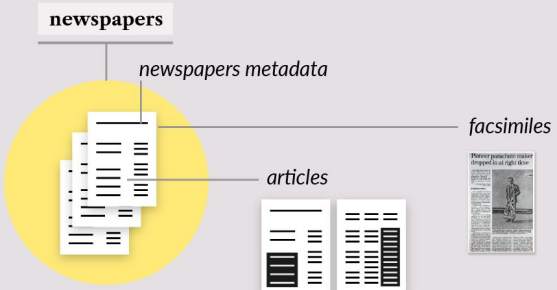
EPFL-SHS class, 3 October 2018

Maud Ehrmann & Matteo Romanello
DHLAB

Impresso overview

Objectives

Mining 200 years of historical newspapers.



1
source selection

multiple archives
multiple languages
multiple countries

historians
designers
engineers

4
co-designed web interfaces

exploratory visualisations
advanced search features

3
releases & web services

integration with original archives

2
data processing

manage data
extract text
parse and annotate articles

2.1
preprocessing

metadata mngmt.
text formatting
OCR correction

2.2
text mining

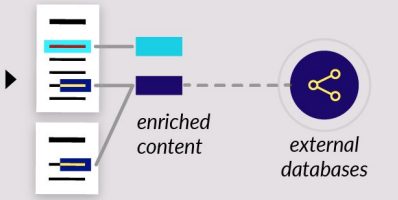
named entities processing
lexical analysis
word alignment
text similarity
topic modeling

2.3
knowledge representation

data modeling
interoperability
interlinking

2.4
benchmarking

annotation
evaluation



Objective 1: Historical media monitoring tool suite

How to adapt NLP tools to historical texts?

1. **Development of multilingual and time-specific NLP components**

   OCR post-correction, lexical processing, named entity processing, topic modelling

2. **Systematic performance assessment**

   summative and formative evaluation, shared task organized within NLP community

3. **Building of a fully traceable and interoperable historical semantic knowledge base**

   semantically indexed, structured and linked data

Objective 2: Visualization interface and visual analytics

How to explore complex and vast amounts of data?

1. **Visualization interface beyond keyword based search**, to accommodate text analysis research tools and allow users to use the system in a reflexive way.
2. **Principle of co-design**: designers, historians and computational linguists will work in close collaboration

Objective 3: Digital history

Investigating the impact of new tooling on historical research and scholarship

1. **Methodological and epistemological questions**

   source criticism & digital scholarship (how to handle digital biases)

2. **Teaching digital history**

   usage of the developed tools in the classroom

3. **Historical use case**

   resistance to the European idea

Synergies

Computational Linguists & Digital Humanists:

Mission: research

Contribution: research in NLP/DH, algorithms, tool implementation

Benefit: research

Archives & Libraries

Mission: preservation, valorisation

Contribution: sources, user needs

Benefit: enriched sources, open tools, support for prototype deployment

impresso

Media Monitoring of the Past

Designers & developers:

Mission: connect

Contribution: design and visualization expertise, interface development

Benefit: tangible products used by many people

Journalists & Publishers

Mission: inform

Contribution: sources, newspaper expertise, journalist and user needs

Benefit: enriched sources, open tools

(Digital) Historians:

Mission: research

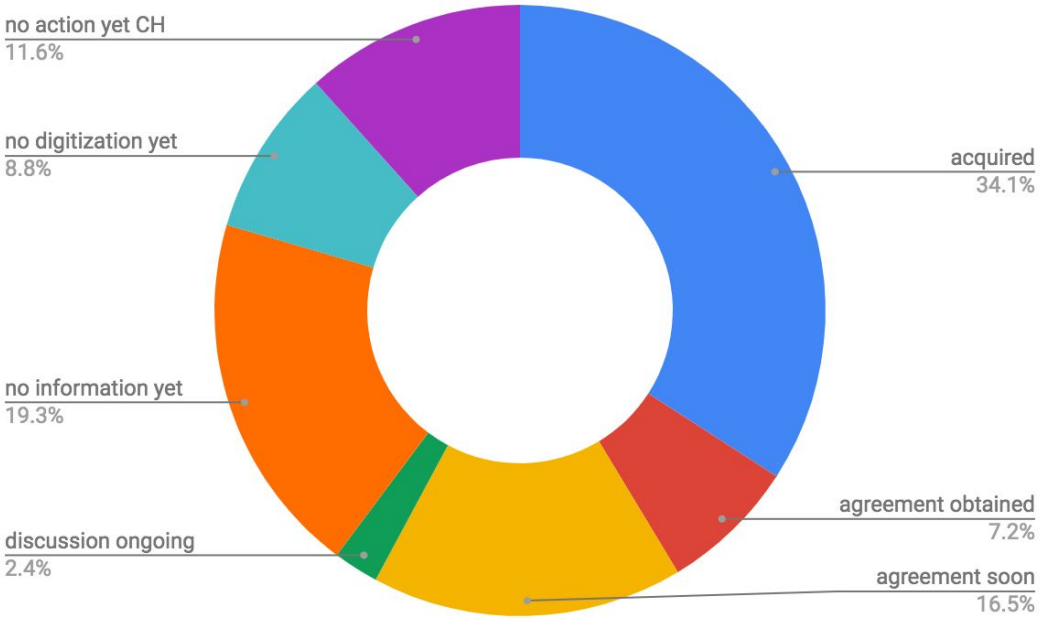
Contribution: research questions & methodology, needs, participation in co-design

Benefit: tools to support historical research

Data acquisition

Acquisition status - CH and LUX

Total: 221 titles



cf. [blog post](#) on impresso collection's state

Data policy

Copyrights

- original datasets will be used under **confidentiality agreements**
- annotations and extracted data and code will be **open source**

Strong focus on standards

- usage of International **Image Interoperability Framework** (IIIF) for images
- serialization of semantic annotations in different formats (RDF, XML)

Non-disclosure agreements

Different categories of data accessibility

public domain	no restriction
semi-closed	internal use only, non-transferrable, visible via the public interface, some functionalities require NDA
closed	internal use only, non-transferrable, visible via the public interface for associated researchers only, NDA required

NDAs signed at 2 levels:

- at consortium level: data providers ⇔ project partners (C2DH, DHLAB, ICL)
- at individual level: project partners ⇔ individual users

User access

Level 1: No restriction, full openness

Level 2: Restricted access, user has to send ID card and signed NDA

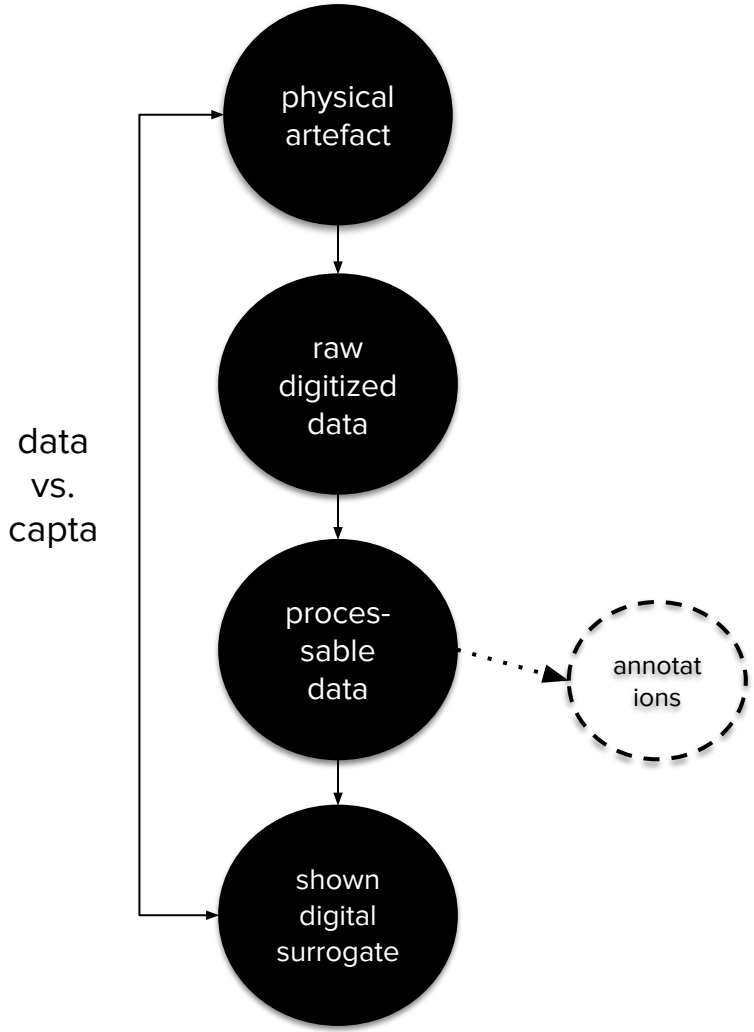
What do we do once we have the data?

A backstage tour

- beyond digitization, primary sources are further ***prepared***
- many “hidden” conversion steps happens

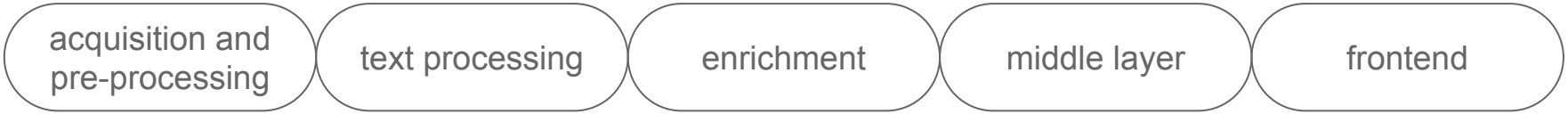
Here, **transparency** as a way to:

- raise awareness about and make comprehensible what happens to sources
- do justice to annoying tasks :)



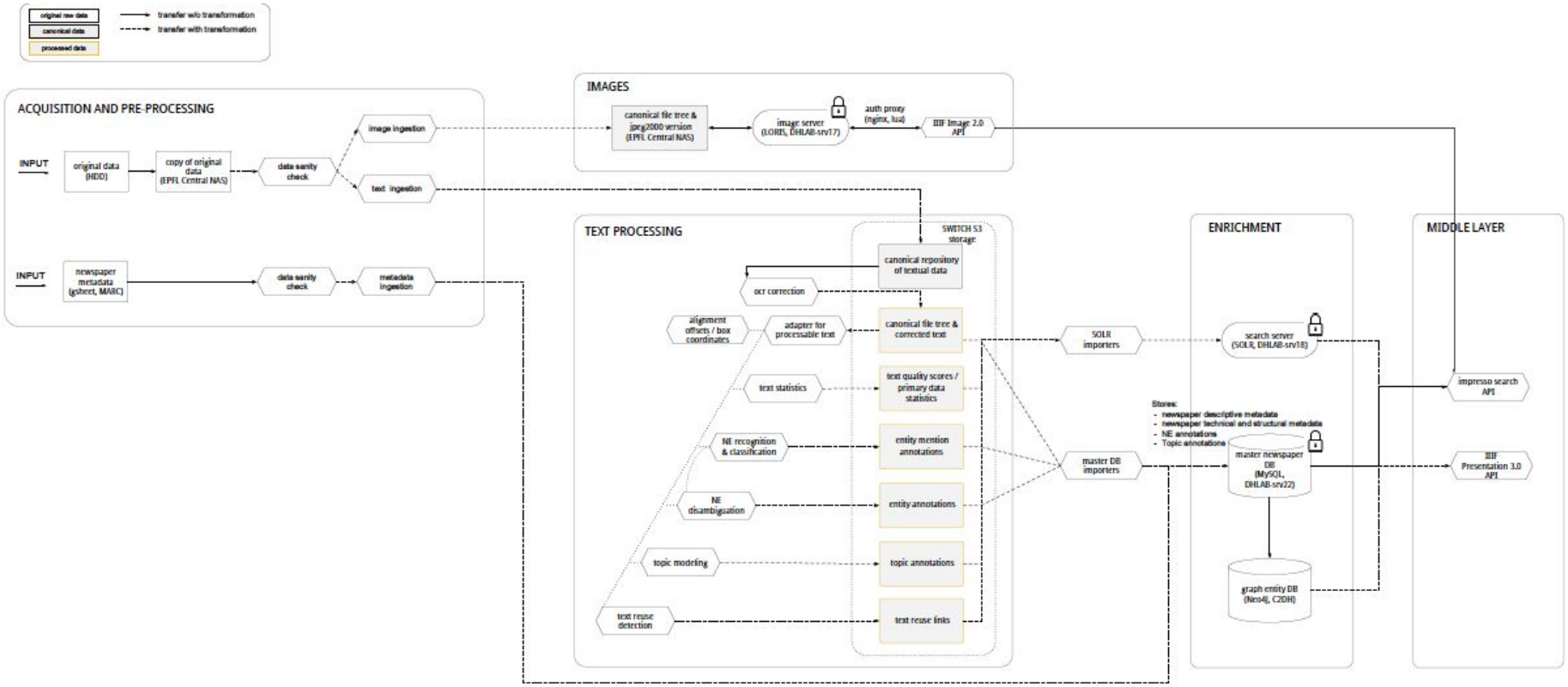
System architecture

How do we organize data, its processing, and visualisation ?

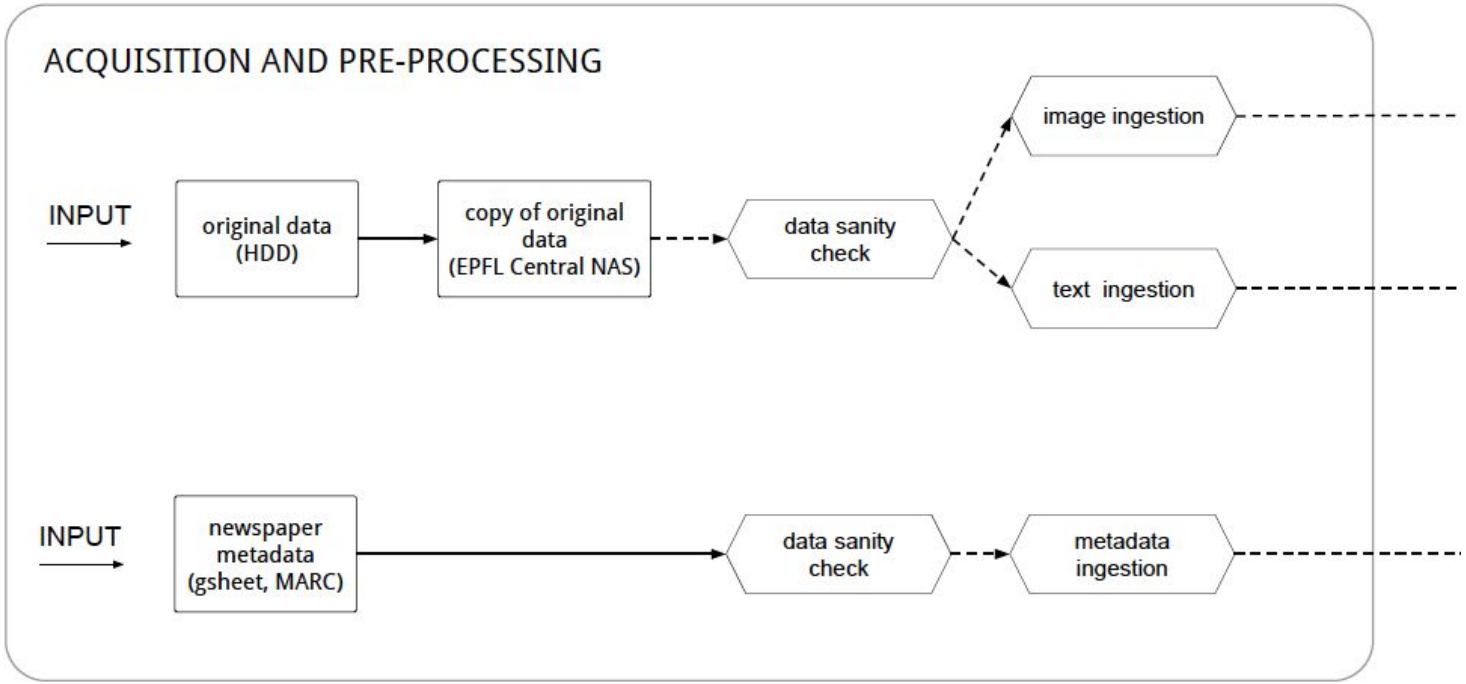


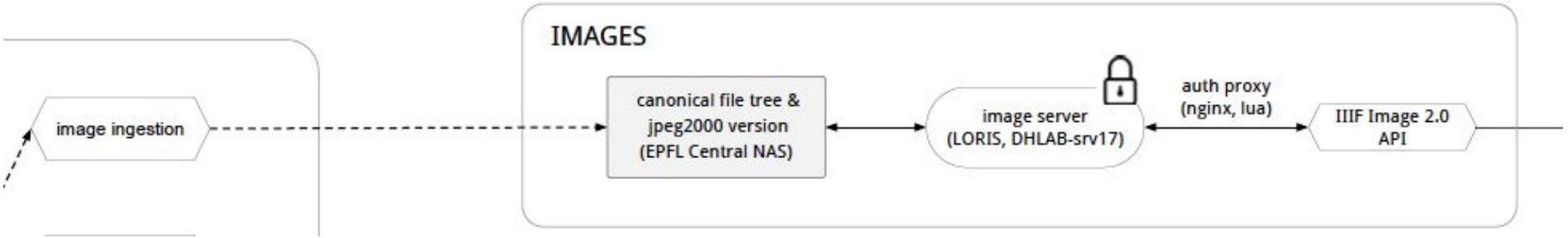
https://docs.google.com/drawings/d/1puphbUKoOes6n2sb2DA-JrQtw9GPLn0mVfWx6xoNZ6I/edit?usp=sharing

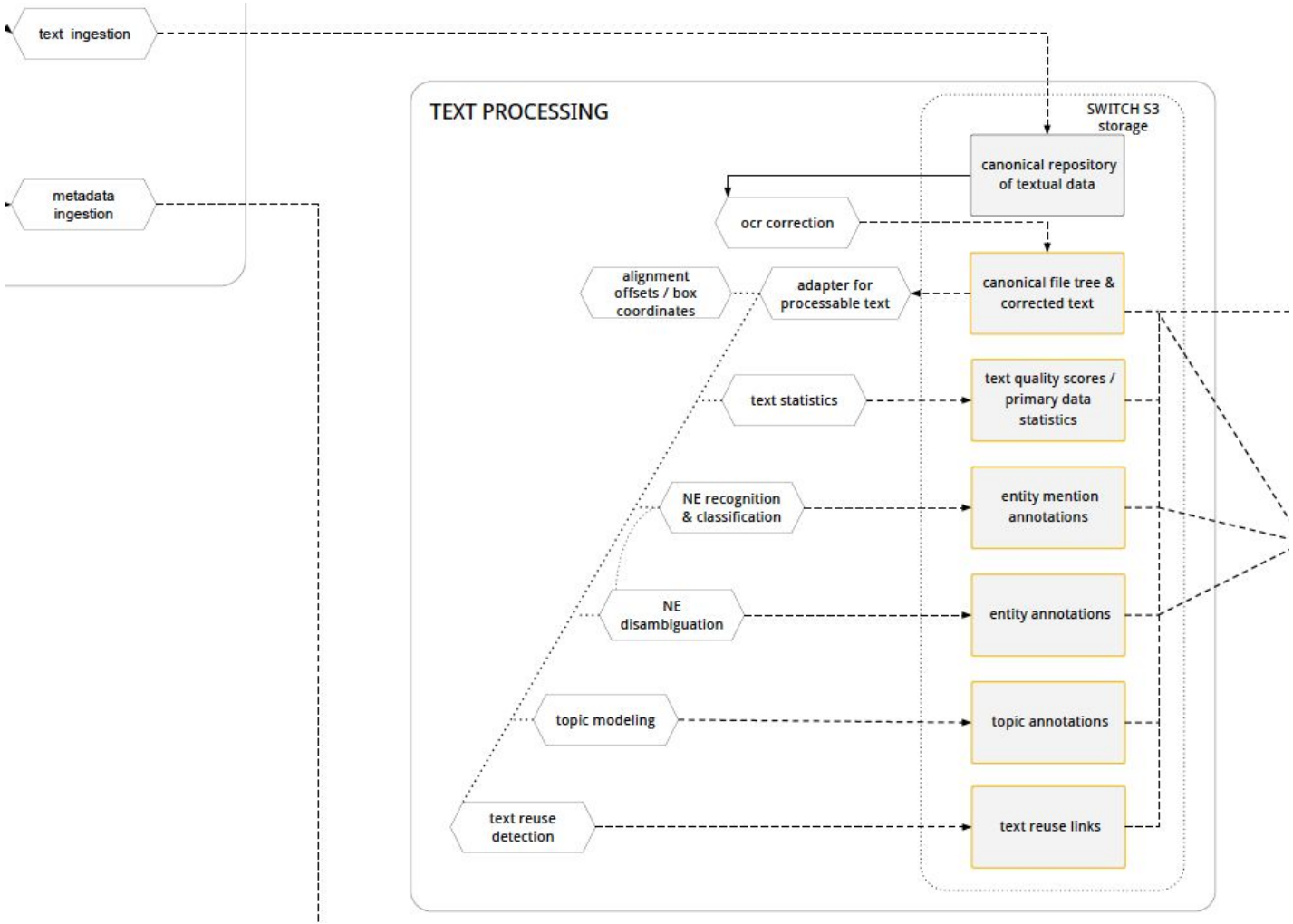
Overview

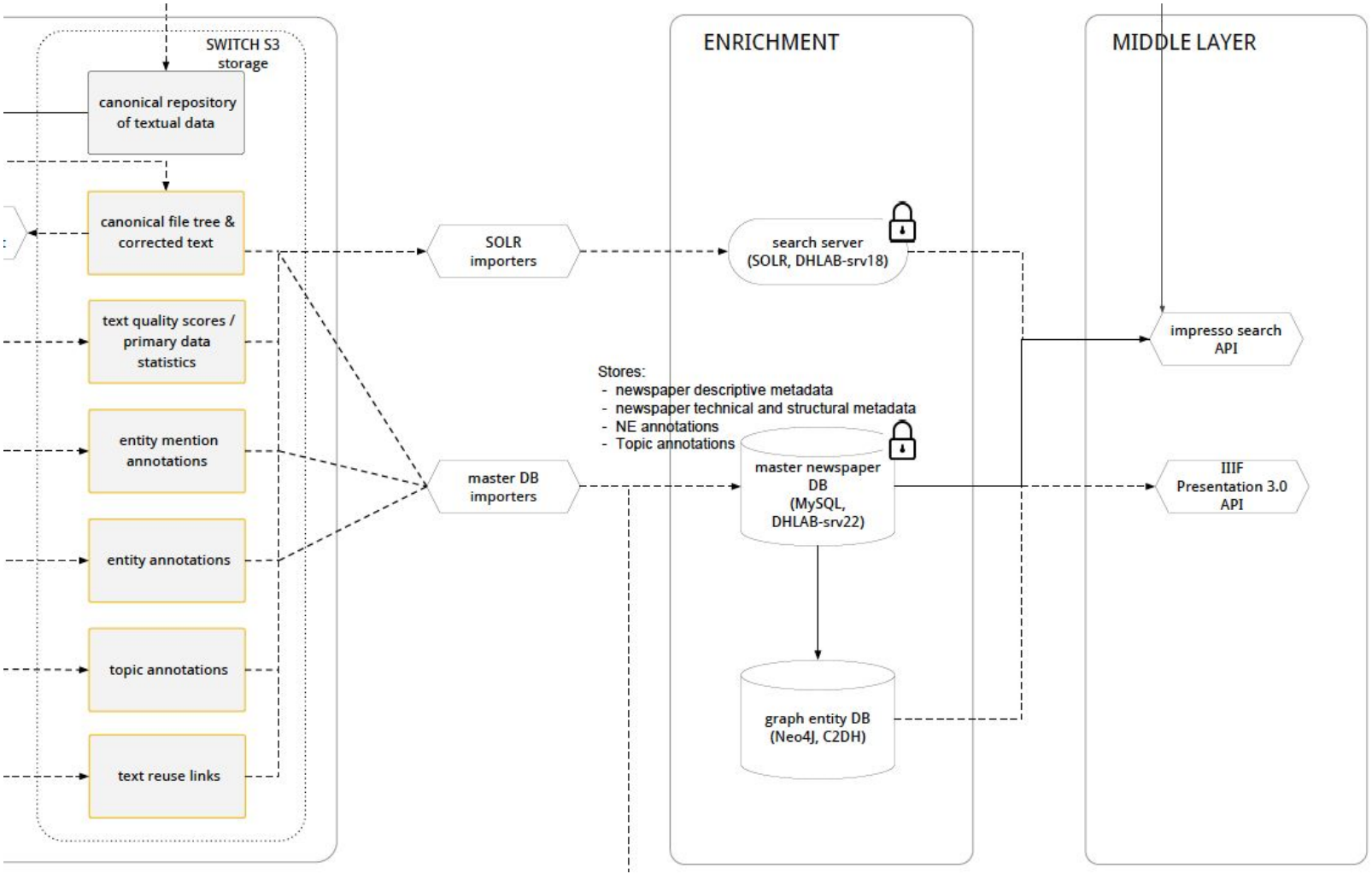

original raw data
canonical data
processed data
transfer w/o transformation
transfer with transformation
ACQUISITION AND PRE-PROCESSING
INPUT
original data (HDD)
copy of original data (EPFL Central NAS)
data sanity check
image ingestion
text ingestion
INPUT
newspaper metadata (gsheet, MARC)
data sanity check
metadata ingestion
IMAGES
canonical file tree & jpeg2000 version (EPFL Central NAS)
image server (LORIS, DHLAB-srv17)
auth proxy (nginx, lua)
IIIF Image 2.0 API
TEXT PROCESSING
SWITCH S3 storage
canonical repository of textual data
ocr correction
alignment offsets / box coordinates
adapter for processable text
canonical file tree & corrected text
text statistics
text quality scores / primary data statistics
NE recognition & classification
entity mention annotations
NE disambiguation
entity annotations
topic modeling
topic annotations
text reuse detection
text reuse links
ENRICHMENT
SOLR importers
search server (SOLR, DHLAB-srv18)
master DB importers
Stores:
- newspaper descriptive metadata
- newspaper technical and structural metadata
- NE annotations
- Topic annotations
master newspaper DB (MySQL, DHLAB-srv22)
graph entity DB (Neo4j, C2DH)
MIDDLE LAYER
impresso search API
IIIF Presentation 3.0 API


original raw data
canonical data
processed data
transfer w/o transformation
transfer with transformation


ACQUISITION AND PRE-PROCESSING
INPUT
original data (HDD)
copy of original data (EPFL Central NAS)
data sanity check
image ingestion
text ingestion
INPUT
newspaper metadata (gsheet, MARC)
data sanity check
metadata ingestion


IMAGES
image ingestion
canonical file tree &
jpeg2000 version
(EPFL Central NAS)
image server
(LORIS, DHLAB-srv17)
auth proxy
(nginx, lua)
IIIF Image 2.0
API


text ingestion
metadata ingestion
TEXT PROCESSING
SWITCH S3 storage
canonical repository of textual data
ocr correction
canonical file tree & corrected text
alignment offsets / box coordinates
adapter for processable text
text statistics
text quality scores / primary data statistics
NE recognition & classification
entity mention annotations
NE disambiguation
entity annotations
topic modeling
topic annotations
text reuse detection
text reuse links


SWITCH S3 storage
canonical repository of textual data
canonical file tree & corrected text
text quality scores / primary data statistics
entity mention annotations
entity annotations
topic annotations
text reuse links
SOLR importers
master DB importers
ENRICHMENT
search server (SOLR, DHLAB-srv18)
Stores:
- newspaper descriptive metadata
- newspaper technical and structural metadata
- NE annotations
- Topic annotations
master newspaper DB (MySQL, DHLAB-srv22)
graph entity DB (Neo4J, C2DH)
MIDDLE LAYER
impresso search API
IIIF Presentation 3.0 API


original raw data
canonical data
processed data
transfer w/o transformation
transfer with transformation

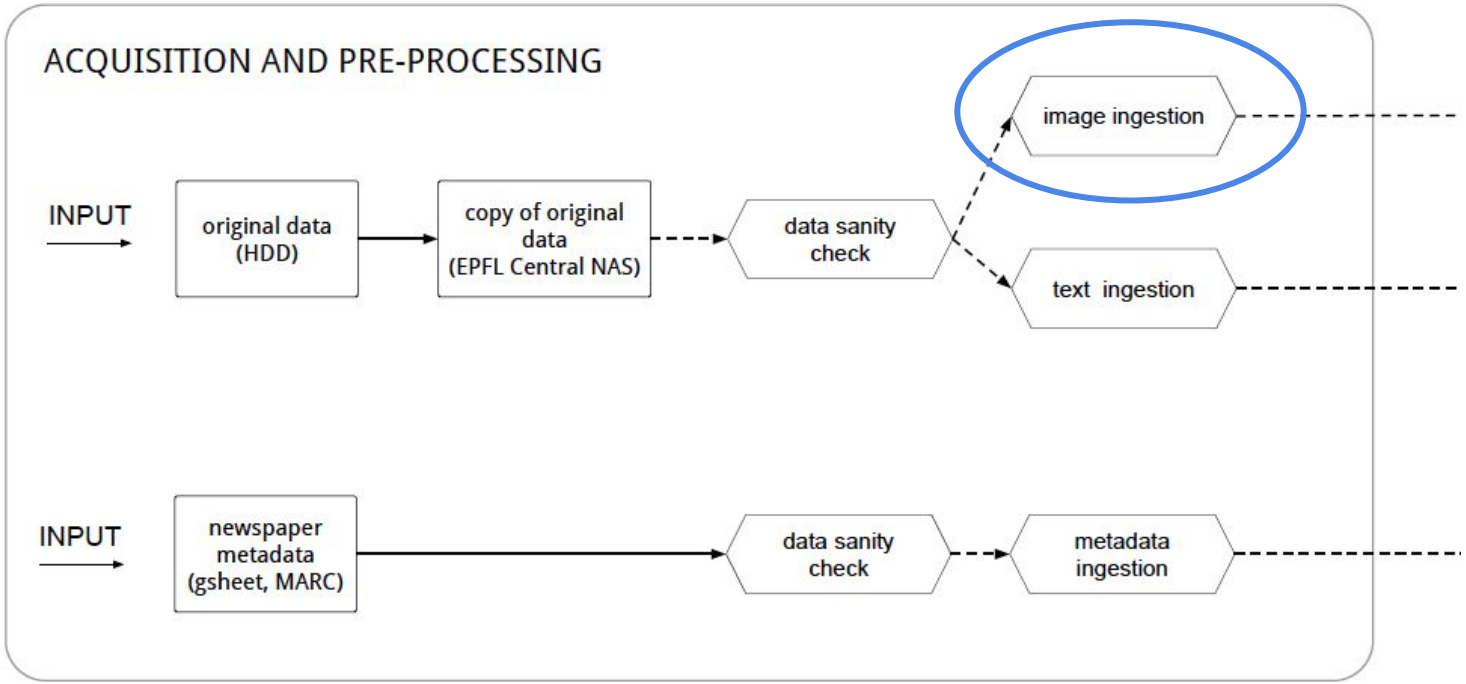

ACQUISITION AND PRE-PROCESSING
INPUT
original data (HDD)
copy of original data (EPFL Central NAS)
data sanity check
image ingestion
text ingestion
INPUT
newspaper metadata (gsheet, MARC)
data sanity check
metadata ingestion

Image processing

given that...	...therefore
not all CH institutions provides IIIF APIs	we need to get images on our servers
original files (tiff or jpeg) are heavy and therefore costly	we need to minimize storage cost by reducing image quality
we might need to re-OCR some material	we need to keep a certain image quality
users will intensively visualize images	the viewing experience must be the best possible

What is the best compromise between storage cost, user experience, and OCR performances ?

Image processing

Choice: JPEG2000

- better compression/quality ratio
- encoding of tiles, which can be delivered quickly by image servers

Evaluation

- dataset: 7 pages from GDL 1850/1900/1950 manually transcribed
- test: different images format with different compression ratio
- best tradeoff between image size and OCR performance:
  JPEG2000 with compression ratio -10, 70% size gain, >1% OCR performance loss.

Image processing - Difficulties

in one archive, image files:

- can be in different formats (tiff, jpg, png) ➜ different conversions steps
- can have different resolutions ➜ recomputation of box coordinates
- can be missing ➜ recovering strategies, keeping counts

IIIF - International Interoperable Image Framework

A common language for computers to talk about images.

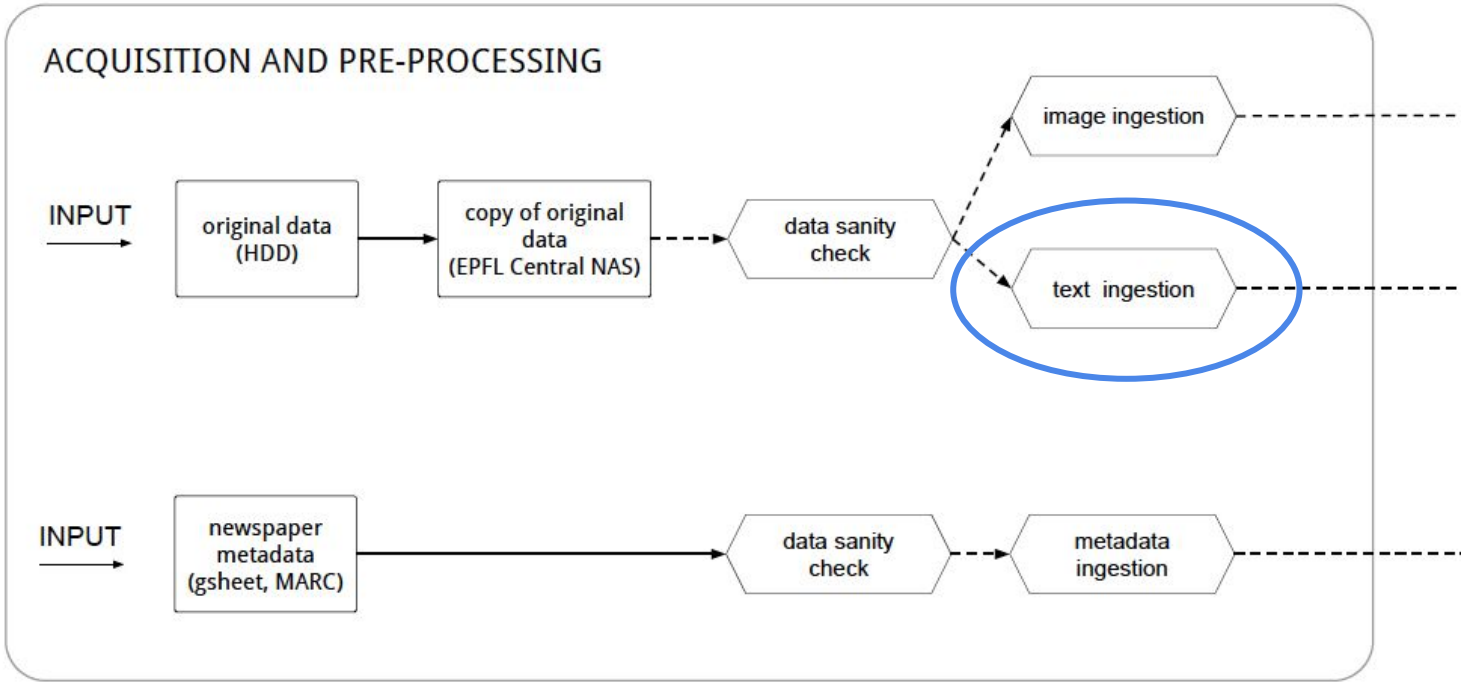
Enable to easily access any part of any page:

https://dhlabsrv17.epfl.ch/iiif_impresso/GDL-1950-04-27-a-p0007/198,5240,131,33/full/0/default.jpg

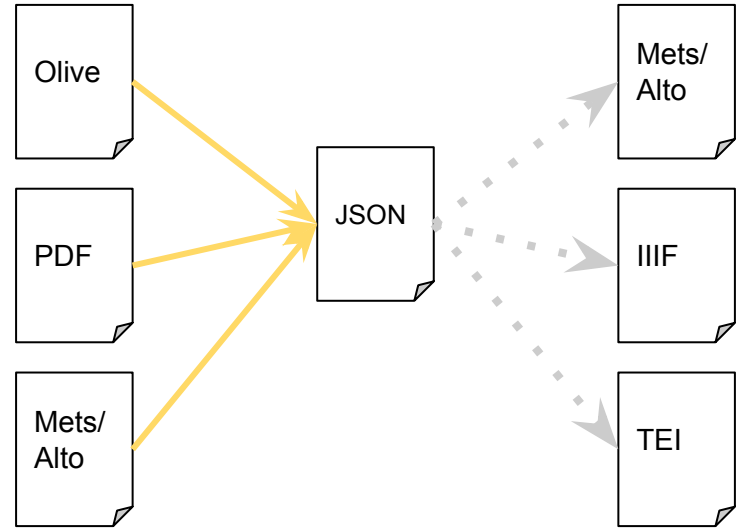
https://dhlabsrv17.epfl.ch/iiif_impresso/GDL-1950-04-27-a-p0007/117,5090,596,863/full/0/default.jpg

https://dhlabsrv17.epfl.ch/iiif_impresso/GDL-1950-04-27-a-p0007/full/full/0/default.jpg


original raw data
canonical data
processed data
transfer w/o transformation
transfer with transformation


ACQUISITION AND PRE-PROCESSING
INPUT
original data (HDD)
copy of original data (EPFL Central NAS)
data sanity check
image ingestion
text ingestion
INPUT
newspaper metadata (gsheet, MARC)
data sanity check
metadata ingestion

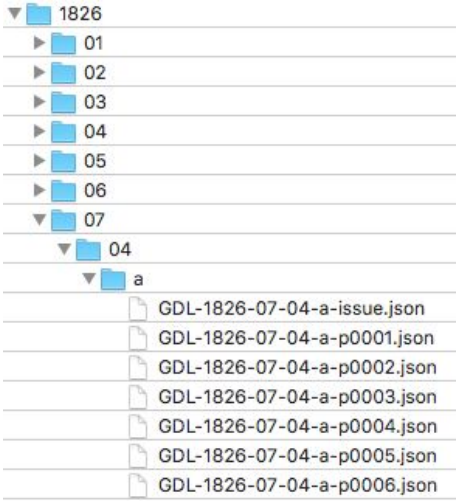
Text acquisition



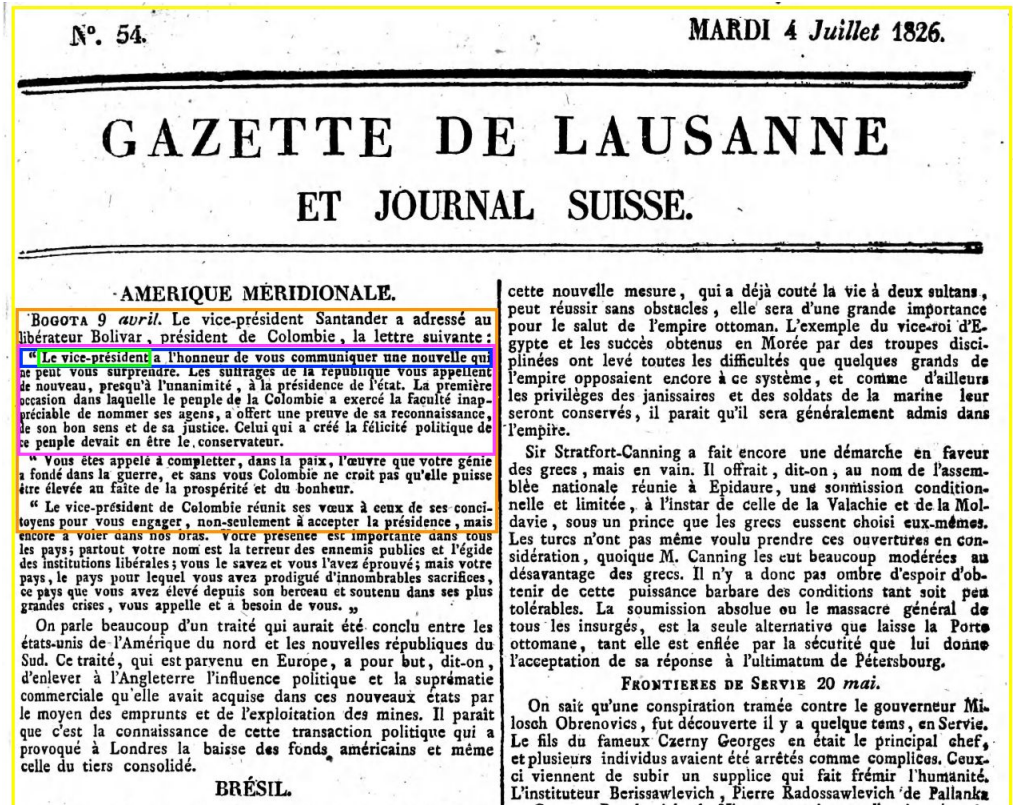
SDSC

SWITCH

S3



Impresso canonical JSON



N°. 54. MARDI 4 *Juillet* 1826.

GAZETTE DE LAUSANNE
ET JOURNAL SUISSE.

AMERIQUE MÉRIDIONALE.

BOGOTA 9 *avril*. Le vice-président Santander a adressé au libérateur Bolivar, président de Colombie, la lettre suivante :

" Le vice-président a l'honneur de vous communiquer une nouvelle qui ne peut vous surprendre. Les suffrages de la république vous appellent de nouveau, presqu'à l'unanimité, à la présidence de l'état. La première occasion dans laquelle le peuple de la Colombie a exercé la faculté inappréciable de nommer ses agens, a offert une preuve de sa reconnaissance, de son bon sens et de sa justice. Celui qui a créé la félicité politique de ce peuple devait en être le conservateur.

" Vous êtes appelé à compléter, dans la paix, l'œuvre que votre génie a fondé dans la guerre, et sans vous Colombie ne croit pas qu'elle puisse être élevée au faîte de la prospérité et du bonheur.

" Le vice-président de Colombie réunit ses vœux à ceux de ses concitoyens pour vous engager, non-seulement à accepter la présidence, mais encore à voler dans nos bras. Votre présence est importante dans tous les pays; partout votre nom est la terreur des ennemis publics et l'égide des institutions libérales; vous le savez et vous l'avez éprouvé; mais votre pays, le pays pour lequel vous avez prodigué d'innombrables sacrifices, ce pays que vous avez élevé depuis son berceau et soutenu dans ses plus grandes crises, vous appelle et a besoin de vous. „

On parle beaucoup d'un traité qui aurait été conclu entre les états-unis de l'Amérique du nord et les nouvelles républiques du Sud. Ce traité, qui est parvenu en Europe, a pour but, dit-on, d'enlever à l'Angleterre l'influence politique et la suprématie commerciale qu'elle avait acquise dans ces nouveaux états par le moyen des emprunts et de l'exploitation des mines. Il paraît que c'est la connaissance de cette transaction politique qui a provoqué à Londres la baisse des fonds américains et même celle du tiers consolidé.

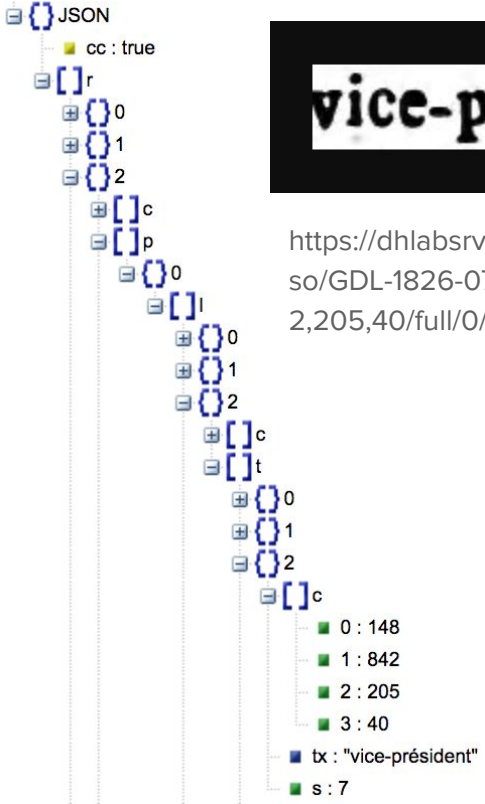
BRÉSIL.

cette nouvelle mesure, qui a déjà couté la vie à deux sultans, peut réussir sans obstacles, elle sera d'une grande importance pour le salut de l'empire ottoman. L'exemple du vice-roi d'Egypte et les succès obtenus en Morée par des troupes disciplinées ont levé toutes les difficultés que quelques grands de l'empire opposaient encore à ce système, et comme d'ailleurs les privilèges des janissaires et des soldats de la marine leur seront conservés, il parait qu'il sera généralement admis dans l'empire.

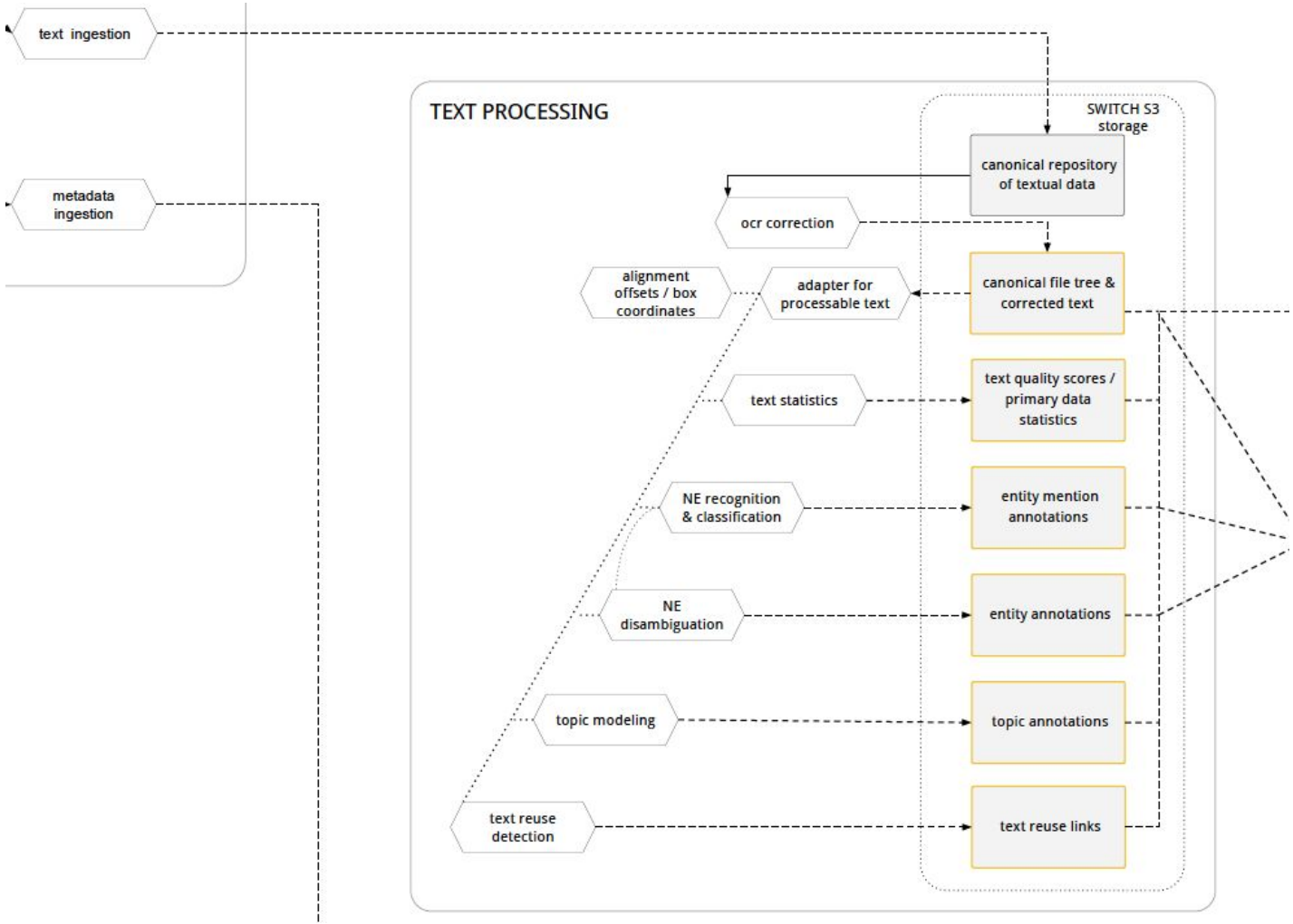
Sir Stratfort-Canning a fait encore une démarche en faveur des grecs, mais en vain. Il offrait, dit-on, au nom de l'assemblée nationale réunie à Epidaure, une soumission conditionnelle et limitée, à l'instar de celle de la Valachie et de la Moldavie, sous un prince que les grecs eussent choisi eux-mêmes. Les turcs n'ont pas même voulu prendre ces ouvertures en considération, quoique M. Canning les eut beaucoup modérées au désavantage des grecs. Il n'y a donc pas ombre d'espoir d'obtenir de cette puissance barbare des conditions tant soit peu tolérables. La soumission absolue ou le massacre général de tous les insurgés, est la seule alternative que laisse la Porte ottomane, tant elle est enflée par la sécurité que lui donne l'acceptation de sa réponse à l'ultimatum de Pétersbourg.

FRONTIERES DE SERVIE 20 *mai*.

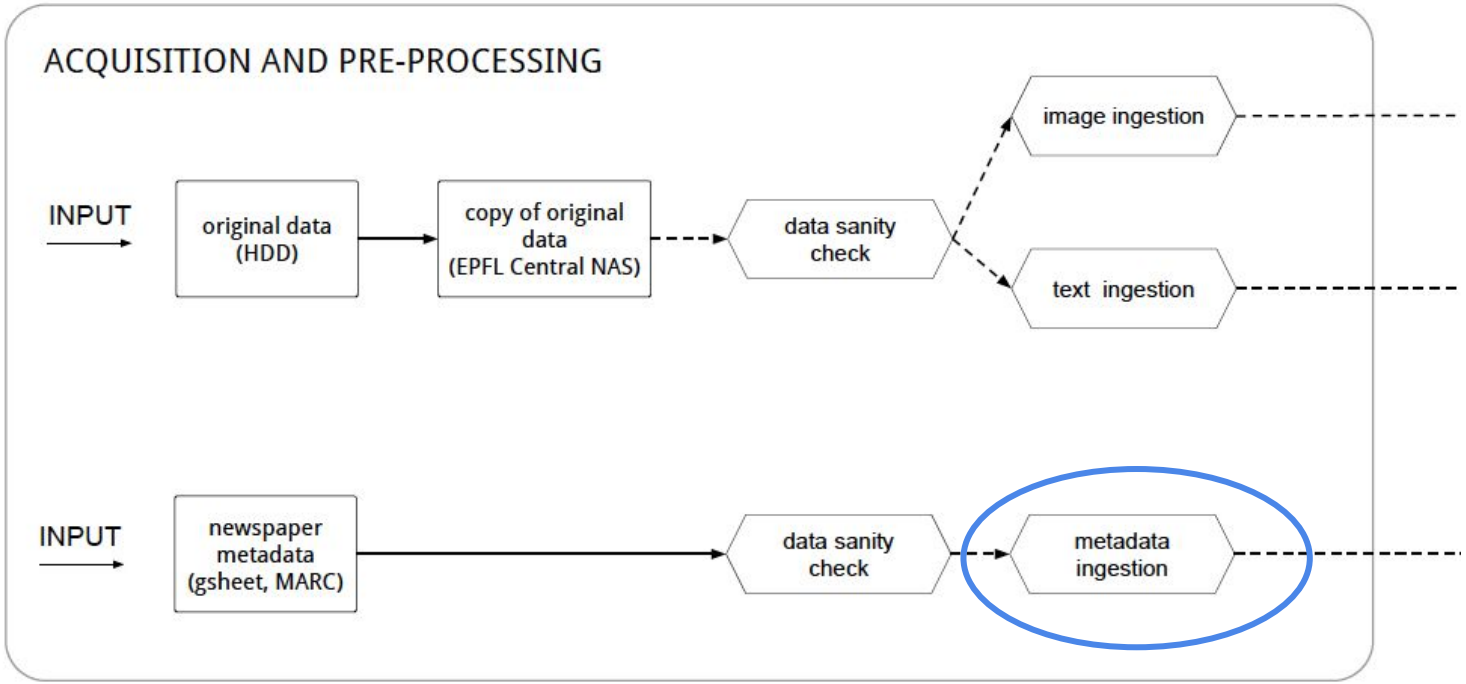
On sait qu'une conspiration tramée contre le gouverneur Milosch Obrenovics, fut découverte il y a quelque tems, en Servie. Le fils du fameux Czerny Georges en était le principal chef, et plusieurs individus avaient été arrêtés comme complices. Ceux-ci viennent de subir un supplice qui fait frémir l'humanité. L'instituteur Berissawlevich, Pierre Radossawlevich de Pallanka



https://dhlabsrv17.epfl.ch/iiif_impresso/GDL-1826-07-04-a-p0001/148,842,205,40/full/0/default.jpg


text ingestion
metadata ingestion
TEXT PROCESSING
SWITCH S3 storage
canonical repository of textual data
ocr correction
canonical file tree & corrected text
alignment offsets / box coordinates
adapter for processable text
text statistics
text quality scores / primary data statistics
NE recognition & classification
entity mention annotations
NE disambiguation
entity annotations
topic modeling
topic annotations
text reuse detection
text reuse links


original raw data
canonical data
processed data
transfer w/o transformation
transfer with transformation


ACQUISITION AND PRE-PROCESSING
INPUT
original data (HDD)
copy of original data (EPFL Central NAS)
data sanity check
image ingestion
text ingestion
INPUT
newspaper metadata (gsheet, MARC)
data sanity check
metadata ingestion

Newspaper metadata

What is it useful for?

1. filter
2. provide contextual knowledge

What is it?

What	Example	Provenance	Disputable
hard facts	title, long title, other title, dates, print run, editor, founder, periodicity, etc.	expert knowledge (librarians, historians)	no
soft facts	political orientation, press type, etc.	expert knowledge	yes
computational hard facts	number of issues, of pages, evolution of volume through time	computation	no

Newspaper metadata

Questions

- which information is useful as a filter vs. as contextual knowledge?
- how to collect expert knowledge? what is easy to get?

Current decisions

- library metadata as authority information
- filtering is enabled for most important not disputable facts
- we cannot provide everything but: the infrastructure to enrich/edit/visualize

Archival holes

External factors

1. publication periodicity
2. the journal stopped for some years/months/days

Preservation factors

3. paper copies were lost and never digitized
4. digitization happened but digital copies were lost or damaged

Transparency about

1 & 2: requires external historical knowledge, is often encoded as metadata

3: requires preservation history knowledge, is rarely encoded as metadata

4: detection possible via quality control of digitization process, is almost never encoded.

Archival holes

Impresso will encode:

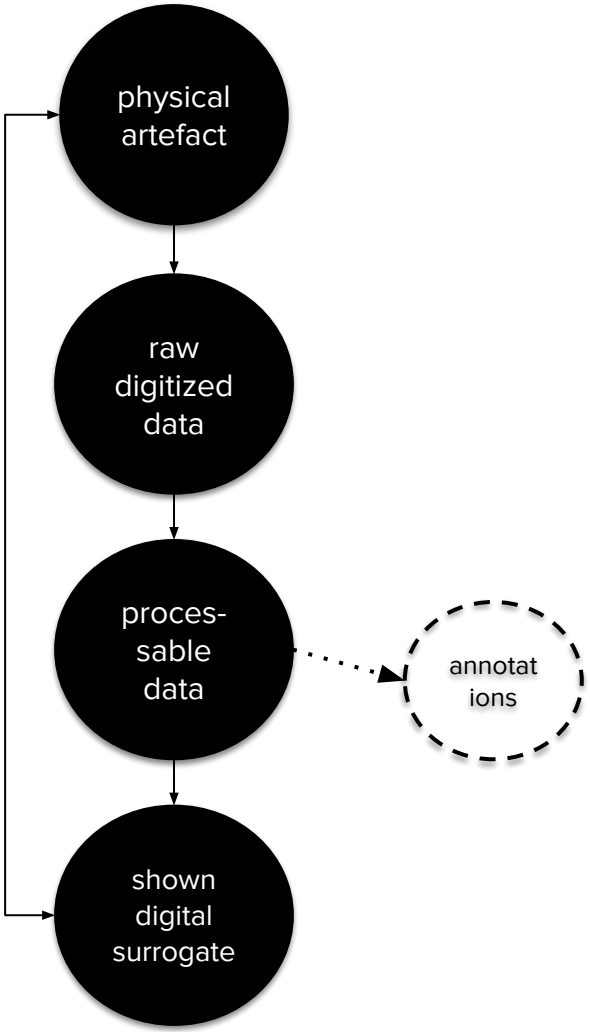
- expected page/issue numbers (what the institution think she has)
- delivered numbers (what exists on the hard disk)
- ingested numbers (what was possible to process)

➜ users know a bit better what they see

Data vs. *Capta*

"Differences in the etymological roots of the terms data and capta make the distinction between constructivist and realist approaches clear. Capta is "taken" actively while data is assumed to be a "given" able to be recorded and observed. From this distinction, a world of differences arises. Humanistic inquiry acknowledges **the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, taken, not simply given as a natural representation of pre-existing fact**."

J. Drucker (2011), "Humanities Approaches to Graphical Display" in *DHQ* 5(1), , http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html#p3



Named entity processing (quick intro)

NE definition

- “elements of interest” generally of type *Person*, *Org* and *Location*
- referential units which underlie text semantics
- proper names & definite descriptions (referential autonomy and unicity)

Origins of NE processing

- **1980's: Text Understanding**
  an over-ambitious project facing technical and theoretical difficulties
- **1990's: Information Extraction**
  focus on elements of interest
  extraction model defined in advance based on application
  analysis of only 10-20% of the text
- **1987-1998: Message Understanding Conference (MUC) cycle**
  financed by U.S. DARPA
  MUC-6 1995, definition of "named entity" subtask

MUC 3 Template example (1991)

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).	
INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

NE tasks

- **recognition**: detecting named entities in texts with correct boundaries
- **classification**: categorizing NE according to a predefined typology
- **disambiguation**: linking NE mentions to a unique reference
- **relation detection**: discovering and specifying relations between NE

NE tasks

Le ministre des Affaires étrangères M. Burkhalter a esquissé les priorités du

Conseil fédéral pour nouer de nouvelles relations bilatérales avec le Royaume-Uni.

(Le Temps 25.06.2016)

NE tasks

Le ministre des Affaires étrangères ***M. Burkhalter*** *a esquissé les priorités du* ***Conseil fédéral*** *pour nouer de nouvelles relations bilatérales avec le* ***Royaume-Uni****.* (Le Temps 25.06.2016)

NE tasks

Le ministre des Affaires étrangères ***M. Burkhalter*** *a esquissé les priorités du*

PERSON

Conseil fédéral *pour nouer de nouvelles relations bilatérales avec le*

ORGANIZATION

Royaume-Uni. (Le Temps 25.06.2016)

LOCATION

NE tasks

http://dbpedia.org/page/Didier_Burkhalter

Le ministre des Affaires étrangères ***M. Burkhalter*** *a esquissé les priorités du*

FUNCTION TITLE **PERSON**

Conseil fédéral *pour nouer de nouvelles relations bilatérales avec le*

ORGANIZATION

Royaume-Uni. (Le Temps 25.06.2016)

LOCATION

Performances

Good when:

- language: English
- domain: news
- typology: simple

NE processing on historical data, challenges

- Noisy inputs

  inherent in the source or from post-processing (OCR)

  e.g. *Constap. iipopjle*, *Buch irest*, *M'' Lucile*

- Language evolution

  spelling variants, old naming conventions

  Härnevi ➜ *Arnevi*, *Kallmar* ➜ *Kalmar*[*]

- Poor resource coverage

  - minor or unknown entities, esp. for ORG type
  - lack of appropriate trigger words

  e.g. *bourgmestre*, *tailleur*, spécialiste en munitions

Diachronic evaluation - Le Temps

- 7 time-series
- 40 articles from GDL, 10 from JDG
- manual annotation of *Person* and *Location* (QUAERO guidelines)

	# words		# pers		# loc		# entities	
	GDL	JDG	GDL	JDG	GDL	JDG	GDL	JDG
1804	33,773	-	417	-	990	-	1,407	-
1826	33,353	14,074	471	184	946	151	1,417	335
1841	40,784	5,558	553	70	1,137	55	1,690	125
1881	55,751	12,360	950	227	912	280	1,862	507
1921	20,117	3,587	377	47	572	136	949	183
1961	23,332	8,301	529	115	556	149	1,085	264
1981	17,759	3,672	258	79	363	56	621	135
TOTAL	299,212	65,139	3,555	722	5,476	827	9,031	1,549

Table 1: Data set statistics.

Systems

• symbolic:	**rule-based** system based on ExPRESS formalism
• machine learning:	**mXS**, supervised learning of extraction patterns
• hybrid:	**AlchemyAPI**, supervised classification and rules
• knowledge graph:	**DandelionAPI**

Systems applied out of the box without adaptation.

Metrics

- Precision, Recall and F-measure
- Slot Error Rate (SER), a measure or error

  Insertion,
  Deletion,
  Substitution of Type,
  Substitution of Boundary,
  Subs. of Type and Boundary.

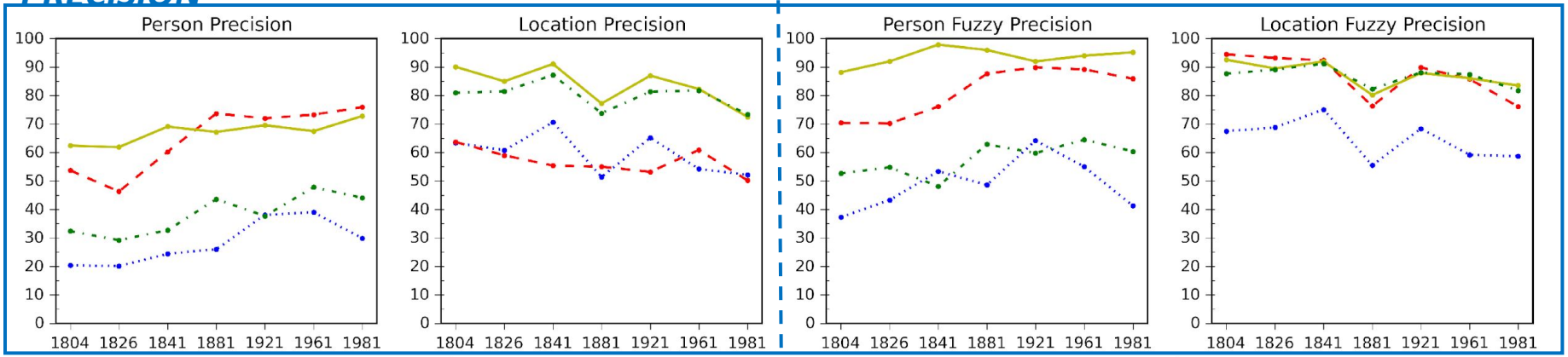
$$SER = \frac{D + I + STB + 0.5 \times (ST + SB)}{R}$$

- Normal and Fuzzy setting

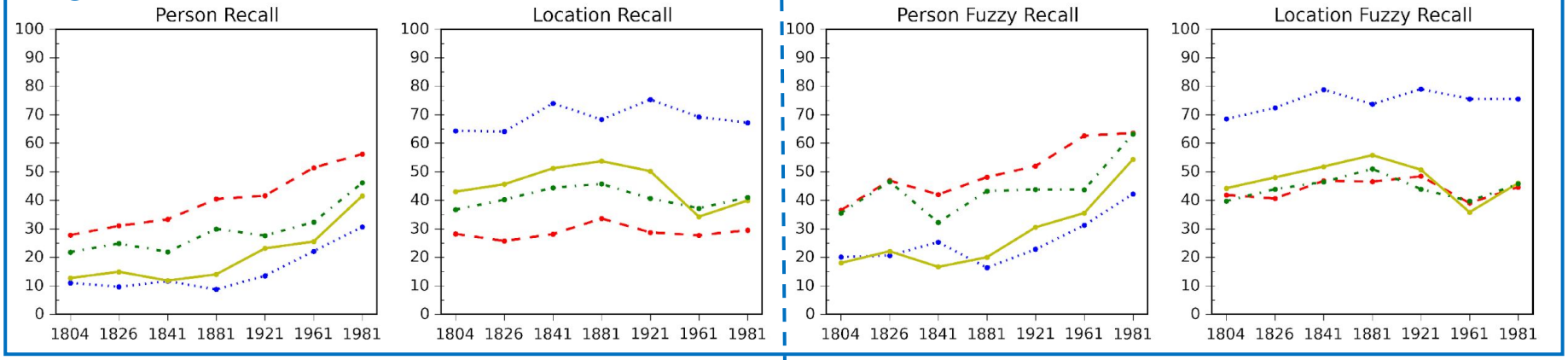
PRECISION

Normal

Fuzzy



RECALL



Some insights

Do NER tool performances gradually decrease when going back in time?

- yes, but with irregularities
- yes, but more for *Person* than for *Location*
- yes, but not the same way for all systems/approaches

cf. paper: Diachronic evaluation of NER systems on old newspapers

Text re-use detection (quick intro)

Text reuse – working definition

Text reuse is the **meaningful reiteration of text**, usually beyond the simple repetition of common language.
Such a broad concept can naturally be understood at different levels and studied in a large variety of contexts.

http://dharchive.org/paper/DH2014/Panel-106.xml

Contexts:

- publishing/teaching ➜ plagiarism
- literary studies ➜ intertextuality (quotes, allusions, paraphrases)
- historical newspapers ➜ text reprinting, copy+paste, information spreading

Passim	D. Smith et al.	Java / Scala	https://www.etrap.eu/research/tracer/
Tracer	M. Büchler	Java	https://www.etrap.eu/research/tracer/
MatchMaker*	R. Snyder	Python	https://github.com/JSTOR-Labs/matchmaker
Tesserae*	N. Coffee et al.	Perl	https://github.com/tesserae/tesserae
…			

Passim

Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers

David A. Smith, Ryan Cordell, and Abby Mullen*

American Literary History, vol. 27, no. 3, pp. E1–E15
http://dx.doi.org/10.1093/alh/ajv029

Step 1: search of candidate document pairs

Step 2: local document alignment

Step 3: passage clustering

http://github.com/dasmiq/passim

Passim algorithm

Step 1: search of candidate document pairs

Goal: reduce total number of comparisons to perform

1.1 **shingling**:

- efficient document indexing via n-grams
- document ➜ set of 5-word sequences (5-grams)
- singleton n-grams are filtered out (> 50%)

1.2 **extraction and filtering of candidate pairs**

- suppress repeated n-grams within same series
- suppress n-grams leading to > 5k document pairs
- filter out document pairs with < 5 shared n-grams

Step 2: local document alignment

Step 3: passage clustering

5-grams

devant le verdict de l'expert sur la responsabilité et réclame, en raison de la violence

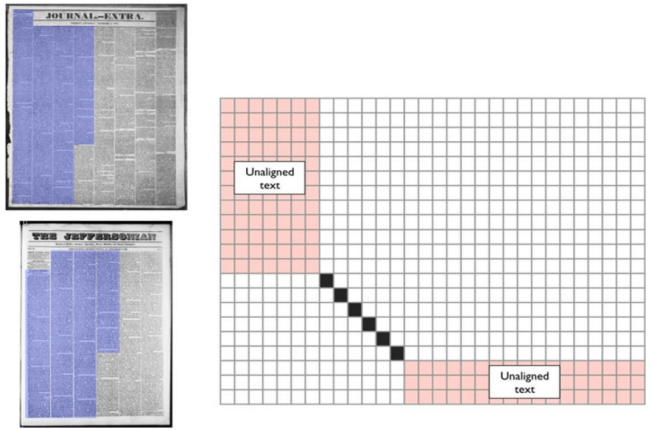
1. devant le verdict de l'expert
2. le verdict de l'expert sur
3. verdict de l'expert sur la
4. de l'expert sur la responsabilité
5. ...

Passim algorithm

Step 1: search of candidate document pairs

Step 2: local document alignment

Goal: given a set of document pairs, obtain a set of aligned passage pairs



*source of illustration: Smith et al. 2015, Fig. 4, p. E9

Step 3: passage clustering

Passim algorithm

Step 1: search of candidate document pairs

Step 2: local document alignment

Step 3: passage clustering

Goal: group similar passages into clusters

- “single-link” clustering
- overlap threshold 50%

Passim – experiment on GDL/JDG

Data Preparation

- JSONlines format
- size of data:
    - 200 years
    - 15 Gb

/GDL-1841-5.jsonl | jq "."


```
{
  "text": "<full_text>''100 LOTERIE D'ARGENT DE FRANCFORT SUR-LE-MEIN. Elle est divisée en 6 classes et compos
ée d'uncapital de Un Million 822,500 florins a\"Empire, au Louis d'or à 11 fl. 2600 billets, dont 13,500 lots
elS 4 primes, outreun grand nombre de billets franc » Les principaux lots sont : 2 lots de fl. 100,000 chacun.
 1 lot de 50,000. 2 lots de 25,000 chacun. 2 lots de 20,000 chacun. Ilot de 15,000. 1 lot de 12,000. 4 lots de
 10,000 chacun. 1 lot de 6,000. 5 lots de 5,000 chacun. 98 lots de 4000 à 1000.94 lots de 600 à 300.5880 lots
de 250 à 100 florins, non compris un grand nombre de moindres gains. Cette loterie est établie et garantie par
 le gouvernement de la ville libre de Francfortbasée sur les principes les plus loyaux et avantageux pour les
joueurs, elle jouit d'une confiance et d'un crédit général et bien mérité. Le plan, offert gratis aux amateurs
, en contient le bilan exact et les conditions ultérieures ; la' première classe sera tirée les 9 et 10 Juin p
rochain ; un billet entier coûte 6 fl ., un demi 3 fl ., un tiers 2 fl ., un quart 1 fl. \" 50, payables compt
ant. Le soussigné offre son entremise aux personnes qui voudront s'y intéresser ; il désignera dans le but de
faciliter les paiemens des mises, aux persoonnes à qui cela pourra convenir une maison de commerce, se trouvan
t le plus à leur portée. J .-B. ZUNDEL i à Schaffouse.</full_text>",
  "page_no": [
    "5"
  ],
  "name": "Untitled Article",
  "date": 1841,
  "series": "GDL",
  "id": "GDL-1841-05-21-a_Ar00504"
}
```


Technical Setup

- one cluster machine (node)
    - 48 cores, ~280Gb RAM
- run with default parameters
    - used 36 cores
    - **3.5 hours** to complete
    - given 100Gb to spark executor/driver

Passim – output examples

GDL 03/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au bateau à vapeur parti d'Ouchy pour Genève à 2 heures 20 minutes. Le bateau était arrêté pour le débarquement el l'embarquement devant Nyon, lorsque tout à coup on entendit une détonation, un nuage de fumée (ou de vapeur ?) sor tit de la machine et le bâtiment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rupture qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soit, le bateau était mis dans l'impossibilité de continuer sa route ; heureusement les voyageurs qui devaient aller plus loin ont pu prendre le train qui passe à Nyon à 5 heures 3 minutes el sont arrivés à bon port, sans autre mal qu'un moment de frayeur. On

JDG 05/12/1863

— Mardi il est arrivé un accident, sans suites fâcheuses, au Guillaume-Tilt, parti d'Ouchy pour Genève a 2 heures 20 minutes. Le bateau était arrêt é Pour le débarquement devant Nyou, lorsque tout à coup on entendit une détonation, un nuage de fu- n) p e (ou de vapeur ?) sortit de la hiadiine, et le hû liment subit une violente secousse. Nous ne savons pas d'une manière précise en quoi consiste la rup'" re qu'il y a eu ; on parle d'une clavette cassée. Quoi qu'il en soii, l e bateau était misdans l'impostJuilite de continuer sa route ; h'ureuiipnient les v < y 'geursqui devaient aller plus l <> iri ont pu piendre le train qui passe à Nyon à 5 heures 3 minutes, et sont arrivés ù bon poil, sans aulre mal qu'un moment de frayeur. Le

Passim – output examples

GDL 28/04/1980

Faux chèque et faux cheval Lausanne, 27 (ATS).-On a confirmé dimanche soir à l'ATS, de source autorisée, une affaire d'escroquerie d'environ trois millions de dollars qui s'est passée en automne dernier à Lausanne et que le quotidien genevois « La Suisse » a révélée dimanche. Signé au siège lausannois d'une banque suisse, un contrat de vente portait sur un cheval de bronze considéré par le vendeur comme une antiquité grecque absolument unique. Après avoir acquis cette pièce rare d'un antiquaire suisse, ce vendeur, un Italien associé à une société ayant son siège à Panama, l'avait revendue à un riche arabe. [...]

JDG 28/04/1980

Faux chèque et faux cheval GROSSE ESCROQUERIE À LAUSANNE Lausanne, 27 (ATS).-On a confirmé dimanche soir à l'ATS, de source autorisée, une affaire d'escroquerie d'environ trois millions de dollars qui s'est passée en automne dernier à Lausanne et que le quotidien genevois « La Suisse » a révélée dimanche. Signé au siège lausannois d'une banque suisse, un contrat de vente portait sur un cheval de bronze considéré par le vendeur comme une antiquité grecque absolument unique. Après avoir acquis cette pièce rare d'un antiquaire suisse, ce vendeur, un Italien associé à une société ayant son siège à Panama, l'avait revendue à un riche arabe. [...]

Passim – output examples

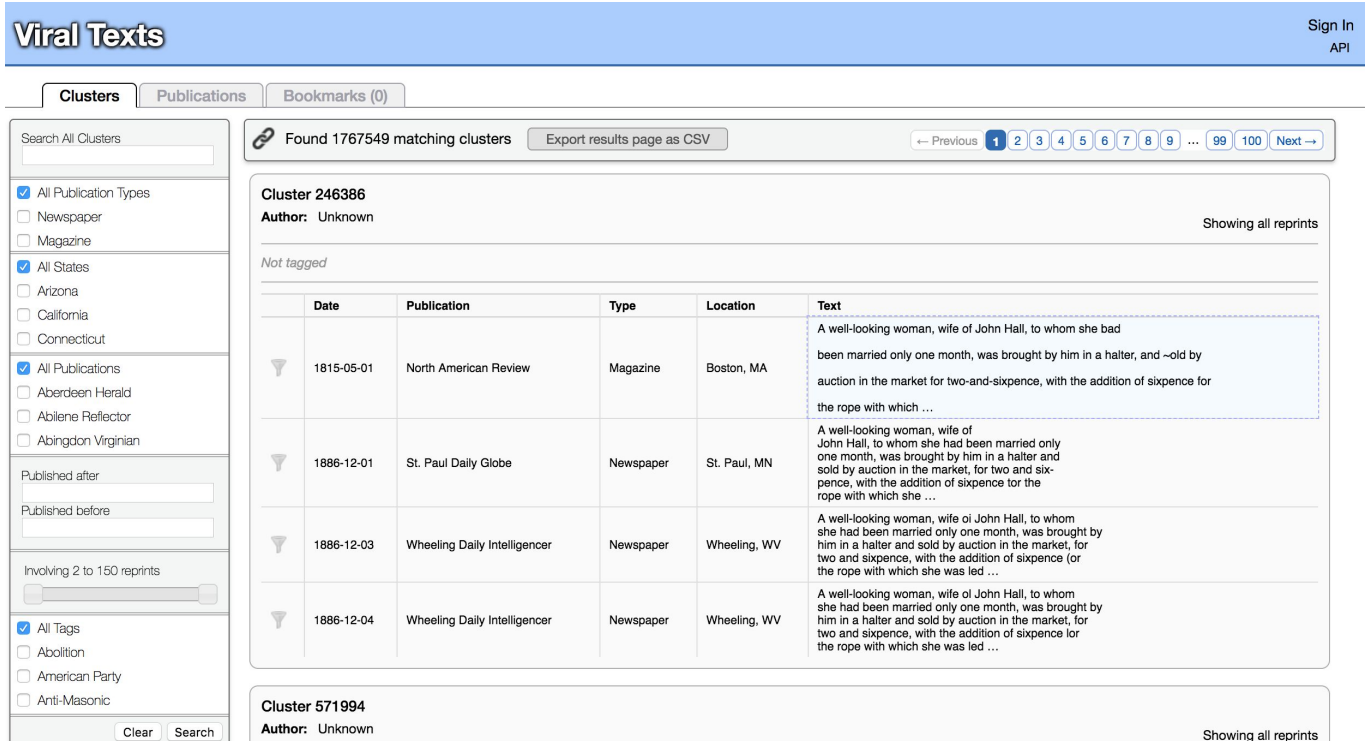
GDL 26/05/1900

un journal a prétendu que des documents relatifs à cette affaire avaient été détournés en vue d'un renouvellement de l'agitation 'dont nous avons tant souffert. Le fait est-il vrai ? Le gouvernement, s'il est vrai, a-til pris des mesures ou entend-il en prendre pour prévenir toute émotion nouvelle ? (Applaudissements.) Le général DE GALLIPPET monte à la tribune au milieu d'une grande attention et dit : L'autre jour, à la Chambre, répondant à M. Alphonse Humbert, j'ai dit que je ne connaissais pas le » documents dont on a parlé, que ces documents n'existaient pas. A ce moment j'avais le droit de tenir un paroil langage. Il était strictement conforme à la vérité. **J'ai le regret de dire aujourd'hui que je me suis trompé. Le lendemain du jour où je m'exprimais ainsi, j'avais un entretien avec le chef d'étatmajor général, et, là, j'avais la douleur d'apprendre non seulement que les documents existaient, mais qu'ils avaient été divulgués par un officier du ministère de la guerre.** (Vive

JDG 26/05/1900

d'un journal est exact, prétendant que des documents relatifs à /' « Affaire » ont été détournés en vue du renouvellement de cette affaire ? — t e général do Galliffet répond qu'il a lo regret de devoir dire qu'il s'est trompé mardi en disant à la Chambre que ces documents n'existaient pas. **Il ignorait alors leur existence, mais le lendemain il eut la douleur d'apprendre, dans un entretien avec le chef de l'état-major, non seulement que ces documents existaient, mais qu'ils avaient été divulgués par un officier du ministère de la guerre.** L'officier

Viral texts explorer

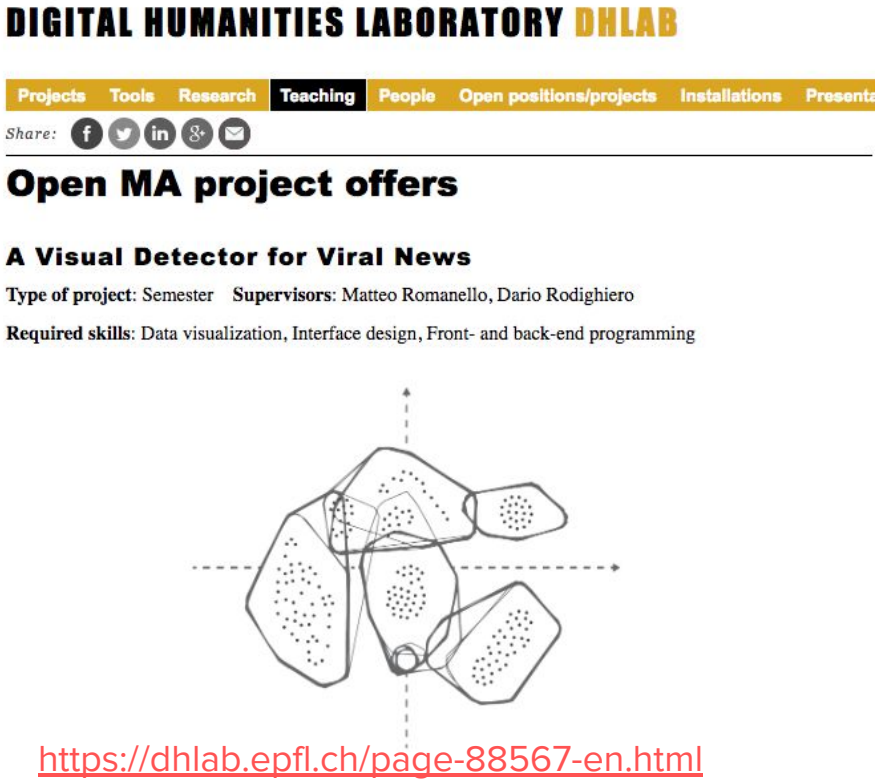


http://viraltexts.northeastern.edu/clusters

Visualization of Text Re-use

Display of text-reuse clusters along three dimensions:

- ***Size***: the number of passages in the cluster (min. 2, max. up to 20)
- ***Time***: the time span covered by the cluster as virals news usually cover a larger time span
- ***Lexical overlap***: the extent to which passages in the same cluster share the same set of tokens



https://dhlab.epfl.ch/page-88567-en.html

Interface co-design

code for the SWOT

mponent
scription

S.1

search autocomplete: based on input, suggests words, date ranges, named entities or article categories

red text: uncompleted features

S.2

timeline
n. results per year.
This acts as
"date range" filter

S.3

metadata filters / quick roundup
how many search results per newspapers titles, languages, page content tags, format tags

explore | search | Alba Rorwacher, Researcher | EN

YOUR SEARCH | PAST SEARCHES

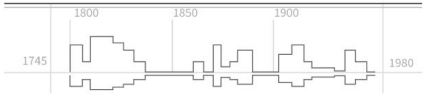
FULL TEXT SEARCH — REFINE ...

... type a text or a date or to start ...

PUBLISHED IN (DATE RANGE)

TIMELINE OF

N. ARTICLES | N. ISSUES



select a date range to show articles published

NEWSPAPER TITLES

53625 **La Gazette de Lausanne**
quotidien suisse de langue française édité à Lausanne

43864 **Le Journal de Gèneve**
quotidien suisse qui a paru du 1826 au 28 février 1998

MORE ...

ARTICLE TYPE

112 **partisans**
50 **satirique**

MORE ...

TOP NAMED ENTITIES

103 **Napoleon**
50 **Zurich, Suisse**
location

MORE ...

LANGUAGE

112 **French**

MORE ...

GROUP RESULTS BY: ISSUE | PAGE | ARTICLES | SENTENCES — ORDER BY: RELEVANCE — DISPLAY RESULTS: AS LIST | AS TILES

SEARCH SUMMARY

Find **53625987** articles in our collection.

SAVE SEARCH — COMPARE ...

1 of 112

Les recherches d'une science jeune: CELLE DE L'INFORMATION

INTERNATIONAL NEWS | TAG AS ...

Gazette de Lausanne, Saturday, May 8, 1954, p. 15
Les recherches d'une science jeune CELLE La presse , son rôle , ses possibilités JEAN-PIERRE AGUET ^^ .- ^ . ^ . ^^ H est à l'heure actuelle de multiples...

ADD TO FAVOURITES | ADD TO COLLECTION... | CONFERENCE REPORTS | EXPORT CITATIONS ... | DOWNLOAD AS ...

2 of 112

Les USA et l'AFN

INTERNATIONAL NEWS | COVER PAGE × | TAG AS ...

Gazette de Lausanne, Tuesday, November 19, 1957, p. 1
Les USA et l'AFN Un reportage de Charles-Henri Favrod en Tunisie A la terrasse de café où je suis assis , me défendant tant bien que mal contre l assaut...

ADD TO FAVOURITES | ADD TO COLLECTION... | GUERRE FROIDE × | EXPORT CITATIONS ... | DOWNLOAD AS ...

3 of 112

FRANCO PRÉFÈRE LE TÊTE-A-TETE AVEC L'AMERIQUE

INTERNATIONAL NEWS | TAG AS ...

Gazette de Lausanne, Friday, April 27, 1951, p. 1
FRANCO PRÉFÈRE LE TÊTE-A-TETE AVEC L'AMERIQUE SOUS LE PRINTEMPS D'ESPAGNE ATTENDANT UN TARDIF REPENTIR FRANCO-ANGLAIS ' ;•; / Par notre envoyé spécial Michel CLERC ' ! ! ,. ' ;...

ADD TO FAVOURITES | ADD TO COLLECTION... | EXPORT CITATIONS ... | DOWNLOAD AS ...

S.4

structured overview of your search query, if logged in search query can be saved and retrieved later

S.5

list of results by article, page navigation preview, issue

S.6

(if logged in) add result to collection (private tags)

S.7

(if logged in) tag article, page, issue (public tags)

search

Alba Rorwacher
Researcher

EN

GROUP RESULTS BY ISSUE PAGE ARTICLES SENTENCES

ORDER BY RELEVANCE

DISPLAY RESULTS AS LIST AS TILES

REFINE ...

1900

1980

es published

aise édité à

SEARCH SUMMARY

Find **53625987** articles in our collection.

SAVE SEARCH

COMPARE ...

1 of 112

Les recherches d'une science jeune: CELLE DE L'INFORMATION

INTERNATIONAL NEWS TAG AS ...

Gazette de Lausanne, Saturday, May 8, 1954, p. 15

Les recherches d'une science jeune CELLE La presse , son rôle , ses possibilités JEAN-PIERRE AGUET ^^ .- ^ . ^ . ^^ H est à l'heure actuelle de multiples...

ADD TO FAVOURITES

ADD TO COLLECTION...

CONFERENCE REPORTS

EXPORT CITATIONS ...

DOWNLOAD AS ...

2 of 112

Les USA et l'AFN

INTERNATIONAL NEWS COVER PAGE × TAG AS ...

Gazette de Lausanne, Tuesday, November 19, 1957, p. 1

Les USA et l'AFN Un reportage de Charles-Henri Favrod en Tunisie A

ADD TO FAVOURITES

ADD TO COLLECTION...

GUERRE FROIDE ×

explore | search | Alba Rorwacher, Researcher | EN

YOUR SEARCH | PAST SEARCHES

GROUP RESULTS BY: ISSUE | PAGE | ARTICLES | SENTENCES — ORDER BY: RELEVANCE — DISPLAY RESULTS: AS LIST | AS TILES

S.8

include / exclude options are given for each text string

FULL TEXT SEARCH — REFINE ...

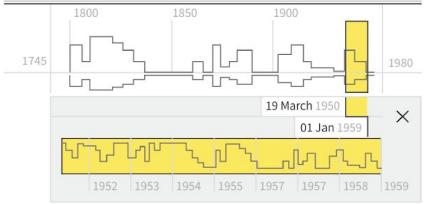
... type a text or a date or to start ...

CONTAINS | EXACTLY | guerre froide

CONTAINS | ... | europe

PUBLISHED IN (DATE RANGE)

TIMELINE OF N. ARTICLES — RESET



S.9

zooming on the selected date range

ADD RANGE ...

IN NEWSPAPER TITLES — RESET

112 **La Gazette de Lausanne**
quotidien suisse de langue française édité à Lausanne

TYPES OF ARTICLES

112 **international news**
50 **politics**

TOP NAMED ENTITIES

103 **Napoleon**
50 **Zurich, Suisse**
location

LANGUAGE OF ARTICLES

112 **French**

SEARCH SUMMARY

Found **112** articles matching exactly **"guerre froide"** AND **europe** published in **La Gazette de Lausanne** between **Saturday, 19 March 1950** and **Friday, 01 Jan 1959**

SAVE SEARCH | COMPARE ...

1 of 112

Les recherches d'une science jeune: CELLE DE L'INFORMATION

INTERNATIONAL NEWS | TAG AS ...

ADD TO FAVOURITES | ADD TO COLLECTION...

Gazette de Lausanne, Saturday, May 8, 1954, p. 15
Les recherches d'une science jeune CELLE La presse , son rôl possibilités JEAN-PIERRE AGUET ^^ .- ^ . ^ . ^^ H est à l'heure de multiples...

> en préjugés sur . ses' , voisins et souvent en guerre « froide » -ou « chaude » . « UNE SEMAINE DANS LE MOND

> , est la presse nationale de plu- sieurs pays parmi lesqu les Etats- Unis , huit pays d'Europe occidentale

S.10

add article to a collection

ADD TO COLLECTION
filter or create collection
favourites
✓ conference reports
guerre froide
CREATE NEW | MANAGE MY COLLECTIONS

2 of 112

Les USA et l'AFN

INTERNATIONAL NEWS | COVER PAGE × | TAG AS ...

ADD TO FAVOURITES | ADD TO COLLECTION...

Gazette de Lausanne, Tuesday, November 19, 1957, p. 1
Les USA et l'AFN Un reportage de Charles-Henri Favrod en Tunisie A la terrasse de café où je suis assis , me défendant tant bien que mal contre l assaut...

> - nomie . Tant qu'a duré la guerre froide , Washington n'a pas eu d'autre souri que de combattre

> en chef des forces du Centre-Europe , « chef atlan- tique sûr » , s'employa à accréditer cet- te version

GUERRE FROIDE ×

S.11

display list of matching text segments in the article

EXPORT CITATIONS ... | DOWNLOAD AS ...

S.12

what you would like to export? Individual articles can be downloaded (text, pdf, images)

3 of 112

FRANCO PRÉFÈRE LE TÊTE-A-TETE AVEC L'AMERIQUE

INTERNATIONAL NEWS | TAG AS ...

ADD TO FAVOURITES | ADD TO COLLECTION...

Gazette de Lausanne, Friday, April 27, 1951, p. 1
FRANCO PRÉFÈRE LE TÊTE-A-TETE AVEC L'AMERIQUE SOUS LE PRINTEMPS D'ESPAGNE ATTENDANT UN TARDIF REPENTIR FRANCO-ANGLAIS ' ;•; / Par notre envoyé spécial Michel CLERC ' ! ! ,. ' ;...

EXPORT CITATIONS ... | DOWNLOAD AS ...

explore | search | Alba Rorwacher, Researcher | EN

YOUR SEARCH | PAST SEARCHES

GROUP RESULTS BY: ISSUE | PAGE | ARTICLES | SENTENCES

ORDER BY: RELEVANCE

DISPLAY RESULTS: AS LIST | AS TILES

L TEXT SEARCH — REFINE ...

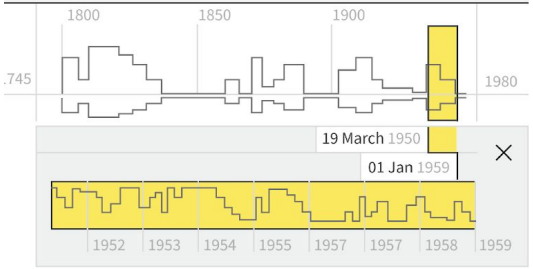
. type a text or a date or to start ...

ONTAINS | EXACTLY | guerre froide

ONTAINS | ... | europe

LISHED IN (DATE RANGE)

ELINE OF N. ARTICLES — RESET



ADD RANGE ...

EWSPAPER TITLES — RESET

112 **La Gazette de Lausanne**
quotidien suisse de langue française édité à Lausanne

SEARCH SUMMARY

Found **112** articles matching exactly **"guerre froide"** AND **europe** published in **La Gazette de Lausanne** between **Saturday, 19 March 1950** and **Friday, 01 Jan 1959**

SAVE SEARCH | COMPARE ...

1 of 112

Les recherches d'une science jeune: CELLE DE L'INFORMATION

INTERNATIONAL NEWS | TAG AS ...

ADD TO FAVOURITES | ADD TO COLLECTION...

Gazette de Lausanne, Saturday, May 8, 1954, p. 15
Les recherches d'une science jeune CELLE La presse , son rôl possibilités JEAN-PIERRE AGUET ^^ .- ^ . ^ . ^^ H est à l'heure de multiples...

> en préjugés sur . ses' , voisins et souvent en guerre « froide » ou « chaude » . « UNE SEMAINE DANS LE MOND

> , est la presse nationale de plu- sieurs pays parmi lesqu les Etats- Unis , huit pays d'Europe occidentale

ADD TO COLLECTION
filter or create collection
favourites
✓ conference reports
guerre froide
CREATE NEW | MANAGE MY COLLECTIONS

2 of 112

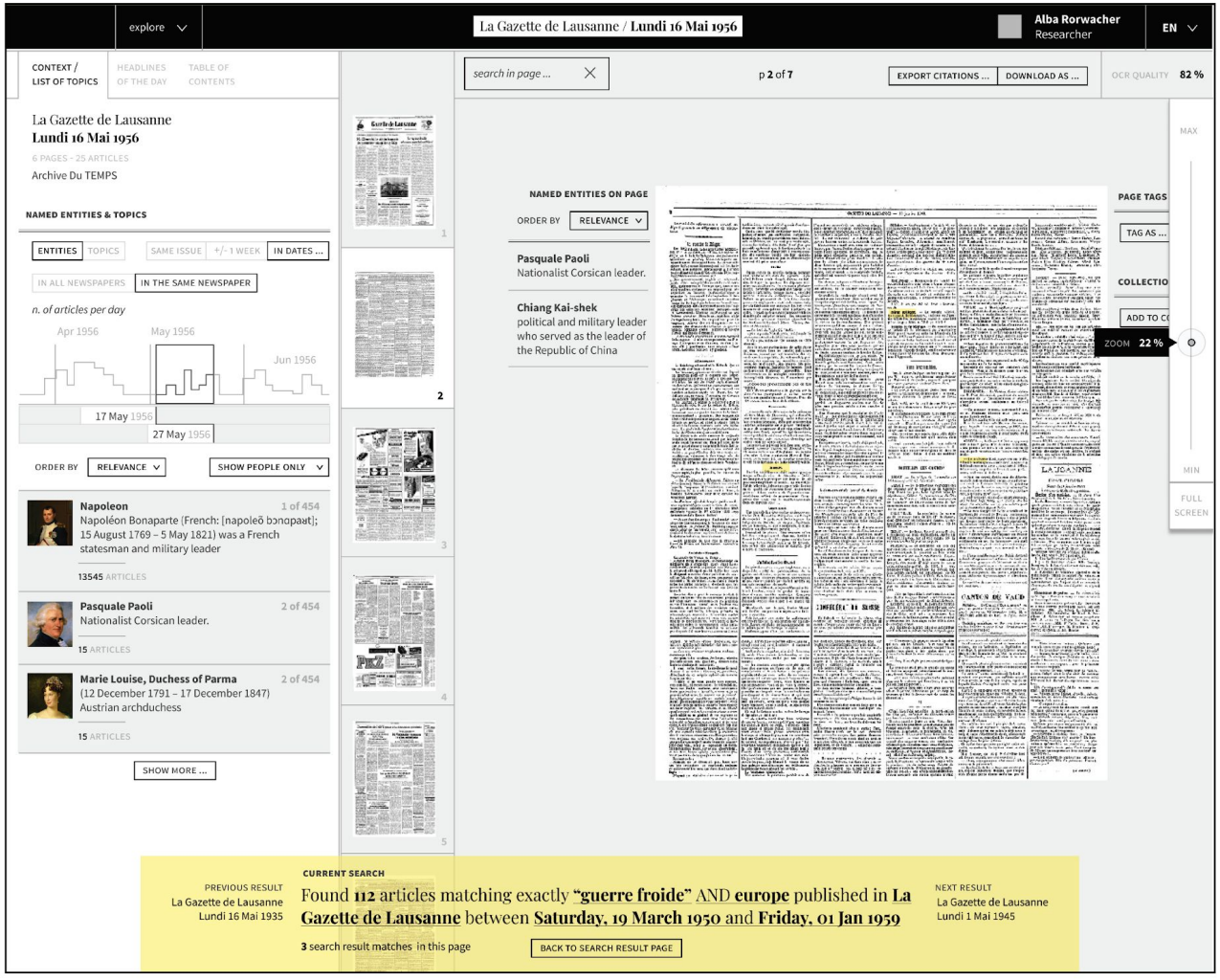
Les USA et l'AFN

INTERNATIONAL NEWS | COVER PAGE × | TAG AS ...

ADD TO FAVOURITES | ADD TO COLLECTION... | GUERRE FROIDE ×

Gazette de Lausanne, Tuesday, November 19, 1957, p. 1
Les USA et l'AFN Un reportage de Charles-Henri Favrod en Tunisie A la terrasse de café où je suis assis , me défendant tant bien que mal contre l assaut...

> - nomie . Tant qu'a duré la guerre froide , Washington n'a pas eu d'autre souri que de combattre



R.1

issue context / switch to table of contents

R.2

timeline of n. articles per day, focus on the date of publication of this issue. This filters related entities in R.3

R.3

entities or topics mentioned in the same issue / the same day (according to user's choice)

R.4

browse through search results (if you reached this page from the search page)

R.5

zoom on newspaper page

R.6

metadata: see tags, entities and collection as "marginalia"

R.7

highlight search results for this issue in the thumbnails

R.2

timeline of n. articles per day, focus on the date of publication of this issue. This filters related entities in R.3

R.3

entities or topics mentioned in the same issue / the same day (according to user's choice)

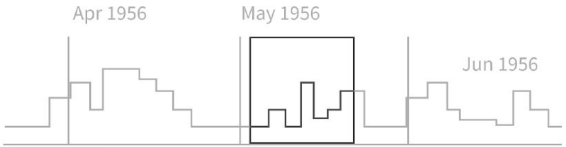
R.4

NAMED ENTITIES & TOPICS

ENTITIES | TOPICS | SAME ISSUE | +/- 1 WEEK | IN DATES ...

IN ALL NEWSPAPERS | IN THE SAME NEWSPAPER

n. of articles per day



17 May 1956

27 May 1956

ORDER BY RELEVANCE

SHOW PEOPLE ONLY

Napoleon 1 of 454
Napoléon Bonaparte (French: [napoleõ bɔnɑpaʁt]; 15 August 1769 – 5 May 1821) was a French statesman and military leader

13545 ARTICLES

Pasquale Paoli 2 of 454
Nationalist Corsican leader.

15 ARTICLES

Marie Louise, Duchess of Parma 2 of 454
(12 December 1791 – 17 December 1847)
Austrian archduchess

15 ARTICLES

SHOW MORE ...

1

2

3

4

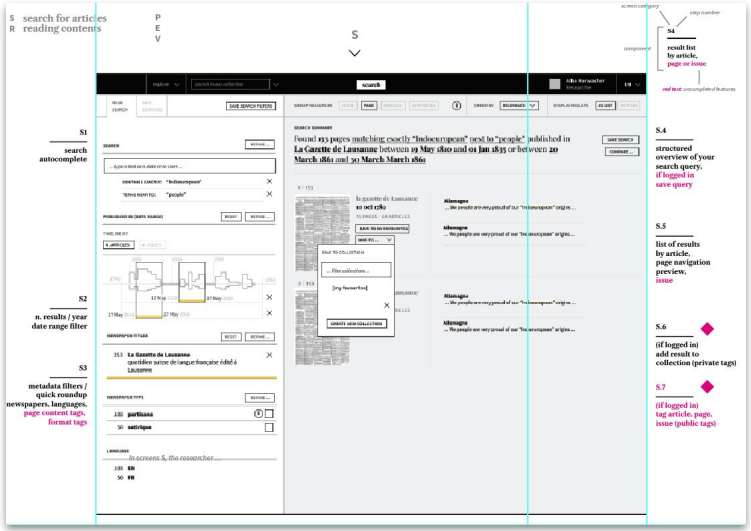
ORDER BY RELEVANCE

Pasquale Paoli
Nationalist Corsican leader.

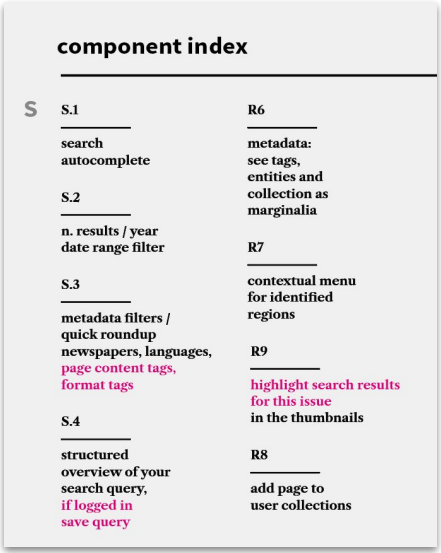
Chiang Kai-shek
political and military leader who served as the leader of the Republic of China

Co-design package

Booklet of interface features,
activity by activity



Component index,
by activity



component index

S

S.1
search autocomplete

S.2
n. results / year date range filter

S.3
metadata filters / quick roundup newspapers, languages, page content tags, format tags

S.4
structured overview of your search query, if logged in save query

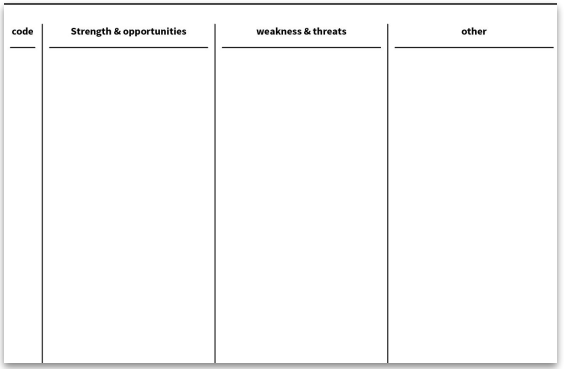
R6
metadata: see tags, entities and collection as marginalia

R7
contextual menu for identified regions

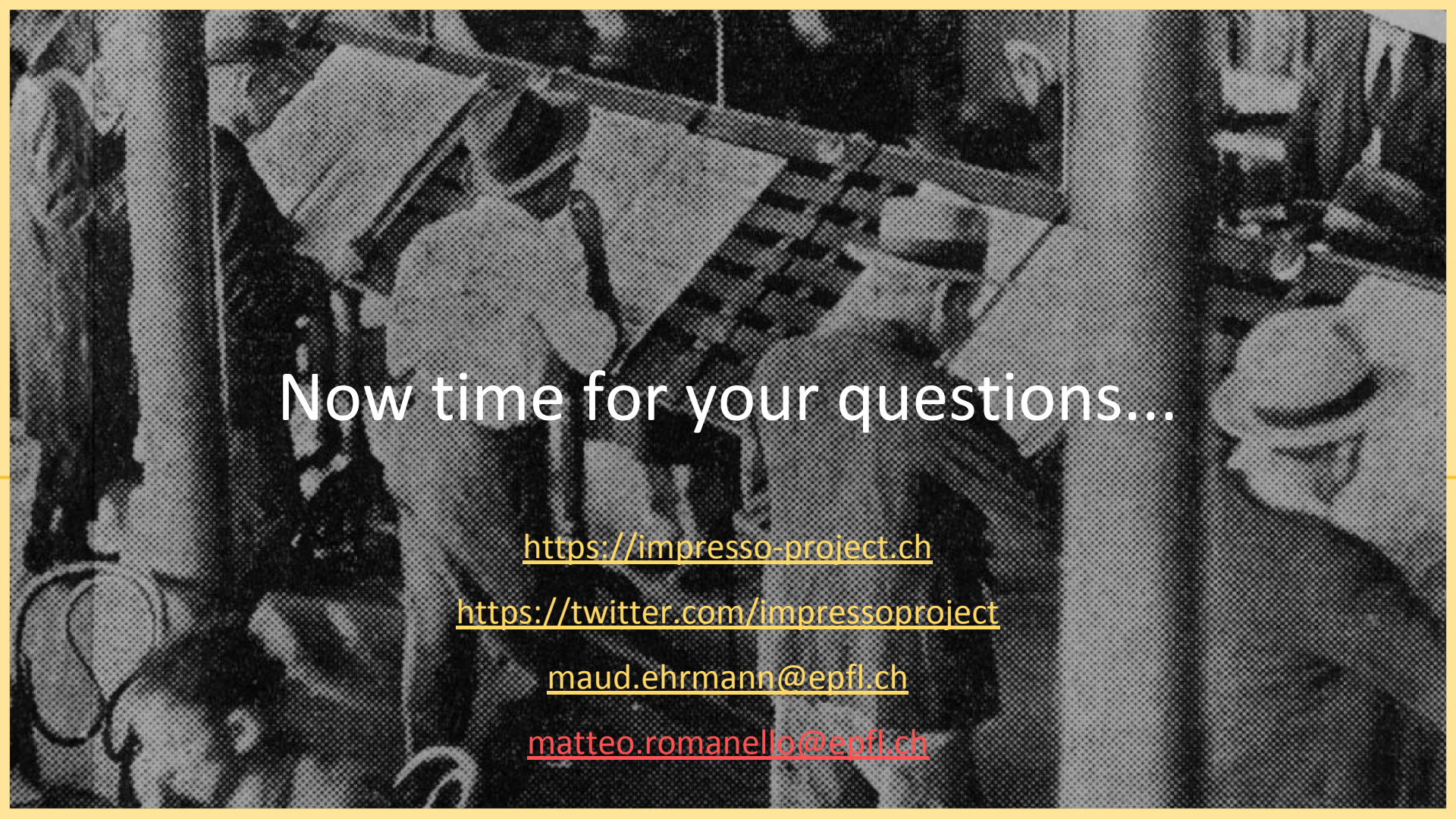
R9
highlight search results for this issue in the thumbnails

R8
add page to user collections

SWOT (**feedbacks**)



code	Strength & opportunities	weakness & threats	other



Now time for your questions...

https://impresso-project.ch

https://twitter.com/impressoproject

maud.ehrmann@epfl.ch

matteo.romanello@epfl.ch

Hands-on

Hands on

1. Getting the data (via ssh)
2. Data cleaning (command line)
3. First exploration (jupyter notebook)

COPY your .ssh/id_rsa.pub here : **goo.gl/YNrK3z**

GitHub repository: https://github.com/impresso/epfl-shs-class

Get started

```
cd ~/Documents

git clone https://github.com/impresso/epfl-shs-class.git

cd epfl-shs-class
```

Get data via sftp

```
cd epfl-shs-class

mkdir data

cd data/

sftp impresso@dhlabsrv4.epfl.ch

sftp> cd sharespace

sftp> ls

sftp> mget GDL-1900.jsonl.bz2

sftp> exit
```

Create conda environment

```
conda create --name shs-class-2018 python

conda activate shs-class-2018

# when done with the class `conda deactivate`

pip install jupyter spacy textacy dask[complete]

python -m spacy download fr_core_news_sm

pip list

jupyter notebook

# a browser window will open
```