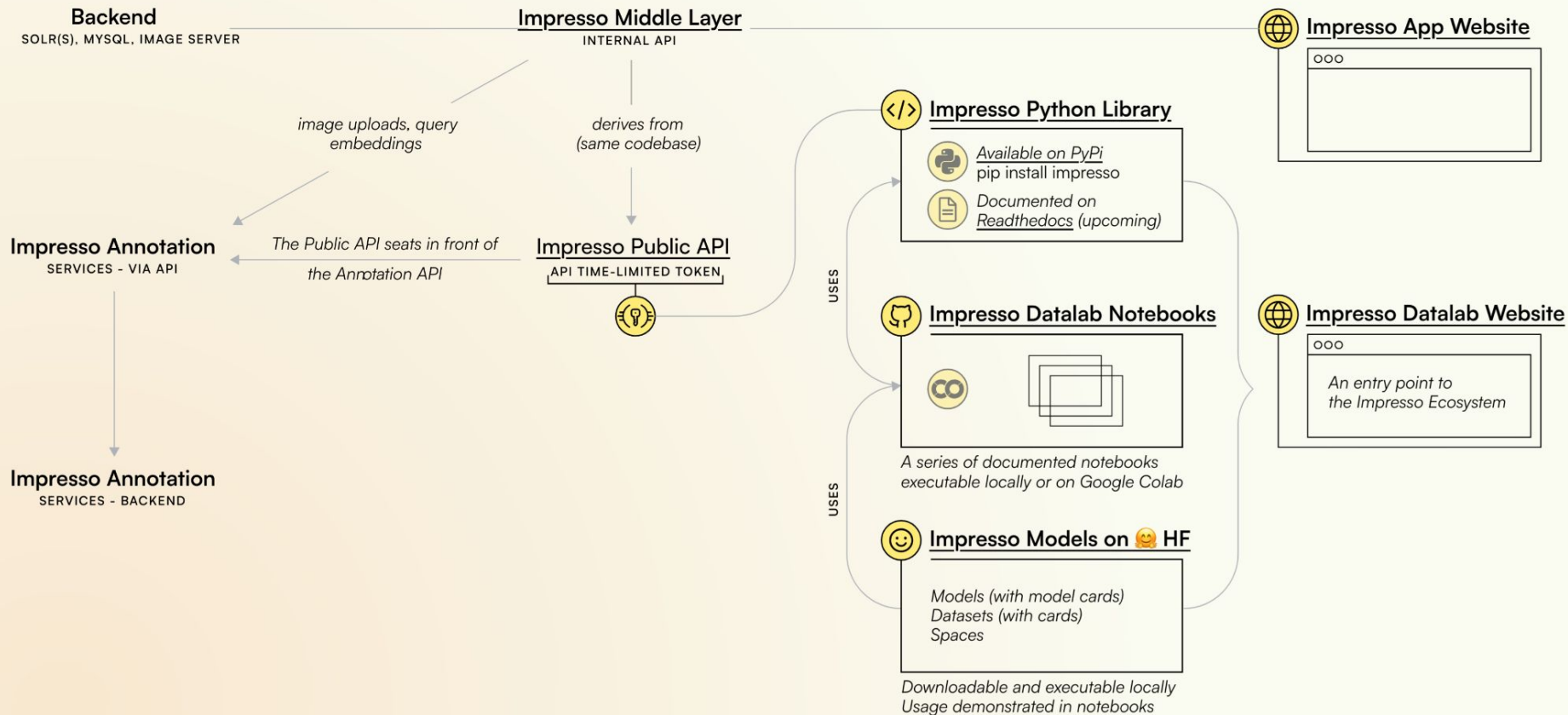
The favorite way proposed to users to interact with the Impresso Public API.

Gives access to Impresso corpus, metadata, enrichments and annotation services

Sharing and storing the notebooks

- on Hugging Face (HF) <https://huggingface.co/impresso-project>
- via Annotation service (through Impresso Public API)





Boost your Media Monitoring

Explore and work programmatically with the Impresso Corpus, Data and Models

Join us in this early stage of development and help us to improve the platform.

The [Impresso project](#) strives to create meaningful links across distinct datasets. The Impresso Datalab is a platform for **programmatic data access** and **annotation services**. It offers access to our data and models via API and a dedicated Python library via Jupyter notebooks. The Datalab enables custom analyses of the Impresso corpus and the semantic indexation of external document collections with the help of models created by the project.

Getting Started

Start your research with Impresso in three easy steps.

Create an Impresso account and learn how to access our API. You can run the notebooks locally or in your preferred environment — whether that's Docker, MyBinder, or Google Colab.

1 REGISTER OR 1 LOGIN THEN

2 ACCEPT OUR TERMS OF USE 3 GET YOUR API KEY


Copy the code below in a blank jupyter notebook to get started


```
# Install the impresso library
%pip install impresso


from impresso import connect


impresso = connect()

results = impresso.search("moon landing")
```

 Open in Colab


 **Interacting with the Impresso Python Library**
by Impresso team →


 Open in Colab


 **Search**
by Impresso team →


Explore and Visualise your Impresso Data

Notebook templates offer complementary views on your Impresso personal collections and external datasets beyond the capabilities of the Impresso Web App.

 Open in Colab

 **Visualising Place Entities on Maps**
by Impresso team →

 Open in Colab

 **Exploring Entity Co-occurrence Networks**
by Impresso team →

Annotate your Documents with Impresso Models


Use Impresso Models for the semantic indexing of your personal document collections, and compare them with Impresso Data.


Check the Impresso models available on HuggingFace and choose the one that fits your needs.


Copy the code below in a blank jupyter notebook to get started


```
# Use a pipeline as a high-level helper
!pip install transformers


from transformers import pipeline
pipe = pipeline("text2text-generation",
model="impresso-project/nel-mgcnre-multilingual")
```


 Open in Colab

 **Named Entity Processing with Impresso Models**
by Impresso team →

 Open in Colab

 **Language Identification using Floret**
by Impresso team →

 Open in Colab

 **News Agencies Recognition and Linking with Impresso BERT models**
by Impresso team →



Generate your API token



You have not accepted our **Terms of Use** yet. Please read the **entire** terms of use document carefully and accept it before using the token.

READ AND ACCEPT THE TERMS OF USE TO GENERATE THE TOKEN

Please login to get your Api Token

LOG IN OR REGISTER

Access tokens programmatically authenticate your identity to the Impresso Datalab, allowing applications to provide you specific data based on your request.

Boost your Media Mo

Explore and work pr
Corpus, Data and M

Getting Started

Start your research with Impresso in t

Create an Impresso account and learn
can run the notebooks locally or in your
whether that's Docker, MyBinder, or Go

1 REGISTER OR 2 LOGIN

2 ACCEPT OUR TERMS OF USE

Copy the code below in a blank jupyter notebook

```
# Install the impresso library
%pip install impresso

from impresso import connect

impresso = connect()

results = impresso.search("moon
```

Open in Colab

Interacting with the Impresso
Library
by Impresso team

Open in Colab

Search
by Impresso team

Plans for Impresso Datalab

	Guest CURRENT PLAN	Impresso User	Coming soon: Student User	Academic User	Coming soon: Academic User+
	Try the Impresso Web app and the Impresso Datalab with access to public domain data.	Adds the ability to work with personal collections in the Impresso Web App and to access our API via the Datalab.	Adds access to copyright-protected data available to students in higher education.	Adds access to copyright-protected data available for general research purposes.	Adds ability to request access to data which requires individual permission.
Explore all features of the Impresso Web App and Datalab.	✗	✓	✓	✓	✓
Create, store and export personal collections.	✗	✓	✓	✓	✓
Generate API keys to access parts of our corpus via the Impresso Datalab.	✗	✓	✓	✓	✓
Requirements					
Agreement to Terms of Use	✓	✓	✓	✓	✓
Creation of an Impresso Account	—	ⓘ	ⓘ	ⓘ	ⓘ
Proof of current student enrollment in higher education	—	—	ⓘ	—	—
Proof of academic affiliation	—	—	—	ⓘ	—
Data access requires approval by content provider	—	—	—	—	ⓘ
Data availability					
Bibliographic Metadata	✓	✓	✓	✓	✓
Bibliographic Metadata public domain	✗	✓	✓	✓	✓
Facsimiles - images of documents created during scanning	✗	🔗	🔗	🔗	🔗
Facsimiles - images of documents created during scanning	✗	👁	👁	👁	👁

ent and help

datasets. The
notation services.
n library via Jupyter
ous and the semantic
reated by the project.

uments with Impresso

semantic indexing of your personal
compare them with Impresso Data.
available on HuggingFace and choose

er notebook to get started

igh-level helper
rs
t pipeline
ext-generation",
/mel-egence-multilingual")

rocessing with Impresso →

ification using Floret →

Recognition and Linking
ERT models →

Boost your Media Monitoring

Explore and work programmatically with the Impresso Corpus, Data and Models

Getting Started

Start your research with Impresso in three easy steps.

Create an Impresso account and learn how to access our API. You can run the notebooks locally or in your preferred environment — whether that's Docker, MyBinder, or Google Colab.

1 REGISTER

OR

2 LOGIN

THEN

3 ACCEPT OUR TERMS OF USE

4 GET YOUR API KEY

Copy the code below in a blank jupyter notebook to get started

```
# Install the impresso library
%pip install impresso

from impresso import connect

impresso = connect()

results = impresso.search("moon landing")
```

Open in Colab



Interacting with the Impresso Python Library

by Impresso team



Join us in this early stage of development and help us to improve the platform.

The [Impresso project](#) strives to create meaningful links across distinct datasets. The Impresso Datalab is a platform for **programmatic data access** and **annotation services**. It offers access to our data and models via API and a dedicated Python library via Jupyter notebooks. The Datalab enables custom analyses of the Impresso corpus and the semantic ~~indexation of external document~~ collections with the help of models created by the project.

Login



Log in to your account

Email address

daniele.guido@uni.lu

Password

.....

LOG IN

Did you forget your password? [Reset your password](#)

Don't have an account?

REGISTER

Any Questions?

Contact us at info@impresso-project.ch

Annotate your Documents with Impresso Models

Use Impresso Models for the semantic indexing of your personal document collections, and compare them with Impresso Data.

Check the Impresso models available on HuggingFace and choose the one that fits your needs.

Copy the code below in a blank jupyter notebook to get started

```
# Use a pipeline as a high-level helper
!pip install transformers

from transformers import pipeline
pipe = pipeline("text2text-generation",
model="impresso-project/nel-mgenre-multilingual")
```

Open in Colab



Named Entity Processing with Impresso Models

by Impresso team



Open in Colab



Language Identification using Floret

by Impresso team

CC-by Marten Düring, Maud Ehrmann, Impresso Team





Your Api token



API access is always subject to the [TERMS OF USE](#). You accepted the Terms of Use **28 October 2024 at 15:06 CET**

eyJhbGciOiJIUzI1NiIsInR5cCI6ImlmFjY2VzcyJ9.eyJ1c2VySWQiOiJsb2NhbmC1kZylslnVzZXJHcm91cHMlOlsicGxhbi1lZHVjYXRpb25hbCIsIk5vUmVkYWNOaW9uIiwiaWF0IjU3RmZmYiOnRydWUslmlhdCI6MTczMDEyNDQwMywiZXhwIjoxNzMTUzMjAzMz0



Please copy your token and keep it in a safe place. This token will be valid for the next 7h 59m 49s

Access tokens programmatically authenticate your identity to the Impresso Datalab, allowing applications to provide you specific data based on your request.

Getting Started

Start your research with Impresso in three easy steps.

Create an Impresso account and learn how to access our API. You can run the notebooks locally or in your preferred environment — whether that's Docker, MyBinder, or Google Colab.

1 ACCEPT OUR TERMS OF USE

2 GET YOUR API KEY

Copy the code below in a blank jupyter notebook to get started

```
# Install the impresso library
!pip install impresso

from impresso import connect

impresso = connect()

results = impresso.search("moon landing")
```

 Open in Colab



Interacting with the Impresso Python Library

by Impresso team



 Open in Colab



Search

by Impresso team



 Open in Colab



Search collections

by Impresso team



Boost your Media Mo

Explore and work pro
Corpus, Data and M

Getting Started

Start your research with Impresso in t

Create an Impresso account and learn I
can run the notebooks locally or in your
whether that's Docker, MyBinder, or Go

1 REGISTER OR 1 LOGIN

2 ACCEPT OUR TERMS OF USE

Copy the code below in a blank jupyter notebook t

```
# Install the impresso library
!pip install impresso

from impresso import connect

impresso = connect()

results = impresso.search("moon
```

Notebook



Interacting with the Impresso Python Library

By Impresso team

🔗 Open in Colab

🔗 Open in GitHub

Last update: 2024 Oct 25

Note: This is a static preview of the Jupyter notebook.

What is this notebook about?

This notebook provides a quick introduction to the Impresso Python library, a module designed to interact with the Impresso Public API. It is an ideal starting point for users looking to explore the Impresso corpus, its metadata, and the associated semantic enrichments. To access the data, you will need an Impresso account. If you do not have one yet, you can register on the [Impresso Datalab Website](#).

What will you learn in this notebook?

In this notebook, you will learn how to:

- Instantiate the client and authenticate with the Impresso API
- Search for content items within the Impresso corpus
- Use Jupyter's auto-assist features for exploring documentation
- Understand the main areas (namespaces) covered by the library
- Retrieve semantic enrichments such as entities and text reuse
- Work with collections
- Work with facets

This notebook will guide you through these core functionalities and help you get familiar with the Impresso library capabilities.

Abstract

This notebook provides a practical introduction to the Impresso Python library for interacting with the Impresso Public API. You will learn to authenticate and search the Impresso corpus for content items, leverage Jupyter's auto-assist features for exploring documentation, retrieve semantic enrichments, and work with collections and facets.

See also

🔗 Open in Colab



Search

by Impresso team



🔗 Open in Colab



Search collections

by Impresso team



Links

[Impresso Datalab Website](#)
[library code base](#)

ent and help

datasets. The
notation services.
n library via Jupyter
pus and the semantic
reated by the project.

uments with Impresso

a semantic indexing of your personal
compare them with Impresso Data.
available on HuggingFace and choose

er notebook to get started

igh-level helper
rs

t pipeline

trained("impresso-project/nel-

```
("generic-nel",  
tokenizer=nel_tokenizer,  
e, device='cpu')
```



+ Code + Text Copy to Drive

Connect

Gemini



- Work with facets

This notebook will guide you through these core functionalities and help you get familiar with the Impresso library capabilities.

▼ Prerequisites

Install the `impresso` python library:

```
[ ] %pip install -q impresso
```

↻ Note: you may need to restart the kernel to use updated packages.

▼ Initialising an Impresso Client

In this cell, we create an instance of the Impresso client and authenticate it with the Impresso API.

The `impresso_session` variable stores an instance of `ImpressoClient`, which establishes a connection to the API using your authentication token. With this object, you can interact with the API to perform operations such as searching for content items, retrieving entities, and fetching facets.

The following command will prompt you to enter your Impresso token if it has not been authenticated recently (it expires after 8 hours).

```
[ ] from impresso import connect
```





+ Code + Text Copy to Drive

Connect

Gemini



▼ Making a first request

Let's start by making a simple request to the Impresso API.

We will search for content items that contain the word **"Titanic"** and order the results by date in ascending order.

In Impresso, a **Content Item** is the smallest unit of editorial content within a newspaper or radio collection. This can be an article (for newspapers) or a radio show or episode (for radio programs). Content items can also vary by type, including articles, advertisements, tables, images, and more. Please note that when a newspaper does not have segmentation (OLR) content items for this title correspond to pages.

```
[ ] search_results = impresso_session.search.find(  
    q="Titanic",  
    order_by="date",  
)  
search_results
```



Search result

Contains **100** items (0 - 100) of **8371** total items.

See this result in the [Impresso App](#).

Data preview:

type	title	size	nbPages	pages	isCC	excerpt	labels	accessRight	year	locations	persons	language
------	-------	------	---------	-------	------	---------	--------	-------------	------	-----------	---------	----------





+ Code + Text Copy to Drive

Connect

Gemini



[] Search result

Contains 100 items (0 - 100) of 8371 total items.

See this result in the [Impresso App](#).

Data preview:

uid	type	title	size	nbPages	pages	isCC	excerpt	labels	accessRight	year	locations	persons	language
volkfreu1869-1872-06-02-a-i0025	ad	Publicité 10 Page 3	151	1	'volkfreu1869-1872-06-02-a-p0003', 'n...	True	93cbeutenbe Immobil- Versteigerung jii <*cbronu...	[article]	OpenPublic	1872	[]	[]	de
NZZ-1876-07-29-a-i0003	page	sind , ft ist von	3982	1	{{'uid': 'NZZ-1876-07-29-a-p0003', 'num': 3, '...	True	sind , ft ist von inen kriegerischen Vorzügen ...	[article]	Closed	1876	{{'uid': 'aida-0001-54- Auch', 'relevance': 1},...	[]	de
indeplux-1908-09-04-a-i0026	ar	Toujours plus grand	73	1	'indeplux-1908-09-04-a-p0003', 'num':...	True	Toujours plus grand La construction des deux* ...	[article]	na	1908	[]	[]	fr

The result of the search request to the API is displayed as a notebook-friendly preview when running in a Jupyter notebook.

basics_ImpressoAPI.ipynb

Fichier Modifier Affichage Insérer Exécution Outils Aide

+ Code + Texte Copier sur Drive

↑ ↓ ↺ ↻ ↵ ↶ ↷ ⋮

Connecter Gemini

Interacting with the Impresso Python Library

What is this notebook about?

This notebook provides a quick introduction to the Impresso Python library, a module designed to interact with the Impresso Public API. It is an ideal starting point for users looking to explore the Impresso corpus, its metadata, and the associated semantic enrichments. To access the data, you will need an Impresso account. If you do not have one yet, you can register on the [Impresso Datalab Website](#).

What will you learn in this notebook?

In this notebook, you will learn how to:

- Instantiate the client and authenticate with the Impresso API
- Search for content items within the Impresso corpus
- Use Jupyter's auto-assist features for exploring documentation
- Understand the main areas (namespaces) covered by the library
- Retrieve semantic enrichments such as entities and text reuse
- Work with collections
- Work with facets

This notebook will guide you through these core functionalities and help you get familiar with the Impresso library capabilities.

Prerequisites

Install the `impresso` python library:

```
[ ] %pip install -q impresso
```

Note: you may need to restart the kernel to use updated packages.

Initialising an Impresso Client

In this cell, we create an instance of the Impresso client and authenticate it with the Impresso API.

The `impresso_session` variable stores an instance of `ImpressoClient`, which establishes a connection to the API using your authentication token. With this object, you can interact with the API to perform operations such as searching for content items, retrieving entities, and fetching facets.

The following command will prompt you to enter your Impresso token if it has not been authenticated recently (it expires after 8 hours).

```
[ ] from impresso import connect
```

Gemini

Gemini

Gemini est un outil d'IA performant développé par Google qui vous aide à utiliser Colab.

Vous ne savez pas quoi demander ?

Essayez l'une des requêtes suggérées ci-dessous

How do I filter a Pandas DataFrame?

How can I create a plot in Colab?

Show me a list of publicly available datasets

Saisissez une requête ici


0/60

Les réponses peuvent contenir des informations inexactes ou choquantes qui ne représentent pas le point de vue de Google. [En savoir plus](#)

CC-by Marten Düring, Maud Ehrmann, Impresso Team

Explore and Visualise your Impresso Data

Notebook templates offer complementary views on your Impresso personal collections and external datasets beyond the capabilities of the Impresso Web App.


 Open in Colab



Visualising Place Entities on Maps

by Impresso team



 Open in Colab



Exploring Entity Co-occurrence Networks

by Impresso team



Notebook



Visualising Place Entities on Maps

By Impresso team

[Report an issue](#)

Open in Colab

Open in GitHub

Last update: 2024 Oct 25

Note: This is a static preview of the Jupyter notebook.

Install dependencies

We need the following packages:

- [impresso-py](#)
- [ipyleaflet](#)

```
%pip install -q impresso ipyleaflet
```



Connect to the Impresso API

```
from impresso import connect, OR, DateRange
```



Abstract

This notebook provides a way to analyze and explore the geographic distribution of entities mentioned in Impresso using the Impresso Python Library.

See also

Open in Colab

**Search**

by Impresso team



Links



+ Code + Text Copy to Drive

RAM
Disk

Gemini



Connect to the Impresso API



```
from impresso import connect, OR, DateRange  
impresso = connect()
```

...

Click on the following link to access the login page: <https://impresso-project.ch/datalab/token>

- Enter your email/password on this page.
- Once logged in, a secret token will be generated for you.
- Copy this token and paste it into the input field below. Then press "Enter".

Enter your token:

Search and collect entities

Find top 100 location entities mentioned in articles that talk about nuclear power plants in the first three decades following the second world war in English, French and German.



```
locations = impresso.search.facet(  
    "location",  
    q=OR("centrale nucléaire", "nuclear power plant", "Kernkraftwerk"),  
    date_range=DateRange("1945-05-01", "1975-05-01"),  
    limit=100.
```

✓ 0s completed at 11:01 AM





+ Code + Text Copy to Drive

✓ RAM
Disk✓ [4] See this result in the [Impresso App](#).

Data preview:

	name	type	countItems	countMentions	wikidataId	wikidata.coordinates.latitude	wikidata.coordinates.longitude	wikidata.coordinates.altitude	wikidata.coordinates.p
uid									
aida-0001-54-Lausanne	Lausanne	location	2918317	4642691	Q807	46.533333	6.633333	NaN	
aida-0001-54-Suisse\$2c\$ Moselle	Suisse, Moselle	location	2561532	4268837	Q22036	48.965833	6.579444	NaN	
aida-0001-54-Switzerland	Switzerland	location	2390778	4727170	Q39	46.798562	8.231973	NaN	

Filter out entities that have no coordinates and add a country tag.



```
import pandas as pd
# disable "copy-on-write" warning
pd.set_option('mode.chained_assignment', None)
on_write = True

df = df[df['wikidata.coordinates.latitude'].notna() & df['wikidata.coordinates.longitude'].notna()]

# entity-type == "item" indicates it's a country
entities_with_coordinates['is_country'] = entities_with_coordinates['wikidata.descriptions.fr'].str.contains('pays')
entities_with_coordinates
```



	name	type	countItems	countMentions	wikidataId	wikidata.coordinates.latitude	wikidata.coordinates.longitude	wikidata.coordinates.altitude	wikidata.coordinates.p
uid									
aida-0001-54-Lausanne	Lausanne	location	2918317	4642691	Q807	46.533333	6.633333	NaN	

✓ 0s completed at 11:46 AM



+ Code + Text Copy to Drive

RAM
Disk

+ Gemini



0s

```
import pandas as pd
# disable "copy-on-write" warning
pd.options.mode.copy_on_write = True

df = entities.df
entities_with_coordinates = df[df['wikidata.coordinates.latitude'].notna() & df['wikidata.coordinates.longitude'].notna()]

# entity-type == "item" indicates it's a country
entities_with_coordinates['is_country'] = entities_with_coordinates['wikidata.descriptions.fr'].str.contains('pays')
entities_with_coordinates
```



	name	type	countItems	countMentions	wikidataId	wikidata.coordinates.latitude	wikidata.coordinates.longitude	wikidata.coordinates.altitude	wikidata.coordinates.precision
uid									
aida-0001-54-Lausanne	Lausanne	location	2918317	4642691	Q807	46.533333	6.633333	NaN	0.01666
aida-0001-54-Suisse\$2c\$_Moselle	Suisse, Moselle	location	2561532	4268837	Q22036	48.965833	6.579444	NaN	0.00027
aida-0001-54-Switzerland	Switzerland	location	2390778	4727170	Q39	46.798562	8.231973	NaN	0.00000
aida-0001-54-Fribourg	Fribourg	location	2286590	4264647	Q36378	46.800000	7.150000	NaN	0.01666
aida-0001-54-Paris	Paris	location	2132568	3551470	Q90	48.856944	2.351389	NaN	0.00027
...
aida-0001-54-Mannheim	Mannheim	location	31397	40884	Q2119	49.487778	8.466111	NaN	0.00027
aida-0001-54-Baden-Württemberg	Baden-Württemberg	location	10124	13277	Q985	48.537778	9.041111	NaN	0.01666
aida-0001-54-Haut-Rhin	Haut-Rhin	location	6388	7590	Q12722	47.964167	7.319722	NaN	0.00001
aida-0001-54-Würenlingen	Würenlingen	location	2108	2866	Q64219	47.532200	8.256600	NaN	0.00010
aida-0001-54-Gösgen	Gösgen	location	1680	2268	Q660753	47.373611	7.993056	NaN	0.00027

0s completed at 11:46 AM





+ Code + Text Copy to Drive

✓ RAM
Disk

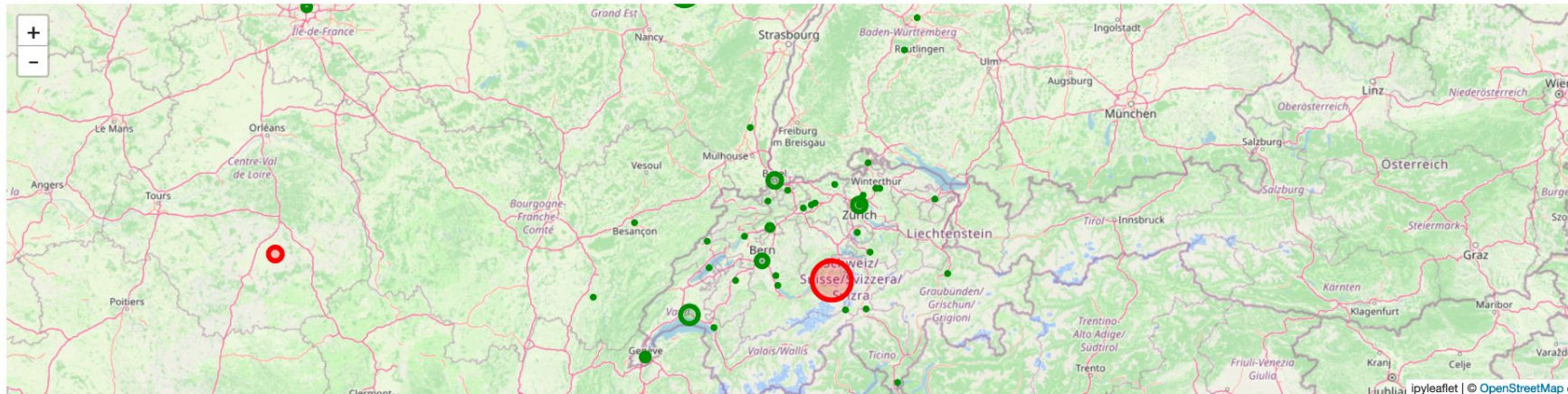
```
marker.popup = build_hover_popup(label, description, row['mentions_count'])

coordinates.append((lat, lon))
markers.append(marker)

# Fit the map to the bounds
map.fit_bounds(find_bounds(coordinates))

# add markers
for m in markers:
    map += m

display(map)
```



Heatman

✓ 0s completed at 11:48 AM



place-entities_map.ipynb

File Edit View Insert Runtime Tools Help Save in GitHub to keep changes



Share

+ Code + Text Copy to Drive

RAM Disk Gemini



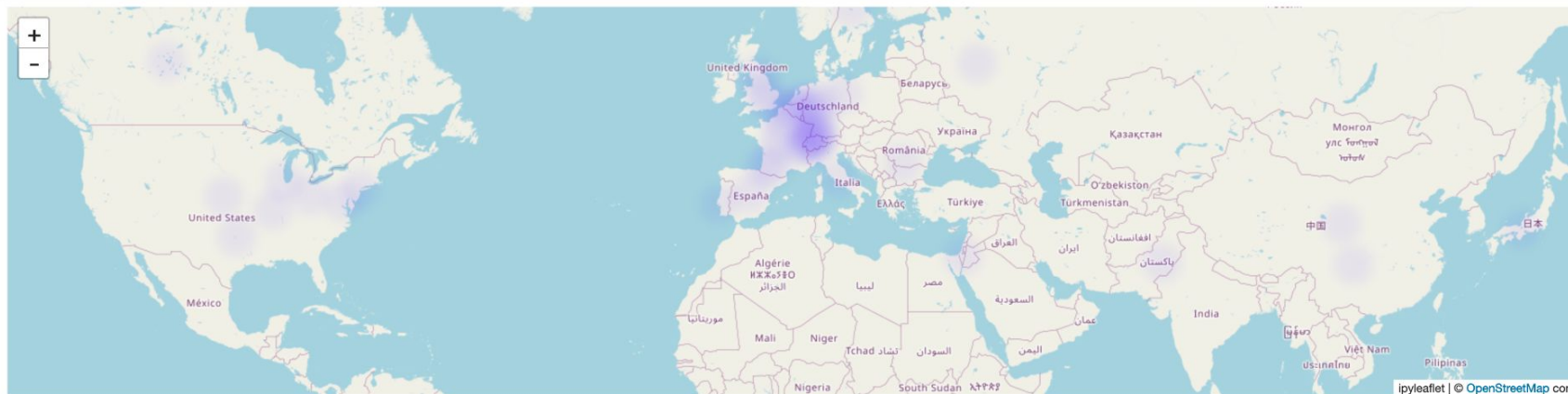
```
map = Map(zoom=0)

locations = []
for index, row in entities_with_coordinates.iterrows():
    lat = row['wikidata.coordinates.latitude']
    lon = row['wikidata.coordinates.longitude']
    # add every coordinate 30 times to make the heatmap more visible
    locations.extend([(lat, lon) for i in range(30)])

heatmap = Heatmap(
    locations=locations,
    radius=20,
    blur=10,
)

map.add(heatmap)

map
```



ipyleaflet | © OpenStreetMap contributors

✓ 0s completed at 11:48 AM

Annotate your Documents with Impresso Models

Use Impresso Models for the semantic indexing of your personal document collections, and compare them with Impresso Data.

Check the Impresso models available on HuggingFace and choose the one that fits your needs.

Copy the code below in a blank jupyter notebook to get started

```
# Use a pipeline as a high-level helper
!pip install transformers

from transformers import pipeline
pipe = pipeline("text2text-generation", model="impresso-project/nel-mgenre-multilingual")
```

 Open in Colab



Named Entity Processing with Impresso Models

by Impresso team




 Open in Colab



Language Identification using Floret

by Impresso team



 Open in Colab



News Agencies Recognition and Linking with Impresso BERT models

by Impresso team



Boost your Media Mo

Explore and work pr
Corpus, Data and M

Getting Started

Start your research with Impresso in t

Create an Impresso account and learn
can run the notebooks locally or in your
whether that's Docker, MyBinder, or Go

1 REGISTER

OR

2 LOGIN

2 ACCEPT OUR TERMS OF USE

Copy the code below in a blank jupyter notebook t

```
# Install the impresso library
%pip install impresso

from impresso import connect

impresso = connect()

results = impresso.search("moon
```

Notebook



Searching Relevant texts within an Embedding space

By Impresso team

Open in Colab

Open in GitHub

Last update: 2024 Oct 29

Note: This is a static preview of the Jupyter notebook.

This notebook demonstrates how to use a pre-trained multilingual embedding model downloaded from Hugging Face to search for relevant texts across languages.

We'll load the model, embed the texts and demonstrate use cases on how to find relevant texts across the German/French language pair within Impresso. Through these use cases we will get familiar with how to use the utilities functions we provide so that you can extend this to your user case of interest.

Recommened Hardware: GPU support, the colab free one (T4) is sufficient. Alternatively, calculations with CPU are possible but much slower.

1. Install Dependencies

First, we need to install `sentence-transformers`

```
!pip install sentence-transformers
```



2. Model Information

In this example, we are using an off the shelf multilingual embedding model

[view in Google Colab](#)

ent and help

datasets. The
notation services.
n library via Jupyter
pus and the semantic
reated by the project.

Documents with Impresso

the semantic indexing of your personal
compare them with Impresso Data.
available on HuggingFace and choose

er notebook to get started

```
High-level helper  
rs  
  
t pipeline  
  
rained("impresso-project/nel-  
  
("generic-nel",  
okenizer=nel_tokenizer,  
e, device='cpu'
```


Insights from Impresso Workshop “Beyond Borders”

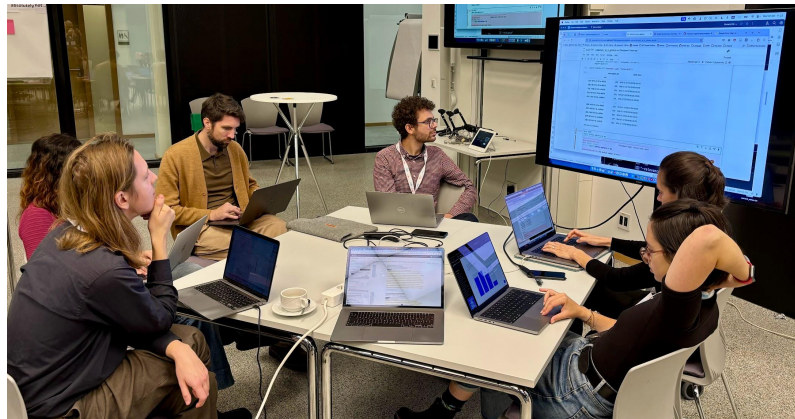
Setup

29. - 30.10.2024 in Luxembourg with
17 DH scholars, developers and librarians
+ Impresso team.



Main goals

- Preview Impresso Datalab for peers
- Develop prototype notebooks
- Kickstart user community

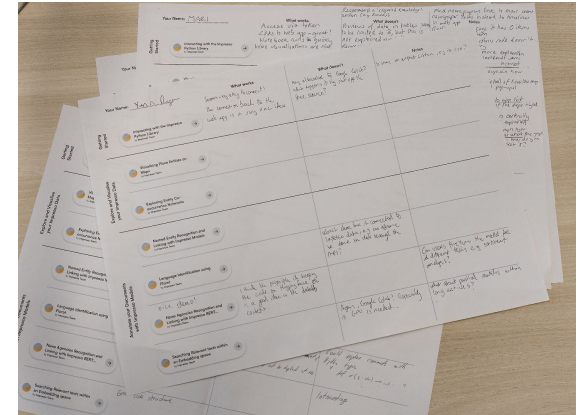
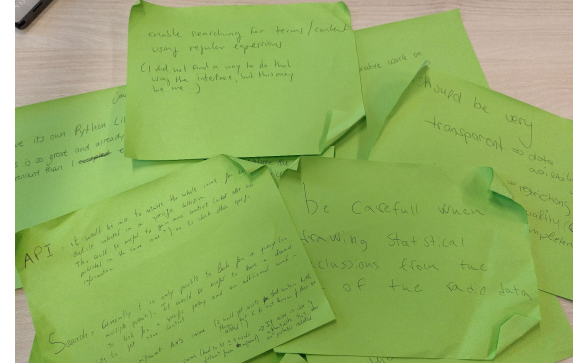


Thematic tracks

1. **Data analysis** - How to explore and analyse Impresso data accessible via API
2. **Cross-lingual alignment** - How to create meaningful links across languages and modalities
3. **Impresso data enrichment using external models** - How to make a relevant subset of Impresso data more useful
4. **External data enrichment using Impresso's models** - How to make other data comparable by applying Impresso's models for semantic enrichment? How to create meaningful links between external data and Impresso?
5. **Generative AI** - How to integrate it for notebook and query generation
6. **Teaching** - How to integrate data labs in academic teaching
7. **Representativity and bias** - How to manage them in historical media collections
8. **Comparison to other Datalabs** - Which ideas could we integrate and what can we learn from others?

Recommendations and main feedback

1. Have a clearly defined focus for the Datalab.
2. Define purpose of individual notebooks well.
3. Break new grounds by developing novel workflows for case studies.
4. Embrace transparency and source criticism.
5. Support different types of users.
6. Support for users with varying skills



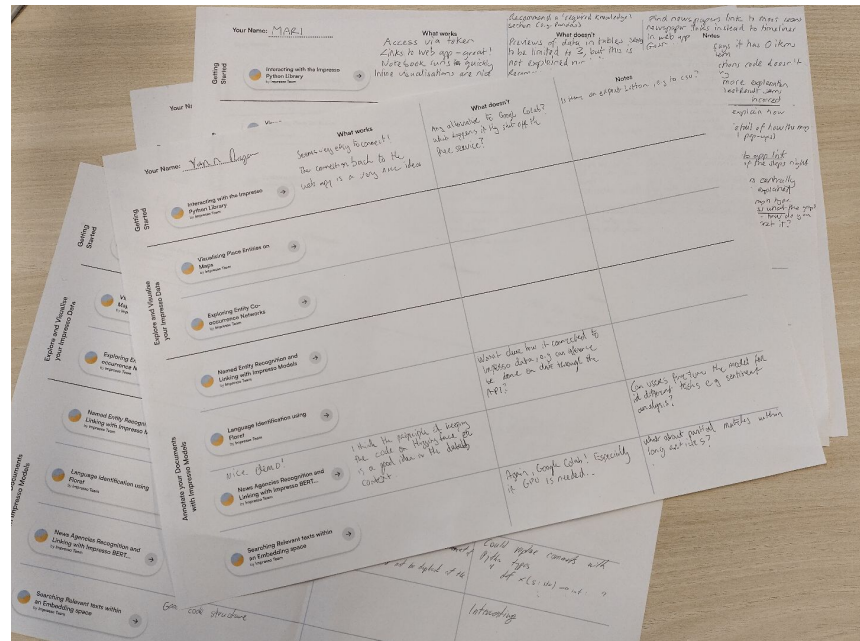
Feedback for preview notebooks

Appreciated

- Link Web App - Datalab
- Impresso Python library
- HuggingFace integration

Invest in

- Support for users with varying skills
- Clear focus for notebooks
- Clarity: Impresso API vs Model usage



Backend
SOLR(S), MYSQL, IMAGE SERVER

Impresso Mi
INTERNAL

image uploads, query
embeddings

derive
(same c

Impresso Annotation
SERVICES - VIA API

The Public API seats in front of
the Annotation API

Impresso
API TIME-LIM

Impresso Annotation
SERVICES - BACKEND



Impresso Python Library



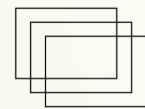
Available on PyPi
pip install impresso



Documented on
Readthedocs (upcoming)



Impresso Datalab Notebooks



A series of documented notebooks
executable locally or on Google Colab



Impresso Models on 🤗 HF

Models (with model cards)
Datasets (with cards)
Spaces

Downloadable and executable locally
Usage demonstrated in notebooks



Impresso App Website

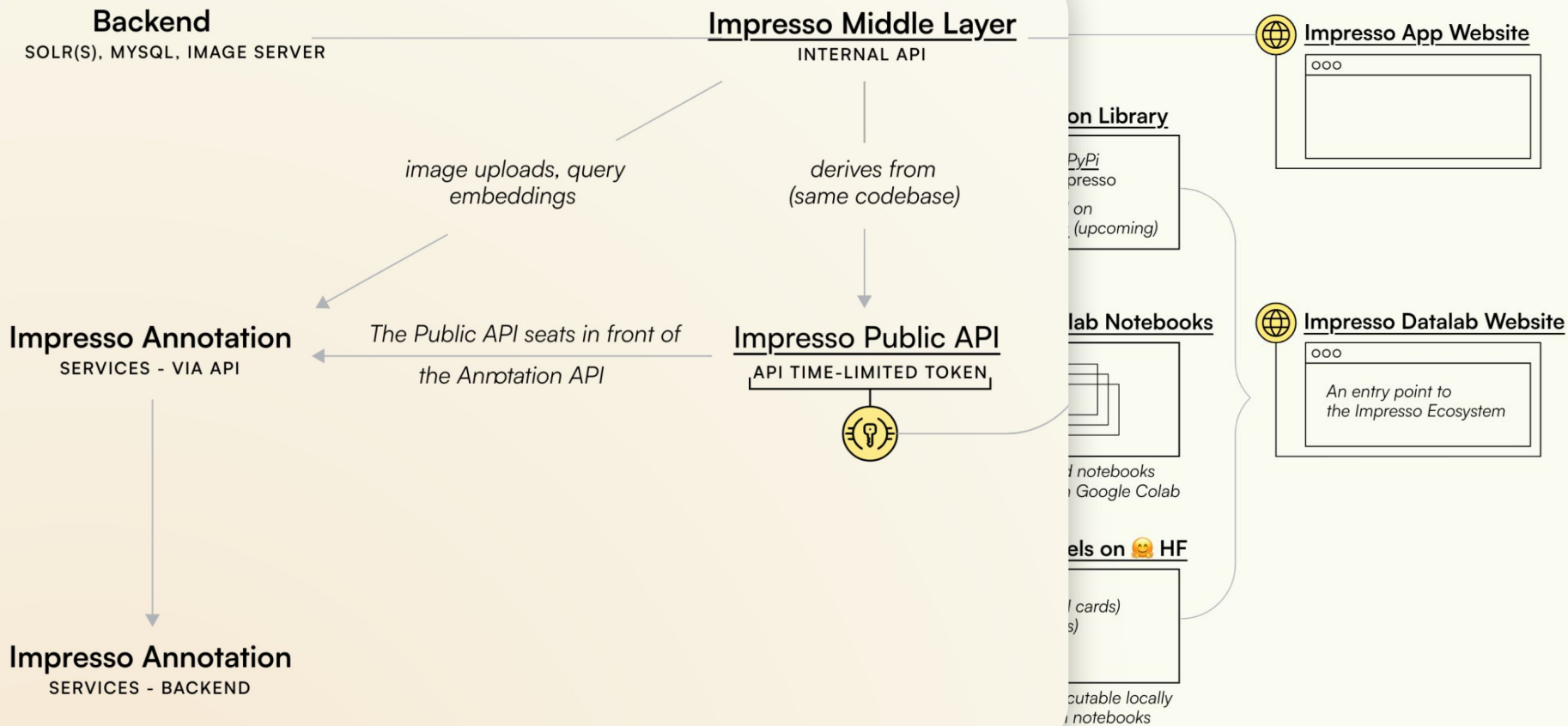
ooo

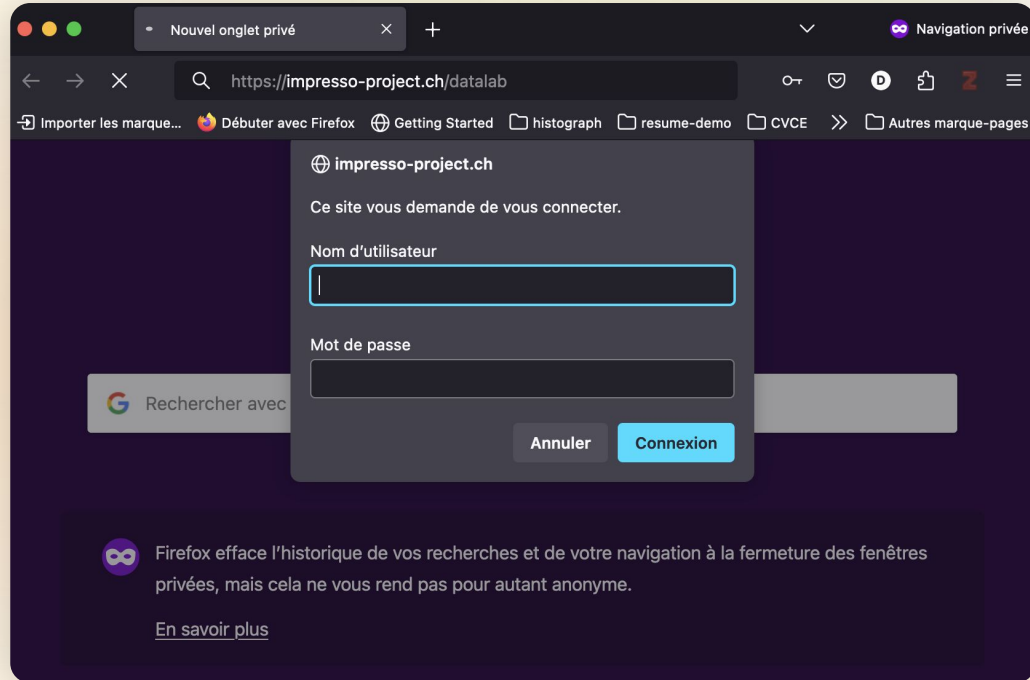


Impresso Datalab Website

ooo

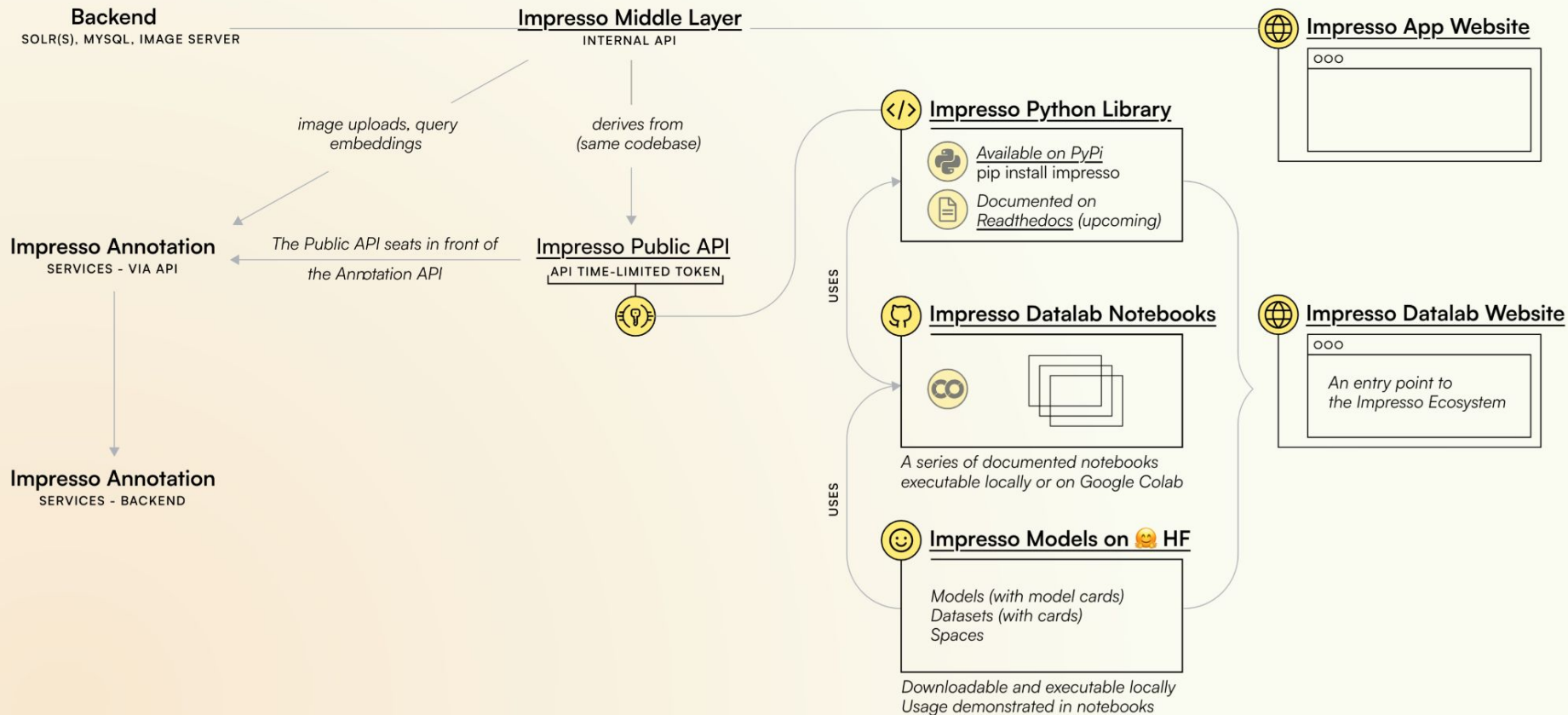
An entry point to
the Impresso Ecosystem





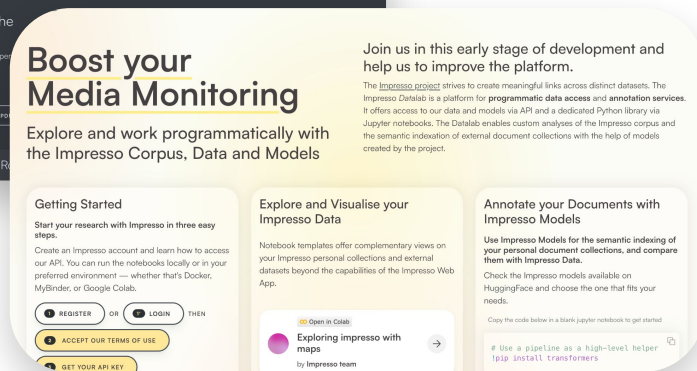
<https://impresso-project.ch/datalab>

impresso / espresso :D



Wrapping up

From critical content mining of newspapers...

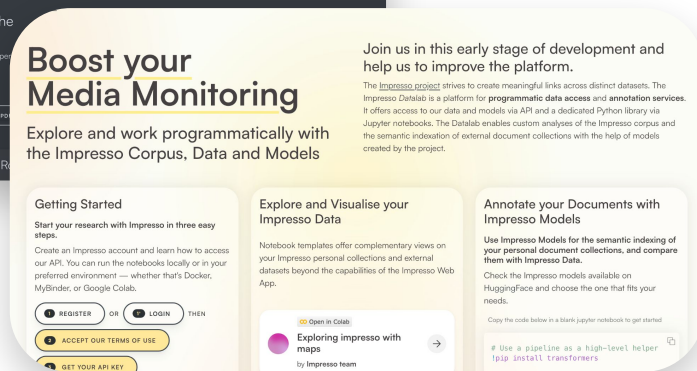


... to the joint exploration of newspaper and radio **beyond borders** to support **comparative** and **data-driven historical research** using **semantically enriched sources** accessible through both **graphical** and **API-based interfaces**.

Wrapping up

Highlights

- **Legal and technical framework** to allow access to **public** and **copyrighted** material, and perform **various operations**.
- **Semantic enrichments** and **connectedness** of data “all round” (linking within/in/out), including across languages.
- **Datalab**, to enable **versatile research workflows** and data exploration: **close integration** of Impresso Web App, Datalab, Models, Services, and documentation.



Outlook

- **First release Impresso Datalab (very soon)**
- **Conference Transmedia History.**
Circulations, Reconfigurations and New Methodologies (27-28 January, Lausanne)
- **Datalab notebook development** for different types of users
- **First release radio data + newspaper corpus expansion (early 2025)**
- **Impresso Datalab Workshop** at DhD2025 (3-7 March 2025, Bielefeld)
- **Workshop: Sustainability, Standards & Infrastructure** (internal, June 2025, CH)



Boost your Media Monitoring

Explore and work programmatically with the Impresso Corpus, Data and Models

Join us in this early stage of development and help us to improve the platform.

The *Impresso project* strives to create meaningful links across distinct datasets. The Impresso Datalab is a platform for programmatic data access and annotation services. It offers access to our data and models via API and a dedicated Python library via Jupyter notebooks. The Datalab enables custom analyses of the Impresso corpus and the semantic indexation of external document collections with the help of models created by the project.

Getting Started

Start your research with Impresso in three easy steps.

Create an Impresso account and learn how to access our API. You can run the notebooks locally or in your preferred environment — whether that's Docker, MyBinder, or Google Colab.

[REGISTER](#) OR [LOGIN](#) THEN

[ACCEPT OUR TERMS OF USE](#)

[GET YOUR API KEY](#)

Explore and Visualise your Impresso Data

Notebook templates offer complementary views on your Impresso personal collections and external datasets beyond the capabilities of the Impresso Web App.

[Open in Colab](#)
Exploring impresso with maps
by Impresso team

Annotate your Documents with Impresso Models

Use Impresso Models for the semantic indexing of your personal document collections, and compare them with Impresso Data.

Check the Impresso models available on Huggingface and choose the one that fits your needs.

Copy the code below in a blank Jupyter notebook to get started

Use a pipeline as a high-level helper
[!pip install transformers](#)



Thank you for your attention

<https://impresso-project.ch>

impresso

Media
Monitoring of the Past

Media Monitoring of the Past – *Beyond Borders*

Marten Düring, C2DH
Maud Ehrmann, EPFL-DHLAB
& Impresso Team

ONB Labs Symposium

25.11.2024 - Vienna

CC-by Marten Düring, Maud Ehrmann, Impresso Team