



**HAL**  
open science

## **IIIF as a Service for Researchers**

Jean-Philippe Moreux

► **To cite this version:**

| Jean-Philippe Moreux. IIIF as a Service for Researchers. 3rd cycle. France. 2021. hal-04675821

**HAL Id: hal-04675821**

**<https://hal.science/hal-04675821v1>**

Submitted on 22 Aug 2024

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

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



Distributed under a Creative Commons Attribution 4.0 International License



# IIIF as a Service for Researchers

## The BnF Use Case

Jean-Philippe Moreux

Bibliothèque Nationale de France,  
Département de la Coopération

**Time Machine Academy on IIIF, September 2021**

<https://www.timemachine.eu/upcoming-time-machine-academy-on-iiif-international-image-interoperability-framework/>



# Outline

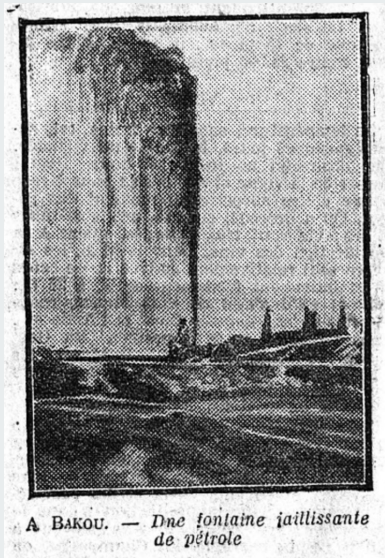
IIIF as an essential part of digitization and enrichment pipelines

IIIF for the dissemination of research results

IIIF for scientific digital mediation

---

# Pipelines



From IIIF as an interoperable format for image in heritage web portals to

...

a standard for document exchange in the digital humanities ecosystem

# The GallicaPix PoC

## Motivations

- Hybrid retrieval PoC (2017) on iconographic material (text, metadata, content-based image retrieval)
- Deep learning demonstrator for heritage collections: locally trained models, AI platforms and tools (commercial, open source)
- Internal and external users
- Use cases: information retrieval and digital humanities
- End to end IIIF support

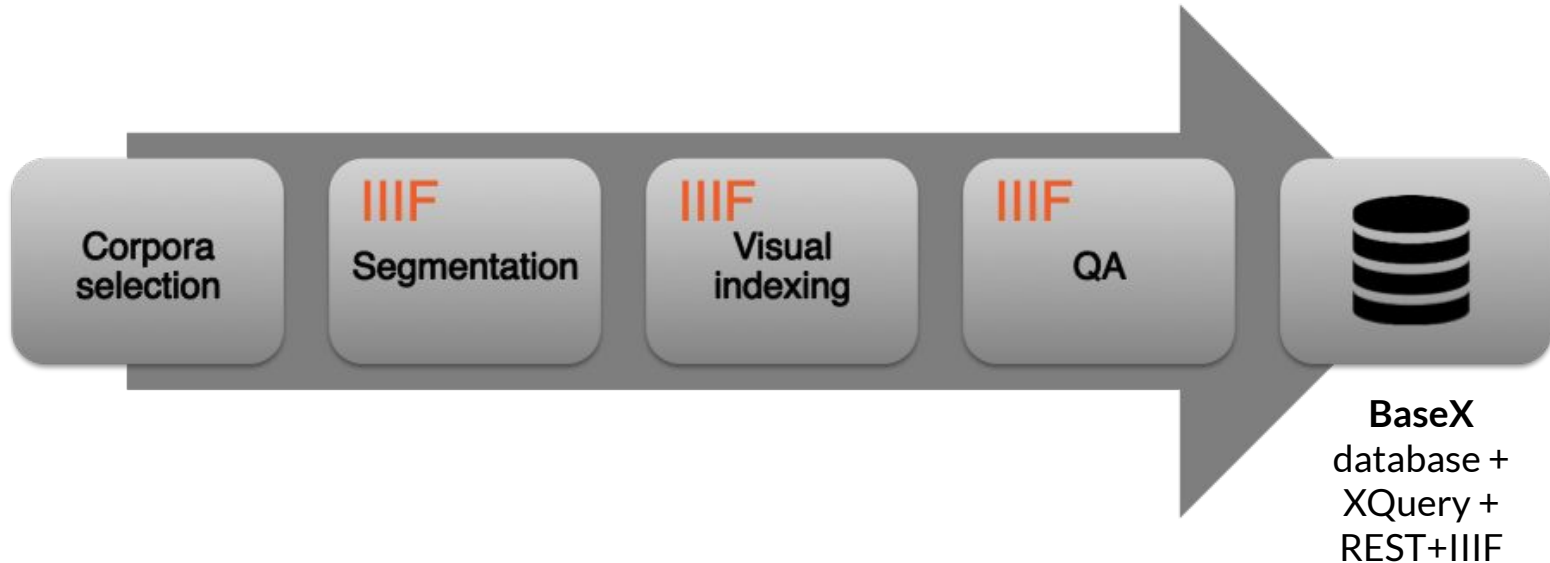
Project: <https://gallicapix.bnf.fr>





# The GallicaPix PoC

## The pipeline



# IIIF pipelines

## The classics




### Illustration Detection in IIIF Medieval Manuscripts using Deep Learning

Fouad AOUINTI – STIH Laboratory, Sorbonne Université  
Victoria EYHARABIDE – STIH Laboratory, Sorbonne Université  
Xavier FRESQUET – IReMus Laboratory, Sorbonne Université

2021 IIIF Annual Conference  
June 22-24, Online

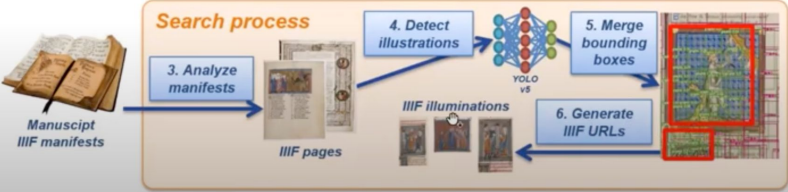
... and the Time Machine Box!



 **Our proposal**

Using the latest version YOLO:

1. **Preprocessing:** as the HBA images are not annotated with bounding boxes surrounding the illuminations, for each page, we randomly generated 10000 rectangles and labeled them
2. **Training:** we trained and validated YOLO on the HBA dataset
3. **Analyze manifests:** for each IIIF manuscript manifest, we extracted the URLs of the page's images
4. **Detect illustrations:** we evaluated each page's image on our model
5. **Merge bounding boxes:** we merged all the bounding boxes obtained to identify the illuminations
6. **Generate IIIF URLs:** for each illumination found on a page, we generate the corresponding IIIF URL



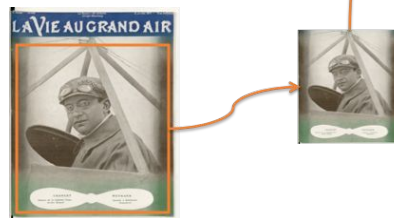
The diagram illustrates the search process. It starts with 'Manuscript IIIF manifests' leading to 'IIIF pages'. From 'IIIF pages', the process moves to '3. Analyze manifests', then to '4. Detect illustrations' which involves a 'YOLO v5' model. The output of '4. Detect illustrations' is 'IIIF illuminations'. From 'IIIF illuminations', the process moves to '5. Merge bounding boxes', then to '6. Generate IIIF URLs', and finally back to 'IIIF pages'.

# The GallicaPix PoC

## Advantages of using IIF in a R&D activity

- API facilitates the development of prototypes: Gallica APIs + Gallica IIF Image
- Interoperable standards like IIF allowed us to work on multiple collections: Europeana APIs + The Wellcome Collection IIF repository
- Instant access to images: no more files!
  - Digging in images with URLs
  - Training datasets, GT... are stored as metadata, not image files
  - Size of images needed for specific task can be tuned with a IIF parameter
  - Commercial APIs are directly feed with IIF URLs
  - Rendering of results (quality control) is very easy: rotating, sizing, cropping with URLs

```
curl -X POST -u "apikey:****" --form
"url=https://gallica.bnf.fr/iiif/ark:/12148/
bpt6k9604090x/f1/22,781,4334,4751/,700/0/
native.jpg" "https://gateway.watsonplatform.
net/visual-recognition/api/v3/classify?
version=2018-03-19"
```



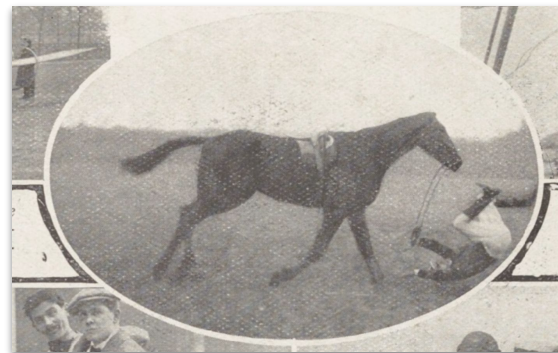


---

# The GallicaPix PoC

## Drawbacks

- Not suitable for very large corpora (pressure on IIIF servers): a local copy is recommended
- A timeout may occur when calling service APIs with an IIIF URL (if the IIIF server is overloaded).
- The processing speed is reduced compared to local processing on a local file.
- If your IIIF servers are not fast enough, your partners will see it and suffer!

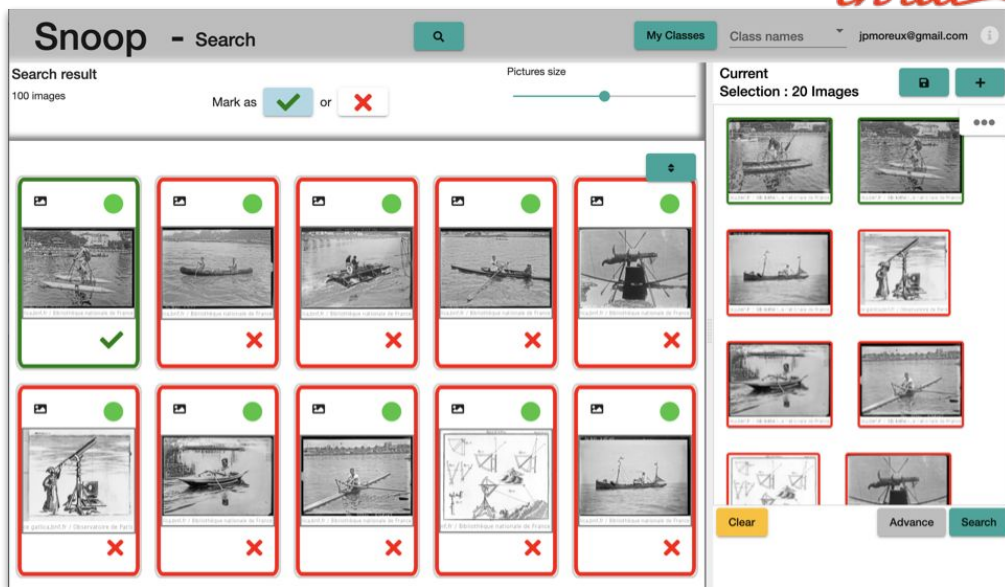


# External collaborations

## Advantages of using IIF in a R&D activity

- GallicaSnoop PoC, 2020 (similarity search):  
1.2M Gallica images downloaded and processed in 15 hours (one thread, waiting time between calls)
- No local storage of images

Project: <https://snoop.inria.fr/bnf/>



# External collaborations

## Advantages of using IIIF in a R&D activity

- REMDEM project, 2020 (identification of scribe on 50k music scores):
  - GT creation with an IIIF compliant annotation tool
  - Each annotation is accessed by the computer scientists as a IIIF URL

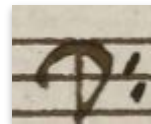
Annotation tool:

<https://www.dicen-idf.org/projet-recherche-opahh-iiif/>



<https://gallica.bnf.fr/iiif/ark:/12148/btv1b52502403/w/f2/1622,2755,145,118/pct:50/0/native.jpg>

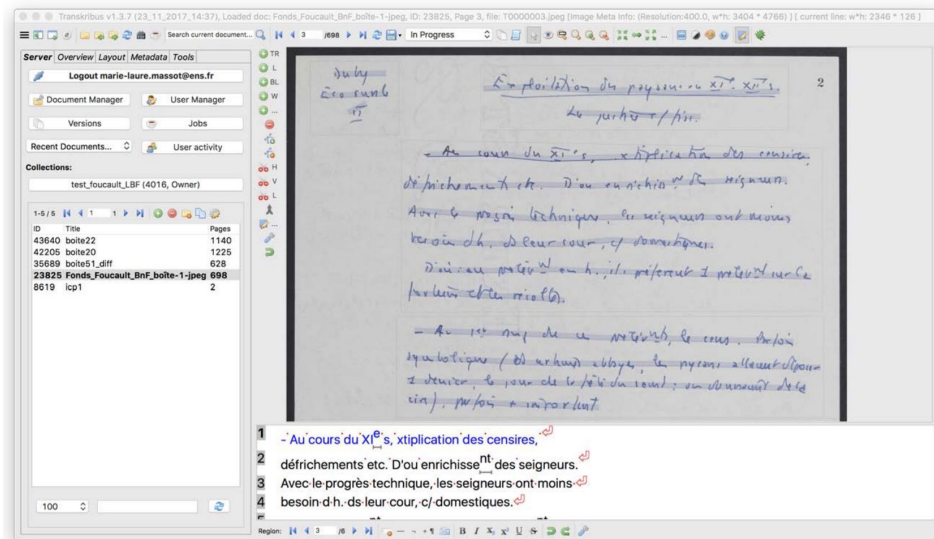
Tag: F clef



# Toolbox

## IIIF compliant production tools: input/output IIIF data

- HTR: Transkribus, eScriptorium
- Annotation: Mirador, [Annotate](#), [labellmg](#), Simple Annotation Server...
- Digital Repository: Archipelago
- ...

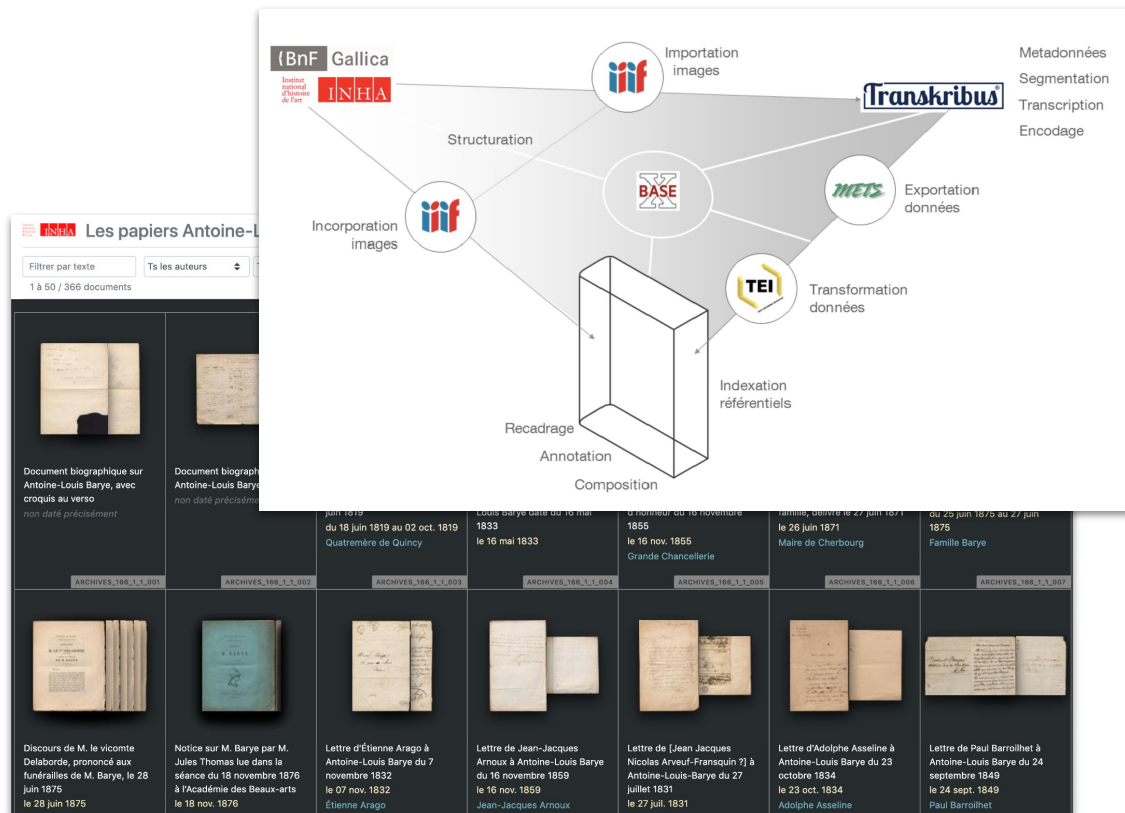


Project: ANR Michel Foucault reading notes (2018-2020)  
<https://odhn.ens.psl.eu/article/foucault-fiches-de-lecture-ffl>

# DH Pipeline

## Transkribus + BaseX

- **INHA project:** *Les Papiers Antoine-Louis Barye*
- From IIF repositories (Gallica, INHA) to web apps using Transkribus for transcription, TEI for data modeling and BaseX as a publishing tool
- At the end: a generic pipeline for DH transcription projects

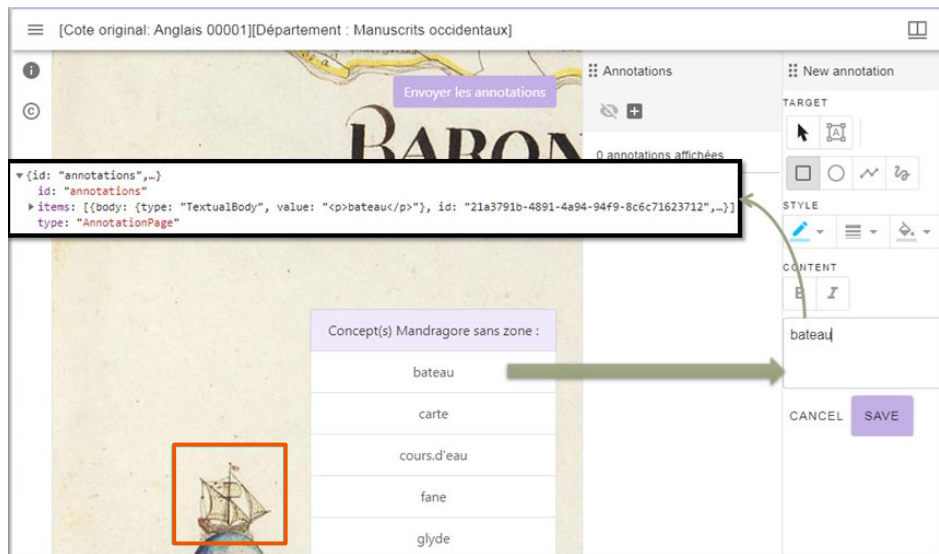


Project: <https://skylab.inha.fr/PENSE/LesPapiersBarye/>

# In-house workflow

## Mandragore-BnF

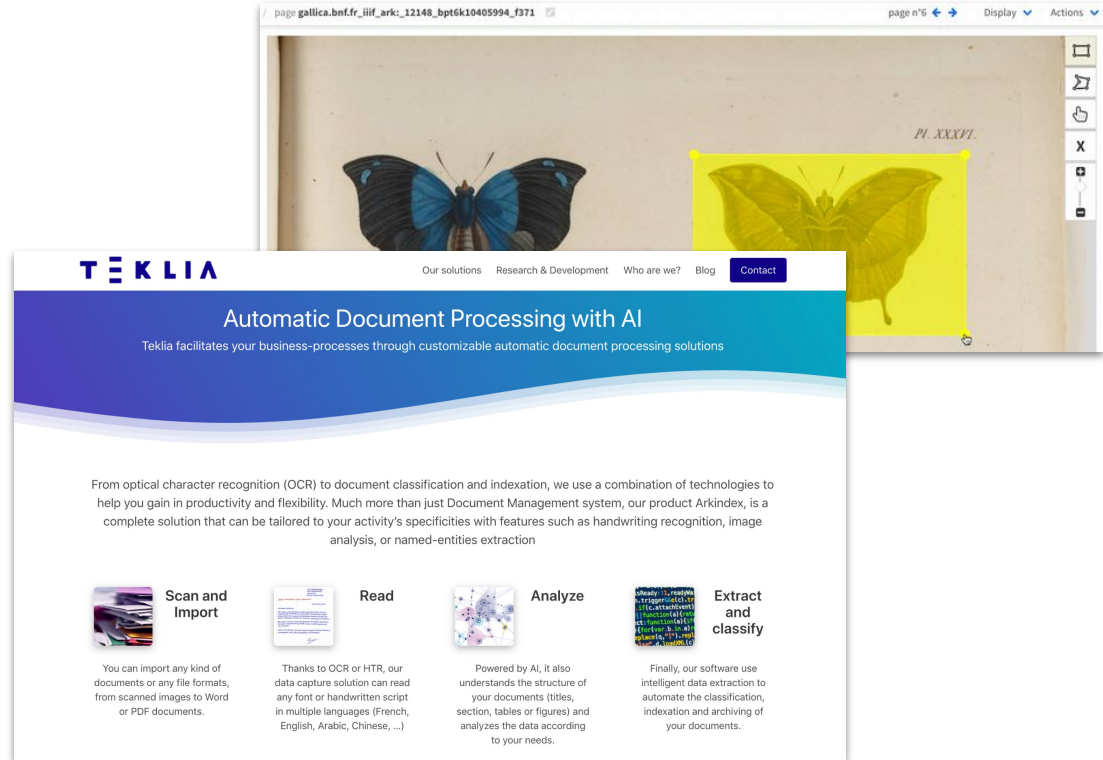
- **Mandragore v2 (2020-2022):** enlightened manuscripts database and its web portal
- Mirador 3 + Annotation plugin embedded in the production workflow
- Validation of the visual concept against the Mandragore ontology



Project: <http://mandragore.bnf.fr>

# Service Provider Heritage Workflow Teklia

- **Arkindex:** automatic document processing with AI
- Platform based on IIIF



The image displays two overlapping screenshots. The top screenshot shows a digital document viewer interface with a butterfly specimen on the left and a yellow rectangular overlay on the right. The URL in the browser is 'page.gallica.bnf.fr\_iiif\_ark:\_12148\_bpt6k10405994\_f371'. The bottom screenshot shows the Teklia website, which features a blue header with the company logo and navigation links. The main content area is titled 'Automatic Document Processing with AI' and describes the company's AI-powered solutions for document processing, including OCR, document classification, and indexing. The website also lists four key services: Scan and Import, Read, Analyze, and Extract and classify.

**T E K L I A** Our solutions Research & Development Who are we? Blog [Contact](#)

## Automatic Document Processing with AI

Teklia facilitates your business-processes through customizable automatic document processing solutions

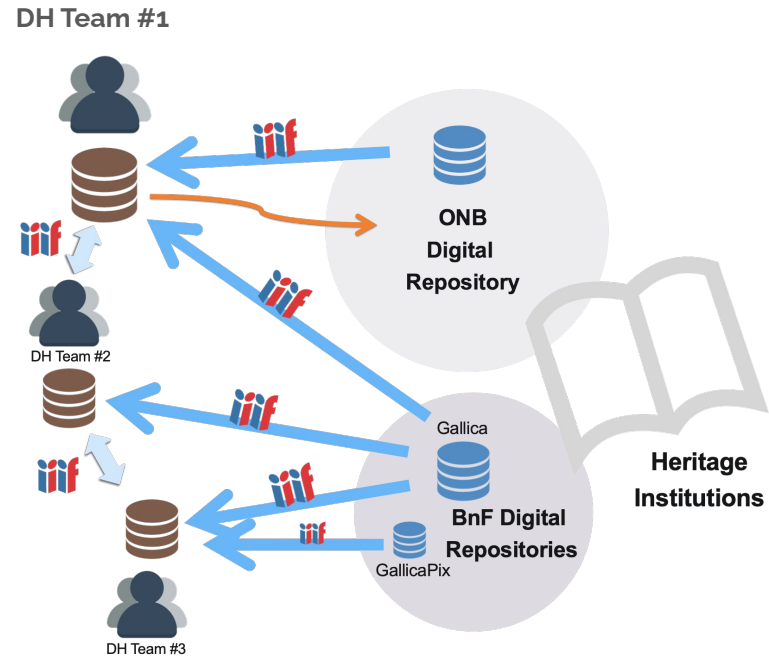
From optical character recognition (OCR) to document classification and indexing, we use a combination of technologies to help you gain in productivity and flexibility. Much more than just Document Management system, our product Arkindex, is a complete solution that can be tailored to your activity's specificities with features such as handwriting recognition, image analysis, or named-entities extraction

- Scan and Import**  
You can import any kind of documents or any file formats, from scanned images to Word or PDF documents.
- Read**  
Thanks to OCR or HTR, our data capture solution can read any font or handwritten script in multiple languages (French, English, Arabic, Chinese, ...)
- Analyze**  
Powered by AI, it also understands the structure of your documents (titles, section, tables or figures) and analyzes the data according to your needs.
- Extract and classify**  
Finally, our software use intelligent data extraction to automate the classification, indexing and archiving of your documents.

<https://tekli.com/>

# Dissemination of Research Results

- Heritage institutions can disseminate more metadata with IIF
- Researchers can share annotations and transcriptions
- Institutions may wish to benefit from the work of researchers →





---

# Dissemination: Datasets and corpora as resources



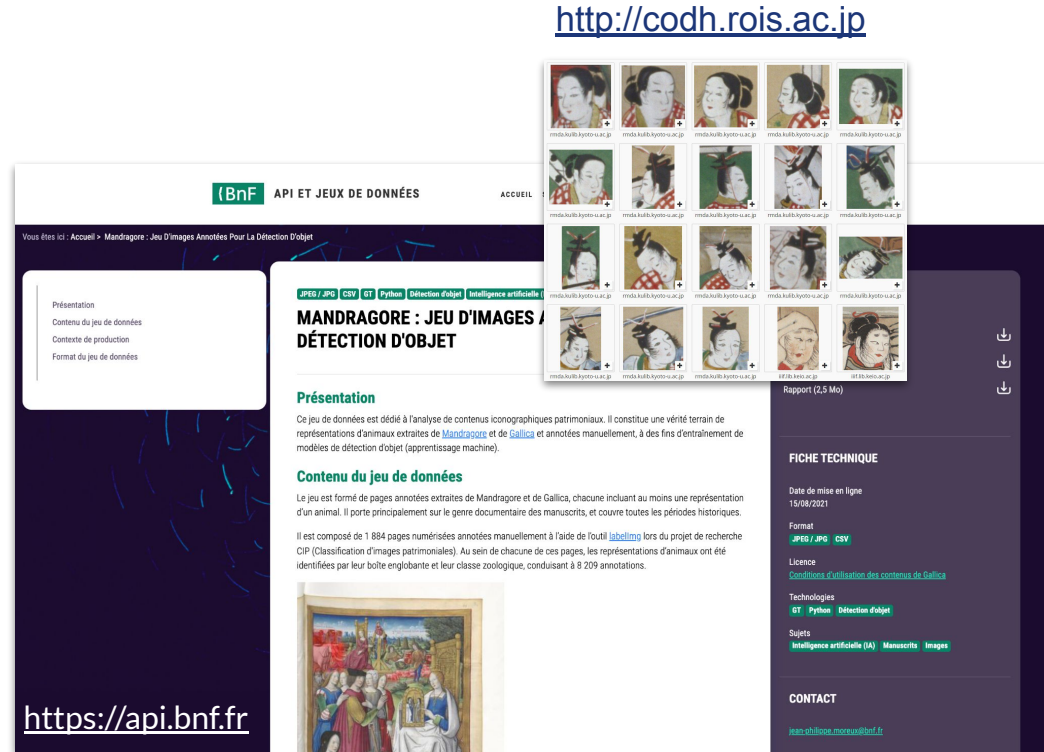
Source: gallica.bnf.fr / Bibliothèque nationale de France

# Gallica R&D

## Training datasets

- Training datasets, GT... are stored as metadata: the formats that CS are used to handling (COCO, Pascal VOC...) but also rich formats like IIIF
- IIIF Collections and IIIF Annotations are used to describe and display documents in datasets
- [IIIF Curation](#) could be used to describe and display parts of images (segmentation /object detection datasets). Content State API is promising for opening documents at the right place/zoom

<http://codh.rois.ac.jp>



The screenshot displays the BnF API and Data Games interface for the 'MANDRAGORE : JEU D'IMAGES / DÉTECTION D'OBJET' dataset. The page features a navigation menu with options like 'Présentation', 'Contenu du jeu de données', 'Contexte de production', and 'Format du jeu de données'. The main content area includes a grid of image thumbnails and a detailed description of the dataset. The sidebar on the right provides technical information, including the date of release (15/08/2021), the format (.JPG, .JPE, .CSV), and the technologies used (GT, Python, Détection d'objet, Intelligence artificielle).

**BnF API ET JEUX DE DONNÉES**

ACCUEIL

Vous êtes ici : Accueil > Mandragore : Jeu D'Images Annotées Pour La Détection D'objet

JPEG / JPE CSV GT Python Détection d'objet Intelligence artificielle

### MANDRAGORE : JEU D'IMAGES / DÉTECTION D'OBJET

#### Présentation

Ce jeu de données est dédié à l'analyse de contenus iconographiques patrimoniaux. Il constitue une vérité terrain de représentations d'animaux extraites de [Mandragore](#) et de [Gallica](#) et annotées manuellement, à des fins d'entraînement de modèles de détection d'objet (apprentissage machine).

#### Contenu du jeu de données

Le jeu est formé de pages annotées extraites de Mandragore et de Gallica, chacune incluant au moins une représentation d'un animal. Il porte principalement sur le genre documentaire des manuscrits, et couvre toutes les périodes historiques.

Il est composé de 1 884 pages numérisées annotées manuellement à l'aide de l'outil [labelimg](#) lors du projet de recherche CIP (Classification d'images patrimoniales). Au sein de chacune de ces pages, les représentations d'animaux ont été identifiées par leur boîte englobante et leur classe zoologique, conduisant à 8 209 annotations.

<https://api.bnf.fr>

Rapport (2,5 Mo)

#### FICHE TECHNIQUE

Date de mise en ligne  
15/08/2021

Format  
.JPG, .JPE, .CSV

Licence  
[Conditions d'utilisation des contenus de Gallica](#)

Technologies  
GT Python Détection d'objet

Sujets  
[Intelligence artificielle \(IA\)](#) [Manuscrits](#) [Images](#)

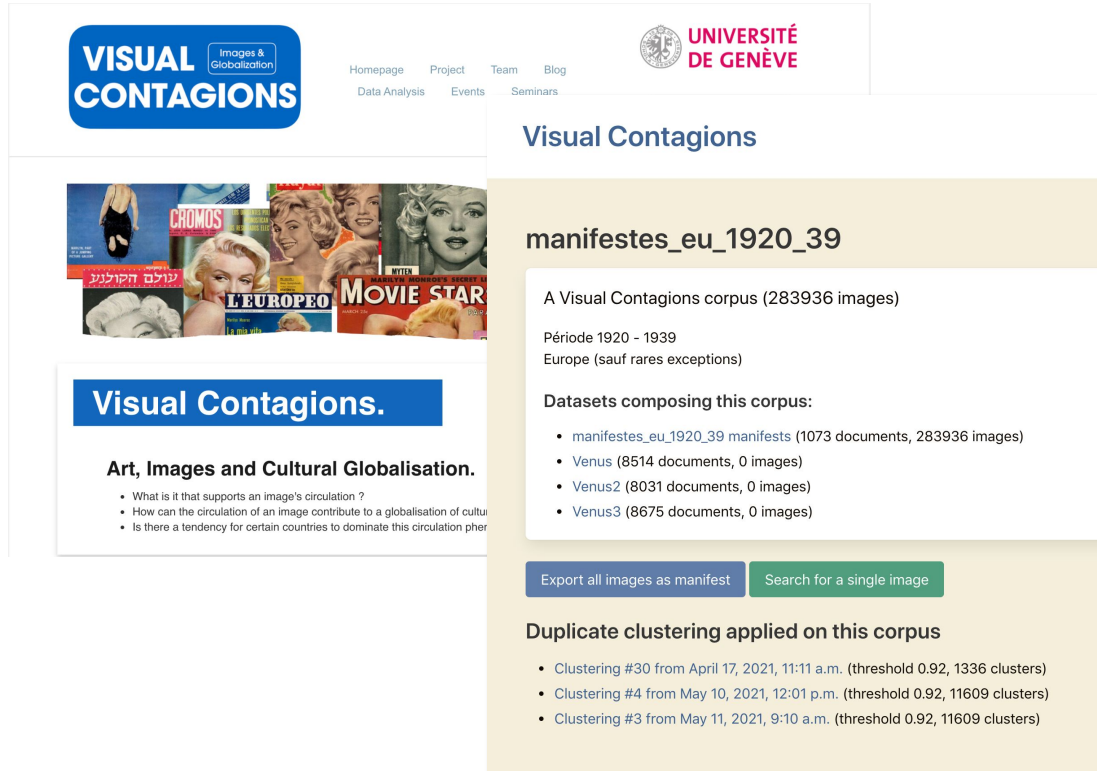
#### CONTACT

[rsz@codh.rois.ac.jp](mailto:rsz@codh.rois.ac.jp)

# Visual Contagions

## Corpora as lists of manifests

- Visual Contagion project: How circulation and dissemination of images take place on a global scale?
- Corpora's project as list of IIF manifests



The screenshot displays the Visual Contagions website. At the top left is the logo "VISUAL CONTAGIONS" with a sub-label "Images & Globalization". To the right is the navigation menu: "Homepage", "Project", "Team", "Blog", "Data Analysis", "Events", "Seminars". The top right corner features the "UNIVERSITÉ DE GENÈVE" logo. The main content area shows a collage of magazine covers including "CHROMOS", "LEUROPEO", and "MOVIE STAR". Below the collage is a blue box with the text "Visual Contagions." and "Art, Images and Cultural Globalisation." followed by three bullet points: "What is it that supports an image's circulation?", "How can the circulation of an image contribute to a globalisation of culture?", and "Is there a tendency for certain countries to dominate this circulation phenomenon?". On the right side, a panel titled "Visual Contagions" displays details for the corpus "manifestes\_eu\_1920\_39": "A Visual Contagions corpus (283936 images)", "Période 1920 - 1939", "Europe (sauf rares exceptions)", "Datasets composing this corpus:" followed by a list: "manifestes\_eu\_1920\_39 manifests (1073 documents, 283936 images)", "Venus (8514 documents, 0 images)", "Venus2 (8031 documents, 0 images)", and "Venus3 (8675 documents, 0 images)". At the bottom of this panel are two buttons: "Export all images as manifest" and "Search for a single image". Below the buttons, it states "Duplicate clustering applied on this corpus" and lists three clustering events: "Clustering #30 from April 17, 2021, 11:11 a.m. (threshold 0.92, 1336 clusters)", "Clustering #4 from May 10, 2021, 12:01 p.m. (threshold 0.92, 11609 clusters)", and "Clustering #3 from May 11, 2021, 9:10 a.m. (threshold 0.92, 11609 clusters)".

Project: <https://www.unige.ch/visualcontagions>

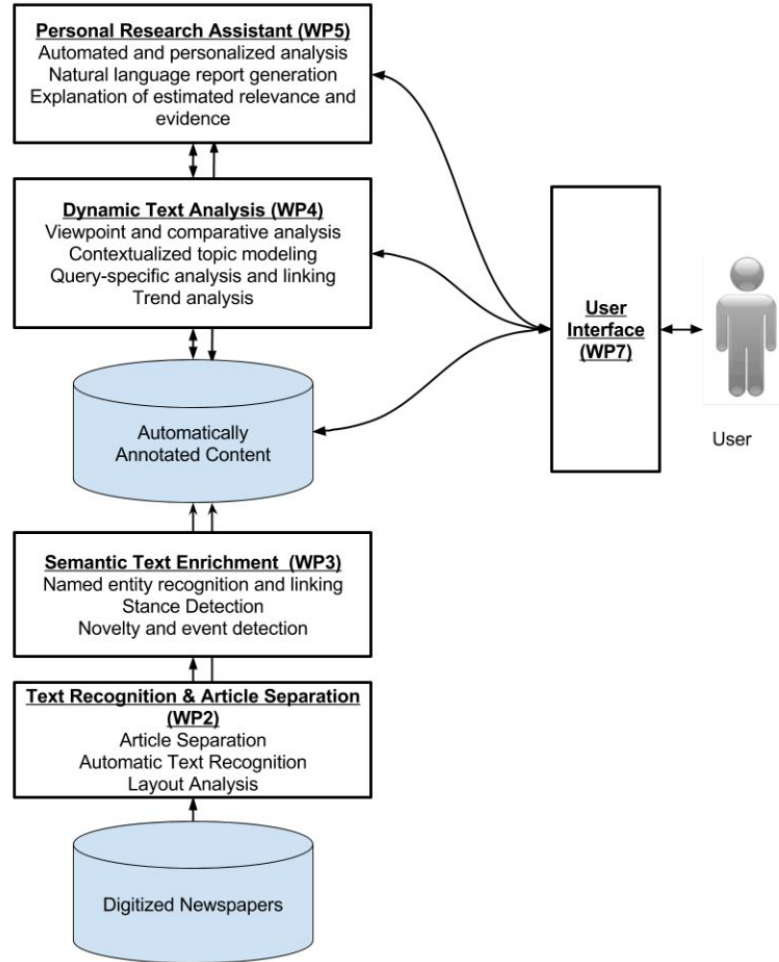
# NewsEye European project

## Motivations

- OCR, HTR, article separation, NER...
- Support qualitative and quantitative analysis of newspaper data for the digital humanities.

More than 180,000 issues from 20 newspapers,  
from 1850 to 1950.

Project website: <https://www.newseye.eu>  
Demonstrator: <https://platform.newseye.eu>



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299.

# NewsEye

## Corpora as IIIF Collections

National or thematic corpora used by the project are described as IIIF collections.

- Instant access to the data and reuse (parsing, scripting...)
- Browsing in any IIIF viewer
- IIIF Collections are well suited for periodicals

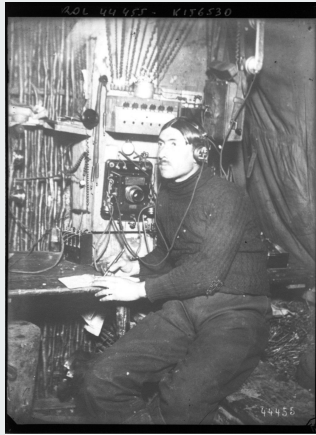
The screenshot displays the BnF IIIF viewer interface. On the left, a sidebar shows the collection details: 'Corpus de presse du projet NewsEye', 'Le Matin (1936)', and 'BnF Le Matin : derniers télégrammes de la nuit'. The main area shows a grid of 10 newspaper pages, with the first page highlighted. The interface includes navigation icons and a 'CLOSE' button at the bottom right.

Le Matin (1936)

NewsEye French dataset of periodicals (opened with Mirador 3)

---

# Dissemination: **IIIF Data in web apps**



# The GallicaPix PoC

## Visual indexing as IIF Annotations

- IIF Gallica manifest enriched with GallicaPix annotations opened in Mirador
- Periodicals exported as IIF Collections
- GallicaPix is now called from Gallica

*Vogue* magazine: tennis

Illustration 39-6 (6/12)  
technique : imp photoméca -  
fonction : repro/photo - genre : -  
annotations (10) : vêtement,  
personne (famille), personne, zoot  
suit, costume, vêtements, gris,  
personne, raquette de tennis, cravate

(BnF) Gallica

Recherche Avancée

TOUS NOS SÉLECTIONS PAR TYPES DE DOCUMENTS PAR THÉMATIQUES PAR AIRE GÉOGRAPHIQUES BLOG

PARUTION PAR DATE	
1920 1921	
Juin	Fév. Mars
Août	Sept.
1	2
3	4
5	6
7	8
9	10
11	12
13	14
15	16
17	18
19	20
21	22
23	24
25	26
27	28
29	30
31	

EN SAVOIR PLUS

SYNTHÈSE

LÉGENDES ET TABLE DES

VERSION TEXTE (OCR)

A DÉCOUVRIR

Gallica est une expérimentation d'indexation hybride de différentes collections de Gallica à contenu iconographique. Les illustrations y sont indexées selon les modalités habituelles de Gallica mais aussi selon des critères (type d'illustration, objets directs dans l'illustration, couleurs, etc.) obtenus par l'application de techniques d'intelligence artificielle. Obtenir plus d'information sur GallicaPix

https://gallica.bnf.fr/ark:/12148/ark:/61904/12148/117

# NewsEye project

## IIIF Data: in/out

- As a consumer:
  - Thumbnails in search results list
  - Viewer in documents show page
- As a producer:
  - Articles as annotations
  - Named entities as tags

The image displays the NewsEye project interface, which is a web-based platform for managing and viewing digital collections. The interface is divided into several sections:

- Search and Navigation:** At the top, there is a search bar with the text "Rechercher" and a "Rechercher" button. Below it, there are tabs for "Documents: Articles" and "Search: Exact".
- Dataset Membership:** A section titled "Dataset Membership" with a button "Create a dataset to add documents to." Below this, there is a "Working dataset:" dropdown menu showing "Mars (714 docs)".
- Issue:** A section titled "Issue" with a "Set relevancy towards Mars:" dropdown menu showing "Delete" and an "Apply" button.
- IIIF exports:** A section titled "IIIF exports" with a button "Open with Mirador" and a "Download IIIF Manifest" button.
- Document Viewer:** The main content area shows a document viewer for the issue "marie\_claire\_12148-bpt6k47011160 ma...". The viewer displays a page from the magazine "MARIE-CLAIRE" featuring a woman holding a bouquet of flowers. The page is annotated with a text box containing a French article snippet and a list of tags: "PER:Mme de Mortsau" and "PER:Félix Vandenesse".



# NewsEye project

## Users datasets as IIF Collection

- Export researchers datasets as IIF Collections:
  - Exports documents as IIF manifests
  - Two layers of annotations can be included in the manifests (All articles in the issue; Only articles that are part of the dataset)

The screenshot displays the NewsEye web interface. At the top, the 'NEWS EYE' logo is on the left, and navigation links for 'Search', 'Datasets', 'Saved searches', 'Experiments', 'FR', and 'Help' are on the right. The main content area is divided into two columns. The left column shows the 'roosevelt' dataset configuration, including a 'Sharing status' of 'Private', 'Export as' options (ZIP, CSV, Excel, JSON, IIF), and a list of linked entities: Locations (145), Persons (69), Organizations (35), and Human Productions (17). The right column displays a list of articles, with three visible: 61. uusi\_aura\_738520\_article\_292, 62. le\_matin\_12148-bpt6k567483h\_article\_511, and 63. le\_gaulois\_12148-bpt6k532010m\_article\_23. Each article entry includes its date and journal. A sidebar on the far right shows snippets of text from the articles, such as 'Rooseveltin, tammik. 24 pöyd.' and 'Requête à M. Roosevelt.'.

NewsEye user datasets

---

# Digital Mediation



Use cases:

- JADIS project
- NewsEye project: Exhibit
- Gallica: Storiies

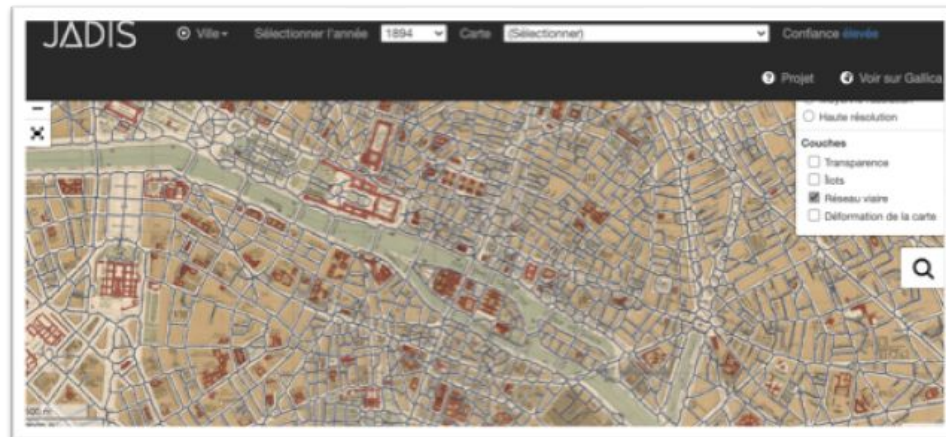
# Showcasing R&D results

## With IIIF

- **JADIS project, 2019 (heritage maps segmentation and georeferencing):**  
Maps of the city of Paris from BnF and BHVP collections
  - Instant access to the Gallica corpora
  - Basic but effective static web site (Github + IIIF images)

Project: <https://bnf.hypotheses.org/9676>

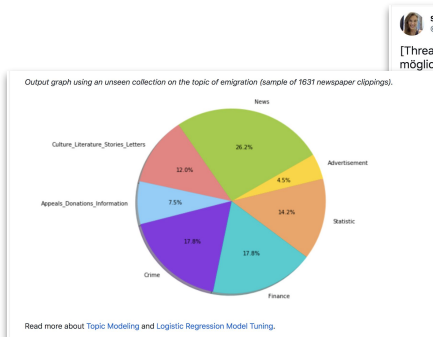
Rémi Petitpierre,  
Master of Science



# Storytelling for Digital Humanities

## How can researchers in humanities and social sciences showcase digital collections ?

- A challenge for any researcher who wishes to publicize research results to the general public
- Different ways to showcase: examples from the NewsEye project



**Sarah Oberbichler** @SOberbichler

[Thread] Todesstrafe für „Hamsterei“ und eine mögliche Antwort, warum wir Klopapier horten. Eine Zeitreise durch die Geschichte des irns mit Hilfe digitaler österreichischer en anno.onb.ac.at

6:48 AM · 15 mars 2020 · Twitter Web App

**Jani Marjanen** @janimarjanen

Our article "The expansion of isms, 1820–1917" is published in [@Journal\\_DMDH](#). It links to previous work on isms/ideology, but thanks to the [@COMHISgroup](#) and the [@NewsEyeEU](#) project we got the opportunity to do something more computationally oriented. [jdmndh.episciences.org/6728](#)

Traduire le Tweet

The expansion of isms, 1820–1917: Data-driven analysis... Words with the suffix -ism are reductionist terms that help us navigate complex social issues by using a simple...

8:38 AM · 21 déc. 2020 · Twitter Web App

### Case Studies

The Case Studies are led by the Digital Humanities Groups who work with the corpora of our three national libraries (Austria, Finland, France) testing existing tools and later those developed by our Computer Science groups. The idea is to delve into some of the current research issues, questions and topics of relevance to our project and other lay (and/or digital) historians, researchers, librarians, students amongst others.

#### Case Study 1: Migration

With the beginning of industrialisation in the eighteenth and nineteenth centuries, intercontinental and internal European migration patterns changed dramatically. The nineteenth century was often even called the 'age of migration' or the century of the 'great drift'. For the first time, we can see a migration industry that specialised in the organisation and processing of mass migration. While the rise in awareness and the subsequent increase in migration studies in all humanities fields has led to important scholarship also in regard to the nineteenth and early twentieth century, the discourses still offer a wide variety of research opportunities using newspapers in a comparative approach. The subtopics chosen for the case study on migration are 'Return migration', 'The business of emigration' and 'Negotiating asylum, aid and accommodation', as these topics all offer various opportunities and challenges.

### Blog

Blog posts are written by project team members. Topics range from conferences we attend, meetings on current efforts of relevance, internal project findings and news and more technical content which can be found in our Digital Humanities Case studies or project related publications. Blog posts will mainly be posted in English but will from time to time feature in the language of the project team member's preference, since we are a multilingual project! Happy reading!

Page 1 of 2 123Next

#### The invention of Mother's Day: origins, media history and gift ideas

Nepha Ousef (University of Bourgogne)

#### Curfew and inflammatory media coverage: spotlight on an exceptional measure

Nepha Ousef (University of Bourgogne)

#### Bringing together what belongs together: Thematic grouping of newspaper clippings using LDA and JSD

Sarah Oberbichler and Eva Plascher (University of Innsbruck)

#### Newspapers as „social“ media: Crowdsourcing and user-generated content in historical perspective

Benedek Kapteer (University of Innsbruck)

#### Of „difficult“ and „modern“ times. The development of journalism in historical newspapers

Benedek Kapteer (University of Innsbruck)

#### „An unsure Leser“: The interaction between newspaper and readership and the journalistic self-image

Benedek Kapteer (University of Innsbruck)

#### From the Spanish flu to Covid-19, the remedies claiming to work miraculously

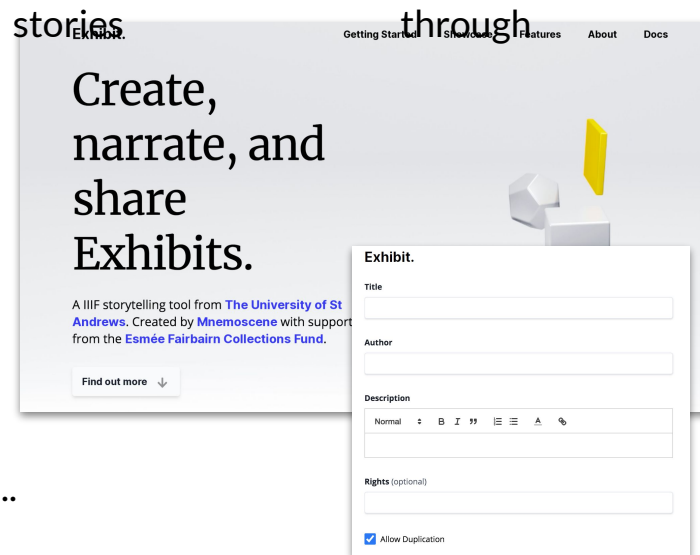
Nepha Ousef (University Paul-Valéry Montpellier)

# Storytelling for Digital Humanities

## Exhibit: a IIIF storytelling tool for an interactive experience

- Shareable, image-based annotation and deep zoom
- A tool developed by [Mnemoscene](#) with support from the [Esmée Fairbairn Collections Fund](#)
- A project initiated during the Covid-19 pandemic to meet the needs of The University of St. Andrews in online and on-campus teaching
- Use of universal Viewer, an open source IIIF viewer

Other tools: [Storiies](#) (Cogapp), [Tesselle](#) (Médialab SciencesPo)...



# Storytelling for Digital Humanities

## An example of narration from a blogpost

<https://exhibit.so/exhibits/UBQVwAjXirDsniiWhI5>

- For Newseye, links between the media treatment of the Coronavirus crisis and, 100 years earlier, that of the “spanish flu”
- Narration from blogpost about remedies claiming to work miracles against the 1918 flu, available in French on Gallica blog.
- Issues from the daily (*Le Matin*, *Le Gaulois*) and weekly press (*Le Pêle-mêle*, *La Baïonnette*, *Le Régiment*)



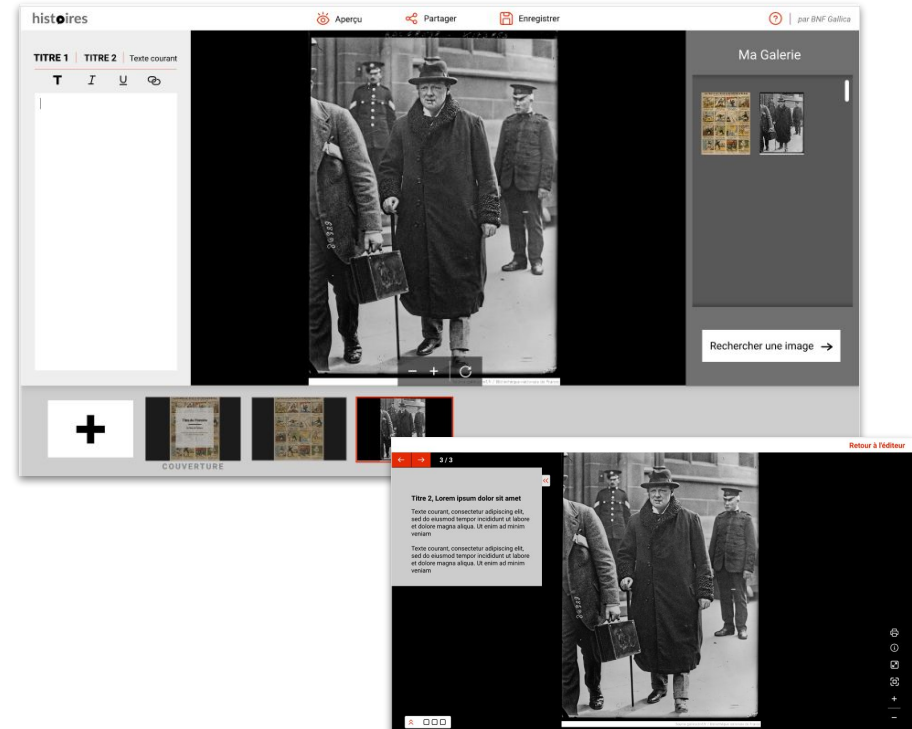
Nejma Omari, université de Montpellier

# Storytelling in Gallica

## From documents to stories

Version 2 of the BnF-Cogapp [Storiies](#) (2022):

- The editor can be launched any Gallica document
- Can mix multiple documents in a story
- Will save user's stories in user's account



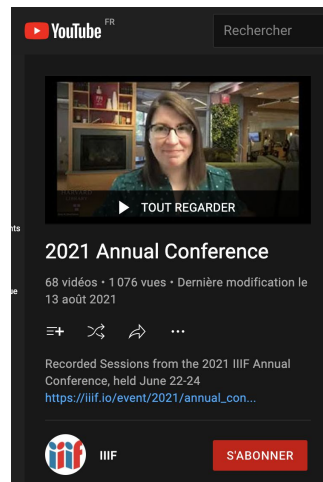


# Thanks!

jean-philippe.moreux@bnf.fr  
henry.huguet@bnf.fr

## Resources:

- <https://api.bnf.fr>
- <https://gallicapix.bnf.fr>
- <https://snoop.inria.fr/bnf/>
- <https://platform.newseye.eu>
- <https://github.com/altomotor/IIIF/>



Some of the projects described were presented at the IIIF 2021 Annual Conference