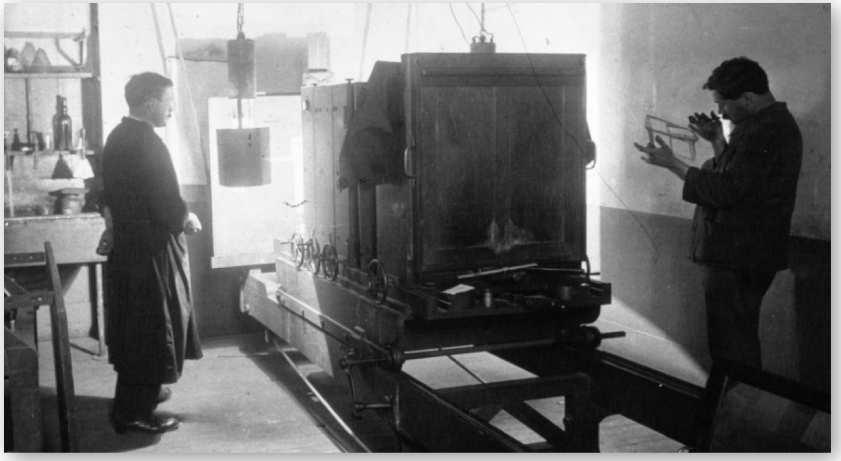
Fouille d’images dans les collections patrimoniales

Avec des techniques IA et le standard IIIF

Congrès ADBU, Bordeaux,
18 septembre 2019

Fouille d'images dans les collections patrimoniales

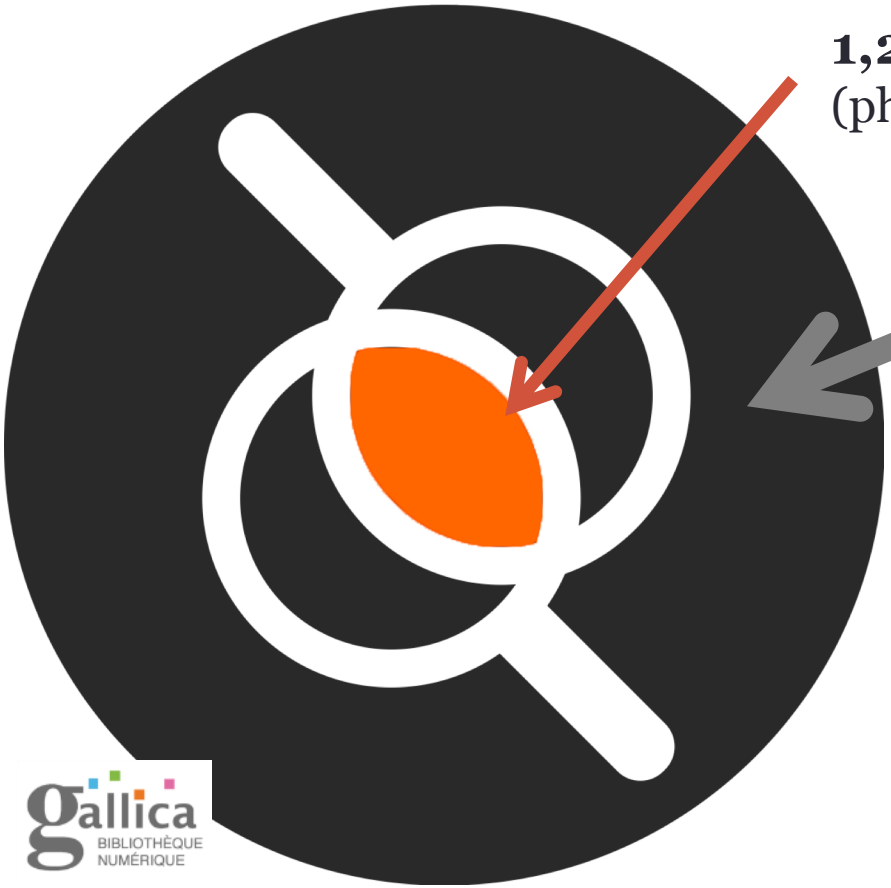
- Introduction : fouille d'images
- Preuve de concept GallicaPix et IIIF
- Expérimentation IA (deep learning) :
  - Classification des genres
  - Reconnaissance visuelle
- Cas d'usage pour la recherche d'information et les humanités numériques
- Conclusion



L'Auto, laboratoire photo (1914)

« Rechercher des images dans les collections patrimoniales ? »

Les bibliothèques comme réservoir d'images inexploitées



1,2 M images cataloguées
(photos, dessins, estampes…)

Large réservoir d'illustrations potentielles :

- **manuscrits**
- **documents imprimés**
- **archives du web**

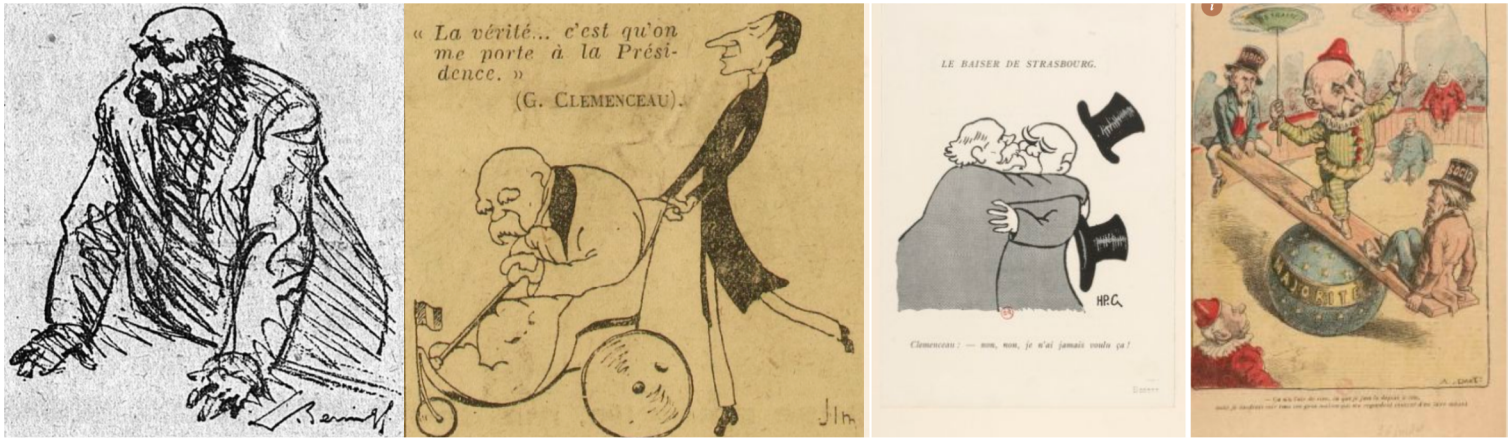
…

Centaines de millions ?

Nouveaux services

Pour la recherche d'information :

"Je cherche des caricatures de George Clemenceau dans toutes les collections."



Nouveaux services

Pour des usages de recherche (par ex. analyse quantitative)

"Je veux compter les visages de femme présents sur la page de une de *Paris-Match* entre 1949 et 1959"

L'Excelsior, 1910-1920

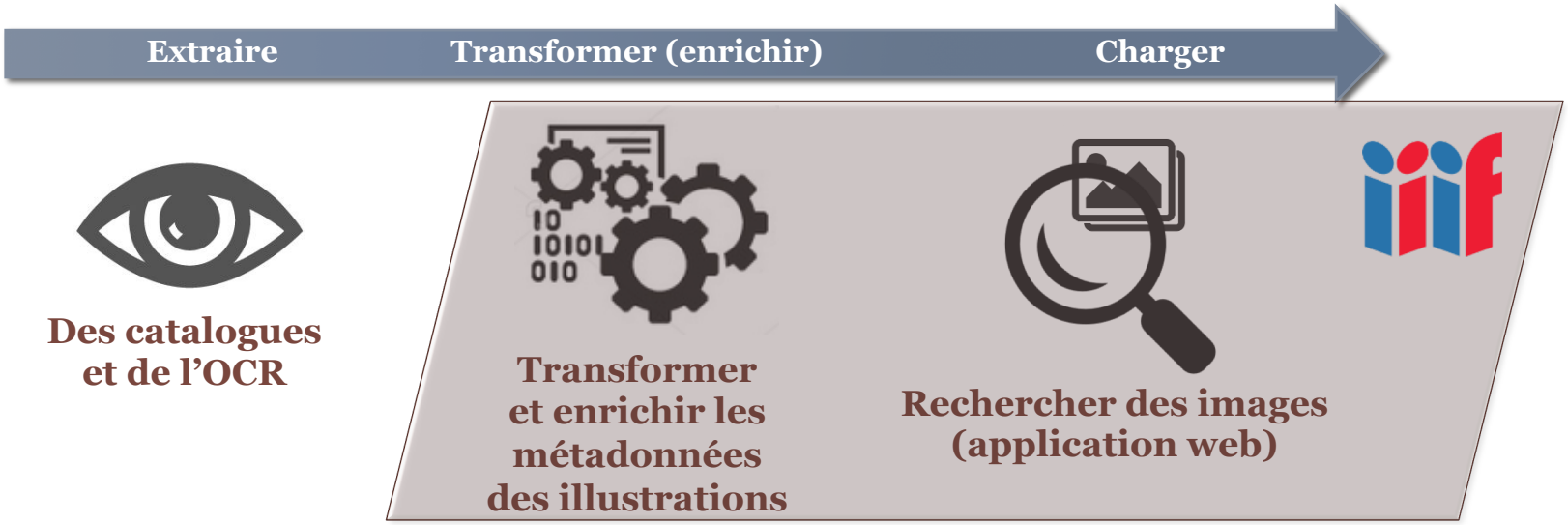
Preuve de concept : GallicaPix

Approche **ETL (Extract-Transform-Load)** pour la reconnaissance et la description automatiques des illustrations

Sur les collections Première Guerre mondiale de Gallica **:** photo, presse, magazines, posters, cartes… (1910-1920)

Avec des techniques d'apprentissage profond (**deep learning**)

IIIF est utilisé pour l'enrichissement et la présentation

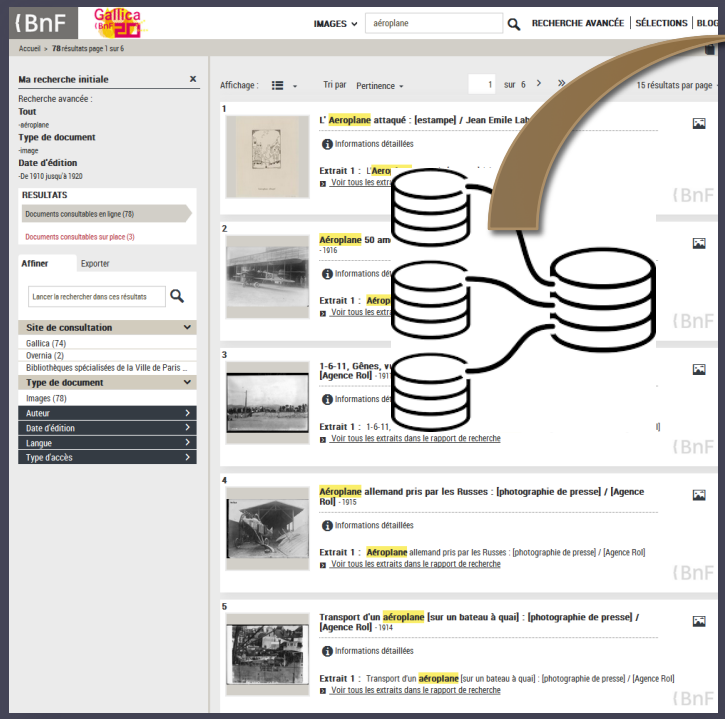


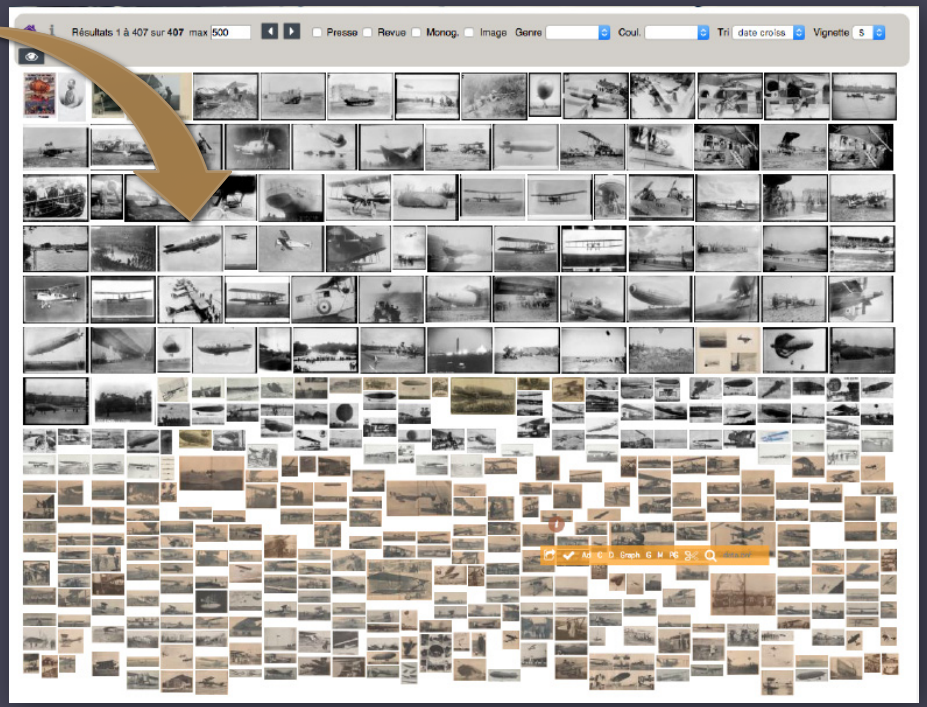
IIIF (*International Image Interoperability Framework*) : protocole d'accès aux images

iiif.io

- Les API et standards facilitent la R&D
- IIIF rend possible la réutilisation des images

SRU, OAI-PMH, IIIF…



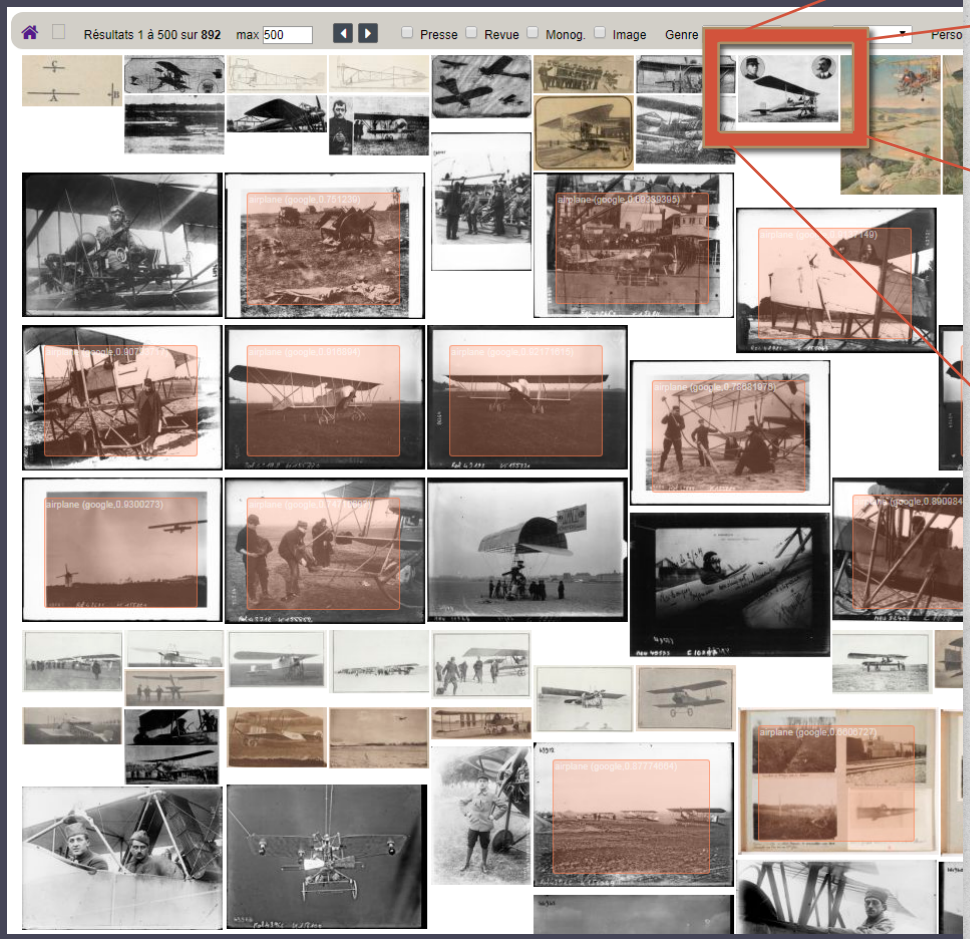


De Gallica (SERP + feuilletage)…

… à GallicaPix

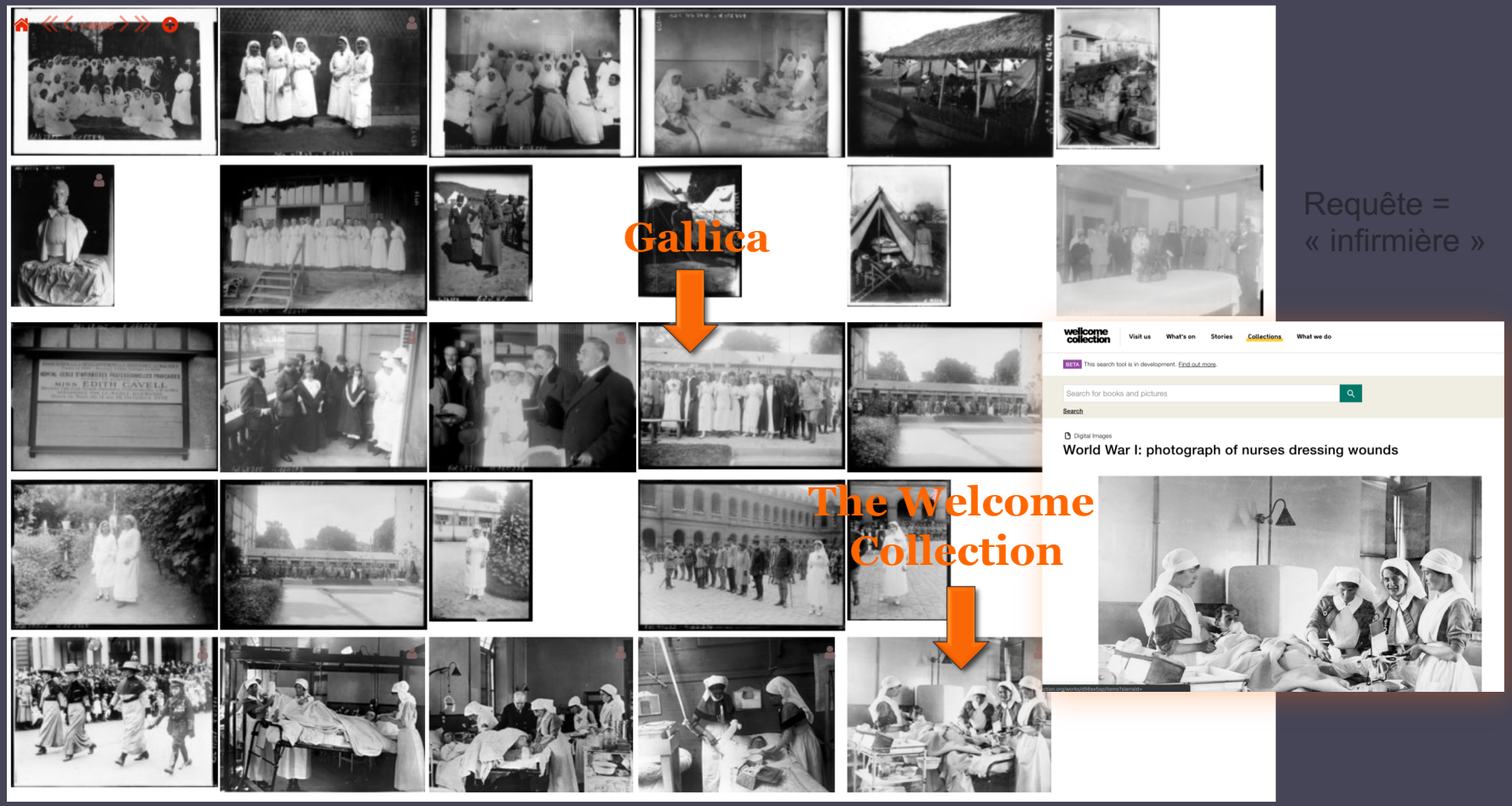
IIIF : valoriser l'image

- IIIF permet « d'entrer » dans le document et de le décrire



IIIF : interopérabilité

- IIIF permet **d'agréger** des contenus iconographiques et de les transformer/enrichir/remédiatiser, etc.



Gallica + The Welcome Collection (via Europeana) dans GallicaPix

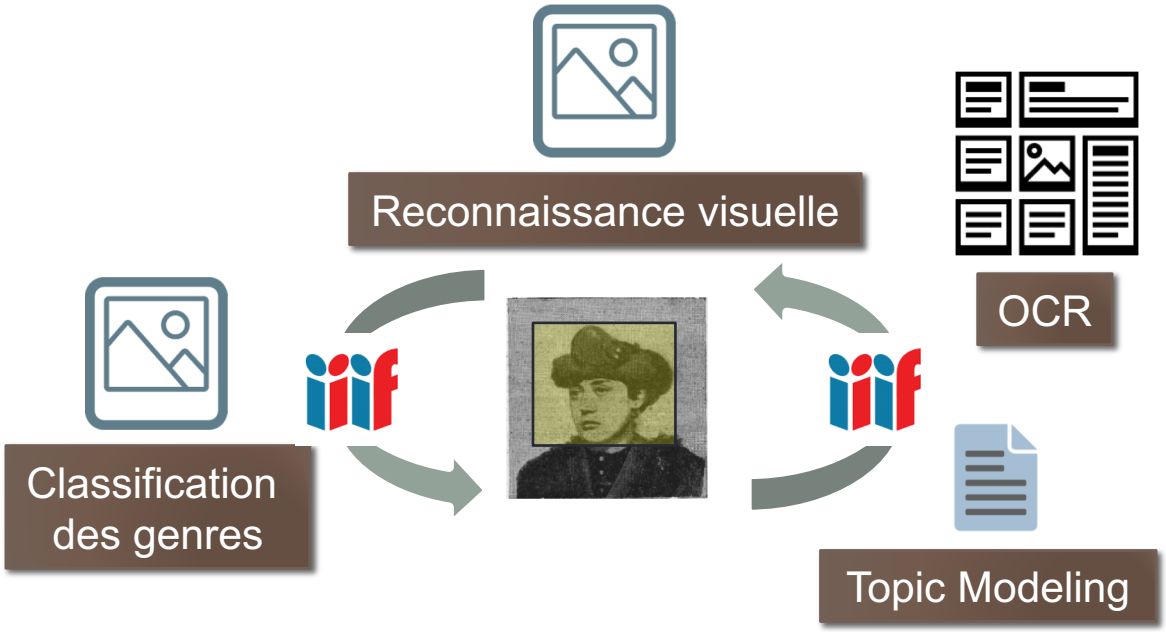
“Enrichir” les illustrations

Reconnaissance automatique des thèmes

OCR : Google Cloud Vision

Classification des genres (dessin, photo…) : réseau CNN (Inception)

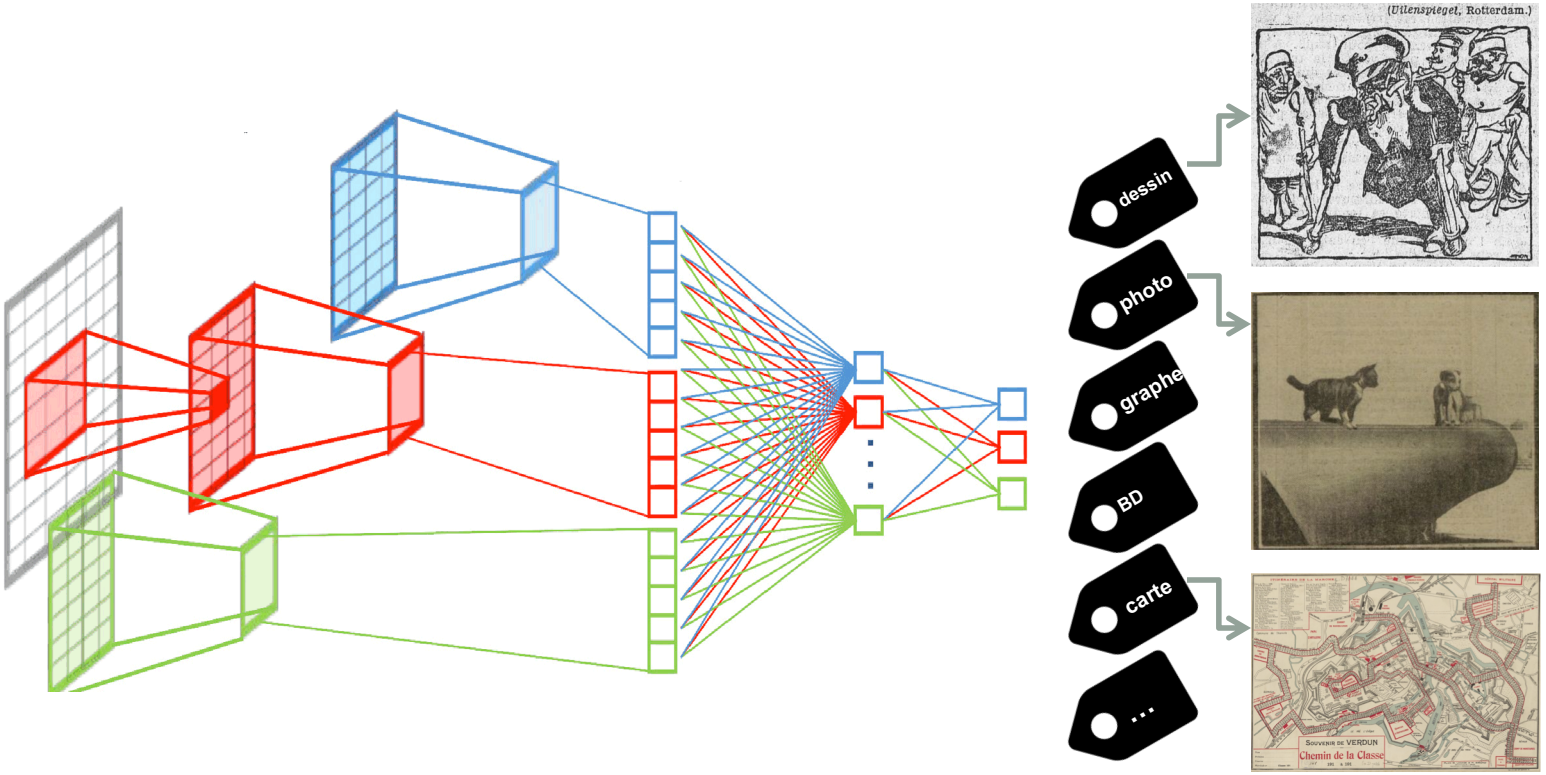
Reconnaissance visuelle : IBM Watson Visual Recognition, Google Cloud Vision, OpenCV/dnn, Yolo…



Classification de genres

Classification avec un réseau de neurones artificiels convolutionnel et une approche par « transfer learning »

Identification du « **genre** » des illustrations (photos, dessins, cartes, BD, graphes et schémas…)



Classification de genres

Transfer learning : seule la dernière couche du réseau est réentrainée sur un jeu de données Gallica (**12 classes)**

4 classes de « bruit » : couverture, page blanche, ornement, texte

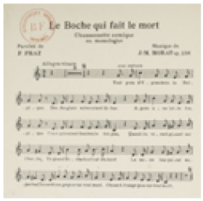
« Bruit »

Drawings (2024)

Photos (2449)

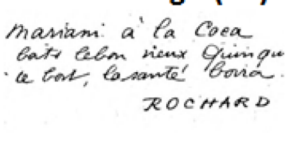
Advertisings (364)

Scores (616)



Comics (212)

Handwritings (64)



Engravings (1133)

Maps (282)

Ornaments (35)

Covers (86)

Blanks (178)

Texts (378)

Reconnaissance visuelle

Services de « détection d'objet » (IBM, Google, Clarifai…)

Génère des paires **concept/niveau de confiance**

Détecte **objets, visages, couleurs**…

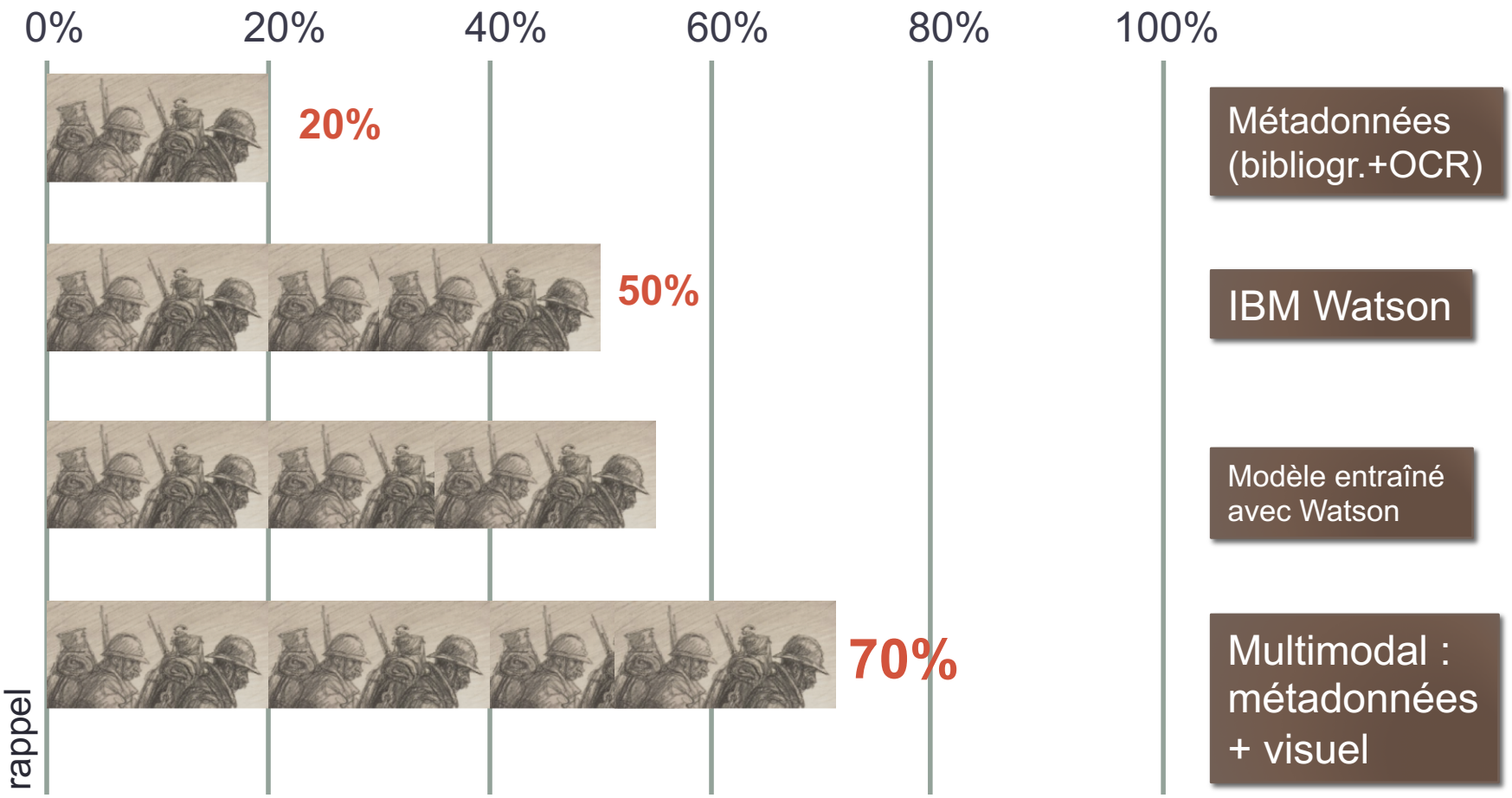
Les concepts lèvent le silence des métadonnées ou de l'OCR, les enjeux de traduction, les effets de l'évolution lexicale ("aéronef"/"avion")

« Les tanks de la bataille de Cambrai, la reine d'Angleterre écoute les explications données par un officiers anglais », 1917

```
black color - 0.90
vehicle - 0.70
coal black color - 0.69
armored vehicle - 0.57
truck - 0.52
...
```


```
"images": [
  {
    "classifiers": [
      {
        "classes": [
          {
            "class": "armored personnel carrier",
            "score": 0.568,
            "type_hierarchy": "/vehicle/wheeled vehicle/armored v
                armored personnel carrier"
          }, {
            "class": "armored vehicle",
            "score": 0.576        },
          {
            "class": "wheeled vehicle",
            "score": 0.705
          }, {
            "class": "vehicle",
            "score": 0.706
          }, {
            "class": "personnel carrier",
            "score": 0.541,
            "type_hierarchy": "/vehicle/wheeled vehicle/personnel
          }, {
            "class": "fire engine",
            "score": 0.526,
            "type_hierarchy": "/vehicle/wheeled vehicle/truck/fire e
          }, {
            "class": "truck",
            "score": 0.526
          }, {
            "class": "structure",
            "score": 0.516
          }, {
            "class": "Army Base",
            "score": 0.511,
            "type_hierarchy": "/defensive structure/Army Base"
          }, {
            "class": "defensive structure",
            "score": 0.512
```

Expérimentation sur la détection de soldats



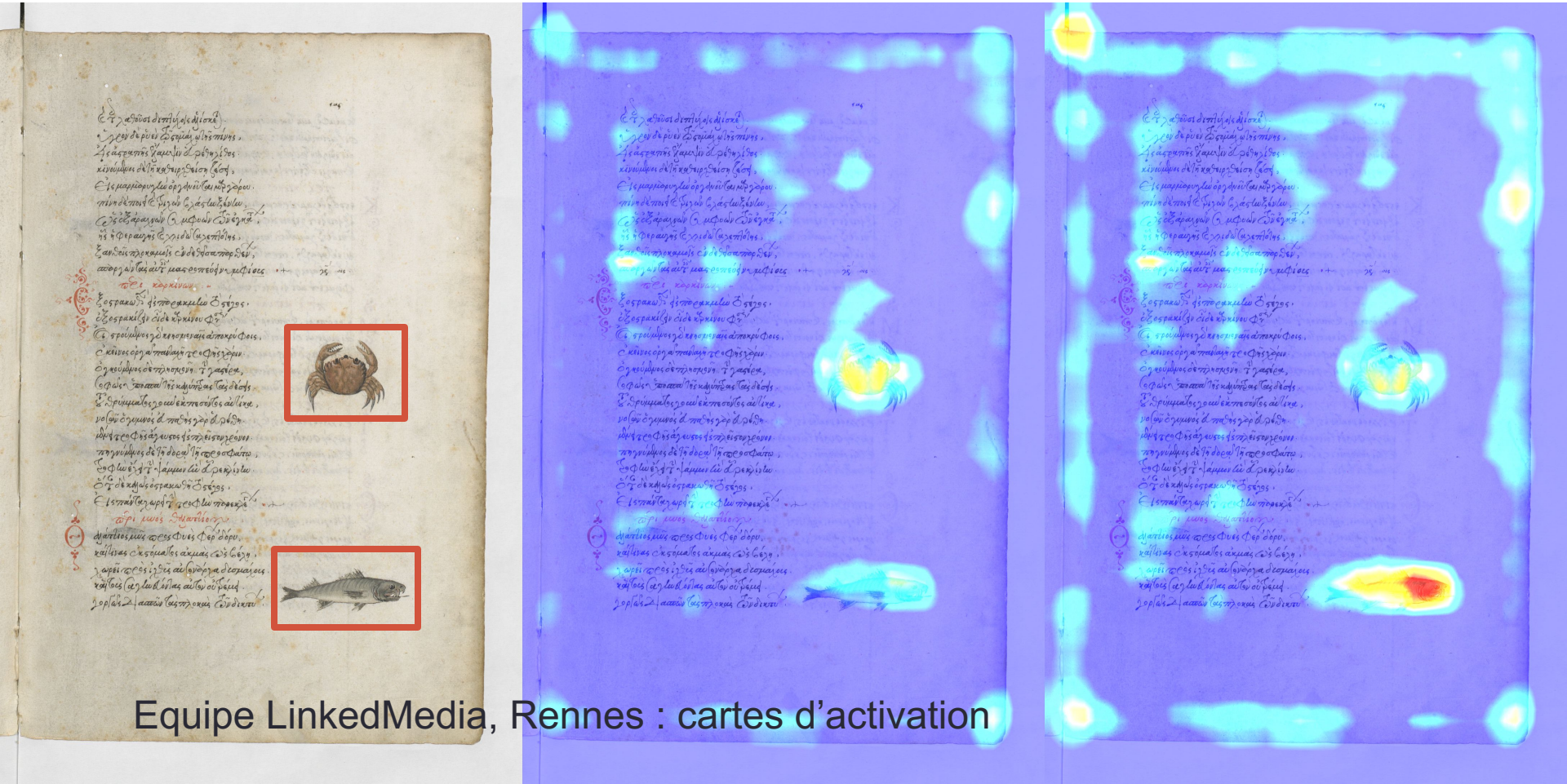
« Rappel » : proportion des documents pertinents proposés parmi l'ensemble des documents pertinents (« exhaustivité », « sensibilité »)

Reconnaissance visuelle : remarques

Les images non segmentées conduisent à des **classes génériques** : « cadre », « document », « document imprimé »…

Reconnaissance visuelle : remarques

Projet INRIA/BnF (convention-cadre ministère de la Culture et INRIA)



Equipe LinkedMedia, Rennes : cartes d'activation

Reconnaissance visuelle : remarques

Ces modèles génériques peuvent opérer sur **des documents patrimoniaux (XIX^e^, XX^e^),** y compris sur des cas « difficiles ».

Reconnaissance visuelle : remarques

Mais des **limitations existent** :

- Généralisation à partir de jeux d'entrainement majoritairement contemporains ➔ **anachronismes**
- Généralisation à partir de jeux nécessaire-ment clos ➔ **erreurs de classification**
- **Scènes complexes** difficiles à interpréter

egway

nicule

wine label

car bombing

10 000 classes permettent de satisfaire des recherches généralistes sur des contenus modernes ou contemporains, mais pas sur le large spectre des collections patrimoniales...

Reconnaissance visuelle : modèles ad hoc

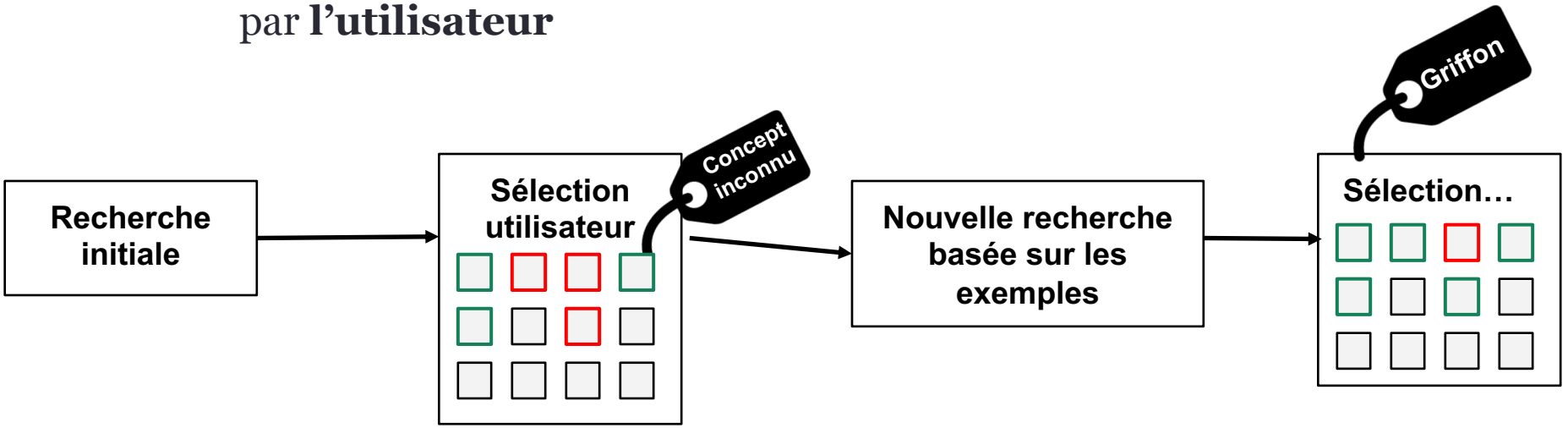
Projet INRIA/BnF (convention-cadre ministère de la Culture et INRIA)

Equipe LinkedMedia, Rennes : base **Mandragore**/Zoologie, 400 classes

Reconnaissance visuelle : bouclage de pertinence

Projet INRIA/BnF (convention-cadre ministère de la Culture et INRIA)

- Principe : sélection itérative de documents pertinents par **l'utilisateur**



Equipe INRIA/Zenith

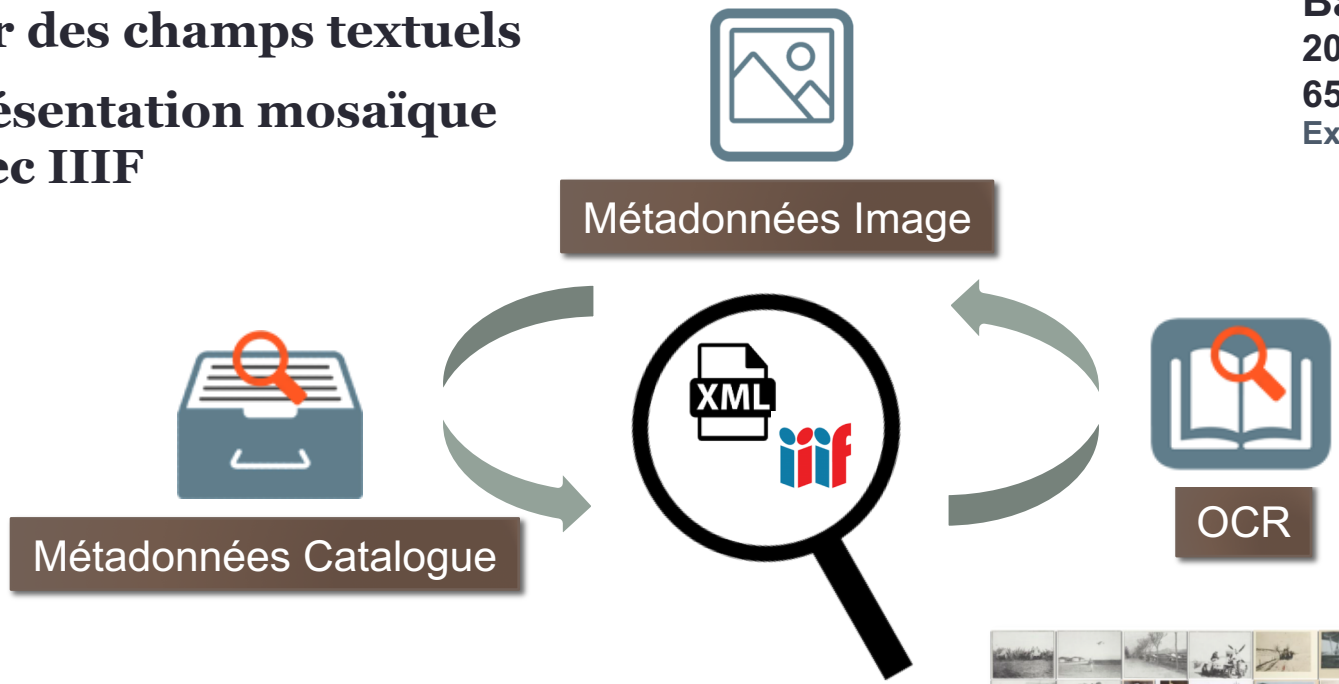
Rechercher

Dans une base XML (BaseX, XQuery)

Sur des champs textuels

Présentation mosaïque avec IIIF

Base 14-18 :
200 k illustrations
65 k pubs illustrées
Extraites de 500 k pages



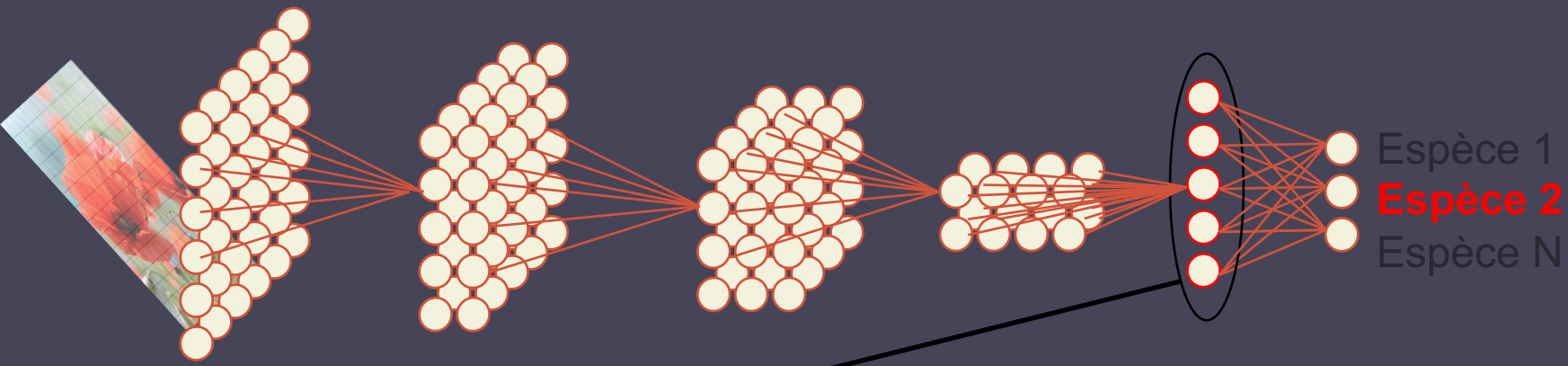
{BnF Gallica Studio

http://gallicastudio.bnf.fr

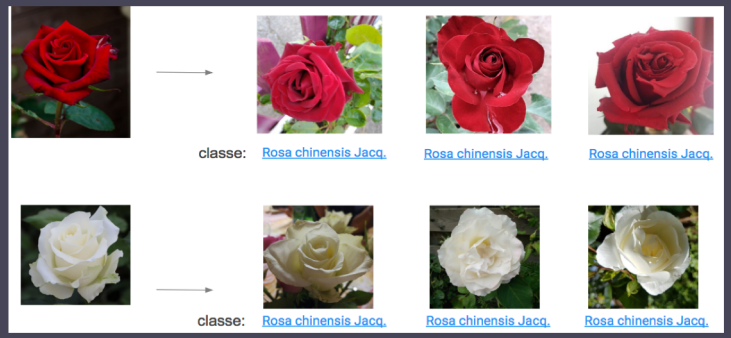
http://demo14-18.bnf.fr:8984/rest?run=findIllustrations-form.xq

Focus : indexation et présentation

SNOOP v3 : deep learning

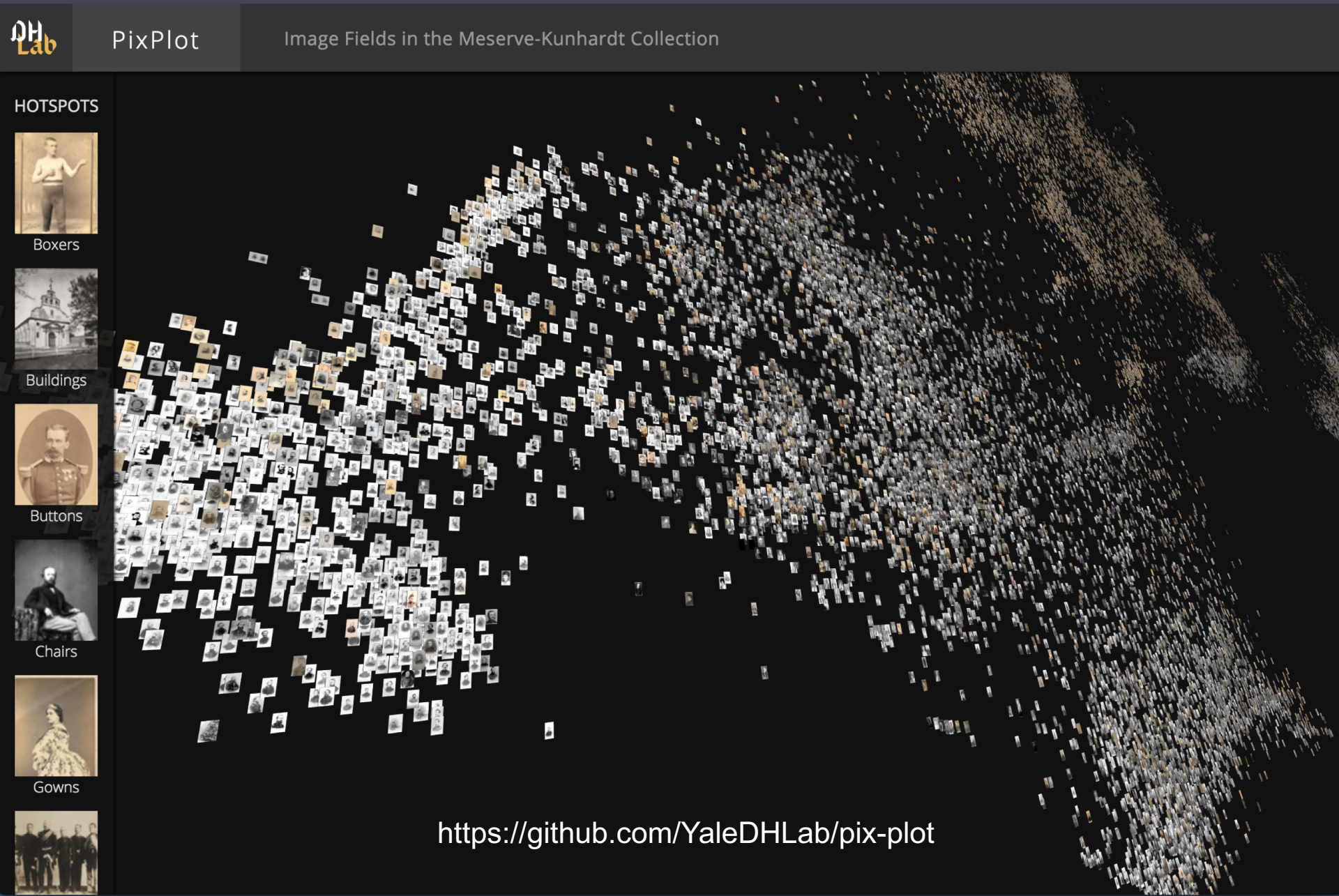


Index de similarité visuelle (dans un **espace numérique** de grande dimension)



Equipe INRIA Zenith / INA : moteur d'indexation SNOOP

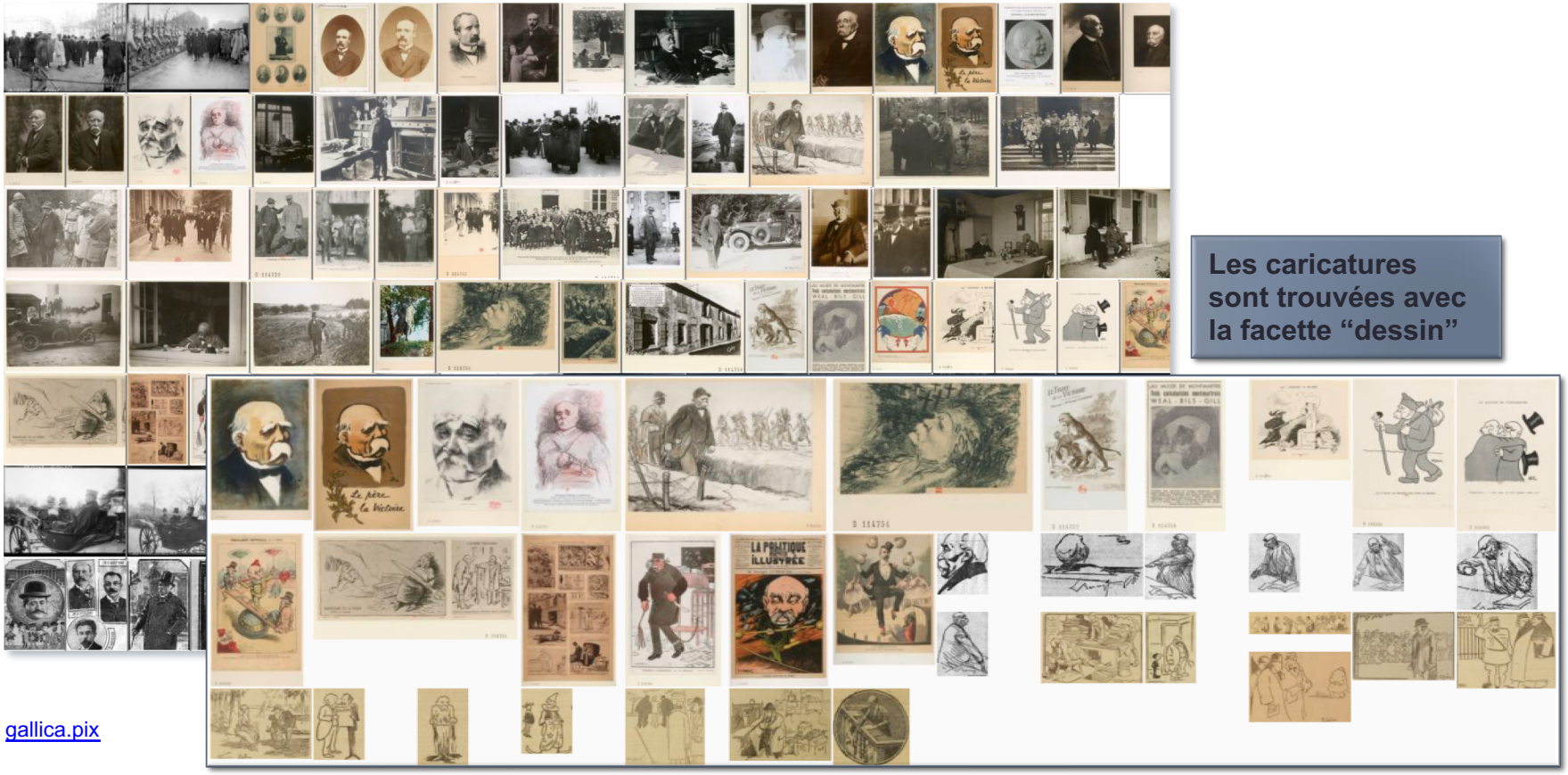
Focus : indexation et présentation



Cas d'usage : recherche encyclopédique sur un nom

Les métadonnées et l'OCR sont utilisés.

"George Clemenceau" : **140** ill. dans Gallica/Images, **> 900** dans GallicaPix



Les caricatures sont trouvées avec la facette "dessin"

gallica.pix

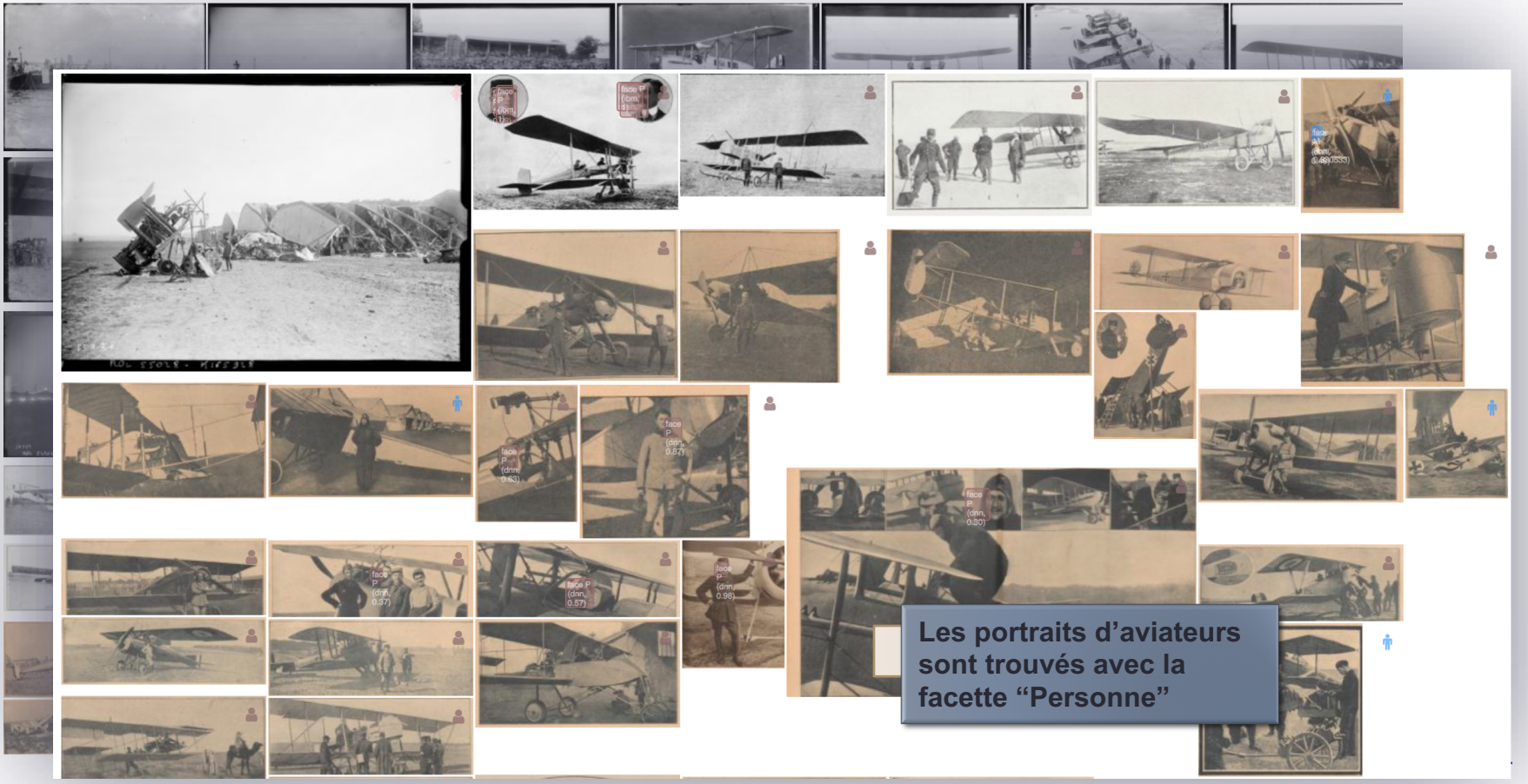
Cas d'usage : recherche encyclopédique sur un concept

Une recherche sur le **mot-clé** "avion" renvoie du bruit…
(portraits d'aviateur, photographies aériennes…)

gallica.pix

Cas d'usage : recherche encyclopédique sur un concept

En utilisant le **concept visuel** avion, le bruit peut être filtré.
(au prix de quelques faux positifs !)



gallica.pix

Cas d'usage: recherche multimodale

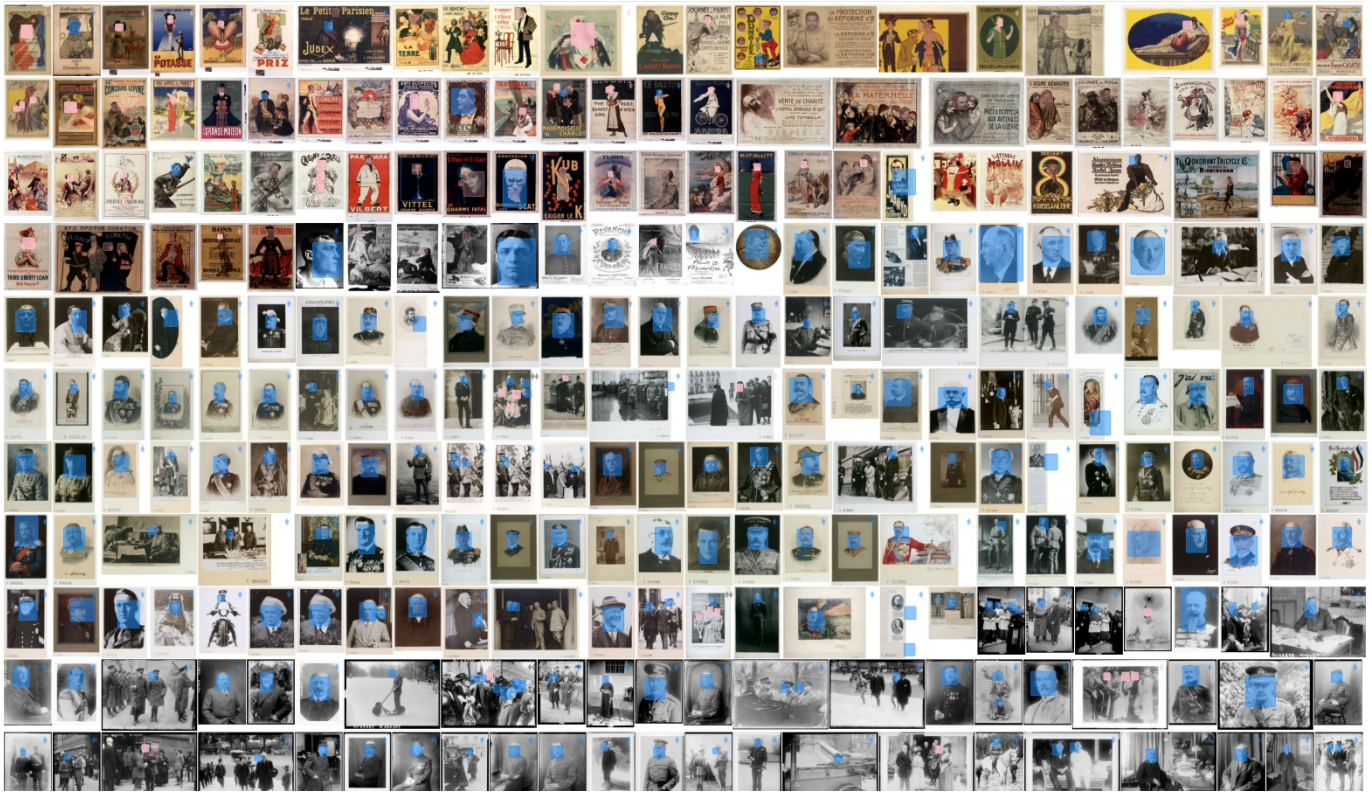
Les **concepts visuels, le texte et les métadonnées** sont utilisés.
Recherche relative aux destructions urbaines consécutives à la bataille de Verdun : `concept=("rue" OU "maison" OU "ruines") ET mot-clé="Verdun"`

gallica.pix

Expérimentation sur la détection de personne

Ces modèles permettent aussi la détection de **visage** et de **genre** :

- IBM Watson : “**Visages**”: rappel = **43 %**, précision = **99,9 %**



Expérimentation sur la détection de personne

Module dnn (*deep neural networks*) dans OpenCV 3.3, modèle ResNet, méthode “Single Shot Multibox Detector” (SSD) :
rappel = **58 %,** précision = **92 %** (CS > 20 %)

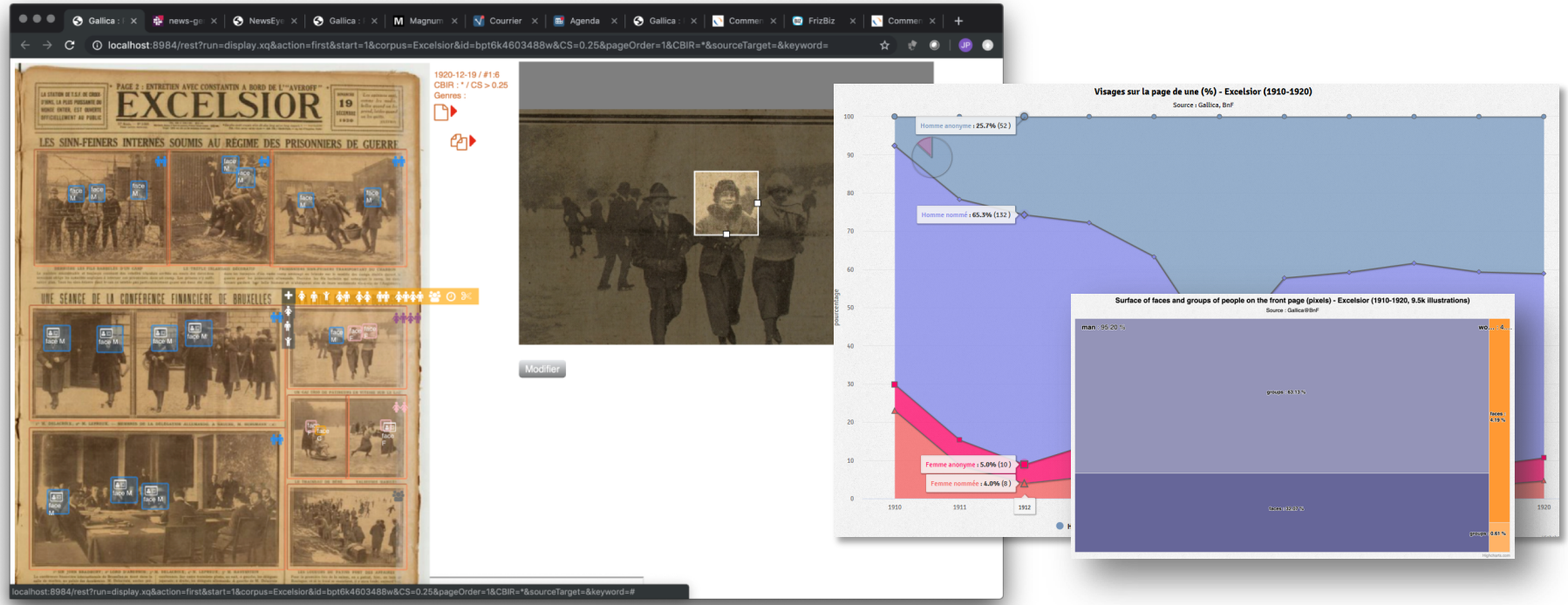
gallica.pix

Modèle Yolo v3

Cas d'usage : humanités numériques

Projet ANR **Numapresse**, collaboration avec la BnF :

- Reconnaissance automatique des **visages et genres** dans *Paris-Match* et *L'Excelsior* (pages de une, période de 10 ans)
- Post-correction manuelle avec l'éditeur GallicaPix (de 1 heure à 2 jours)
- Analyse de données et datavisualisation



Journée d'étude Numapresse « Paris Match : le poids des mots, le choc des photos » (université Paul-Valéry Montpellier III).
https://www.fabula.org/actualites/journee-d-39-etude-numapresse-paris-match-le-poids-des-mots-le-choc-des-photos-universite-paul-_90246.php

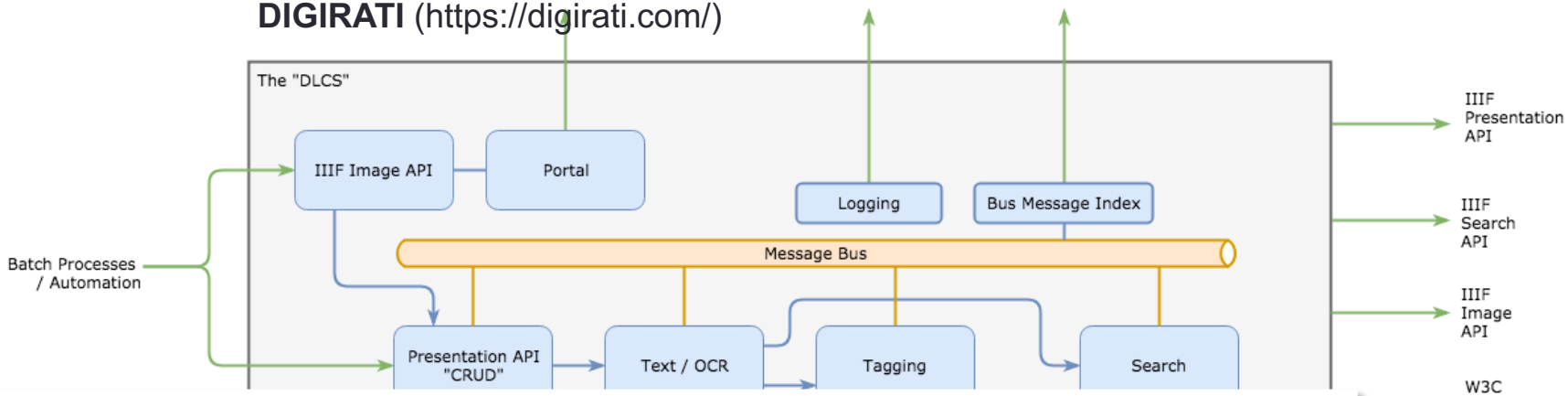
Cas d'usage : humanités numériques

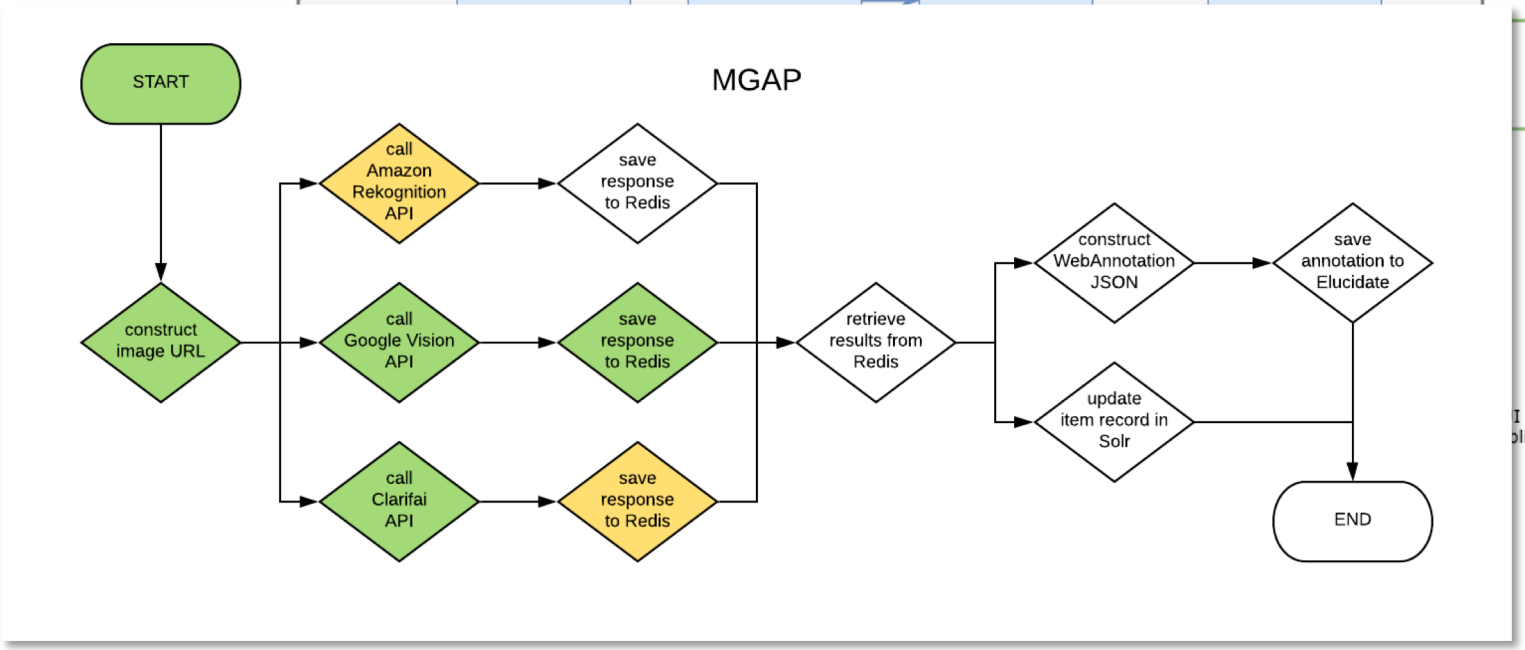
Hackathon DHH, mai 2019, Helsinki : thème "Newspapers and Capitalism", focus sur la **publicité**

- Reconnaissance automatique (Yolo v3) des **moyens de transport** dans le sous-corpus de publicités illustrées de GallicaPix (60 k)
- Post-correction (1 jour), analyse de données et datavisualisation

Production : comment faire ?

DIGIRATI (https://digirati.com/)

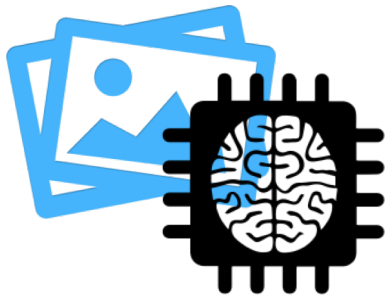




UCLA (https://github.com/UCLALibrary?language=python)

Conclusion

- **Un accès unifié à toutes les illustrations** d'une collection encyclopédique est un service répondant à de réels besoins.
- Il favorise aussi la **réutilisation** des illustrations.
- La maturité des techniques **IA en matière d'indexation visuelle** rend possible leur intégration dans la boîte à outils standard d'une bibliothèque. **IIIF** et **IA** forment un tandem efficace.
- Leurs résultats, **mêmes imparfaits**, aident à rendre visibles les illustrations de nos collections.
- Il n'y a **pas de solution universelle** en matière d'indexation visuelle, mais des applications immédiates sont possibles.



Merci pour votre attention !

jean-philippe.moreux@bnf.fr

Jeux de données, modèles et scripts :

- **api.bnf.fr**
- **github.com/altomator/Image_Retrieval**

GallicaPix :

- **gallicastudio.bnf.fr**
- **http://demo14-18.bnf.fr:8984/rest?run=findIllustrations-form.xq**

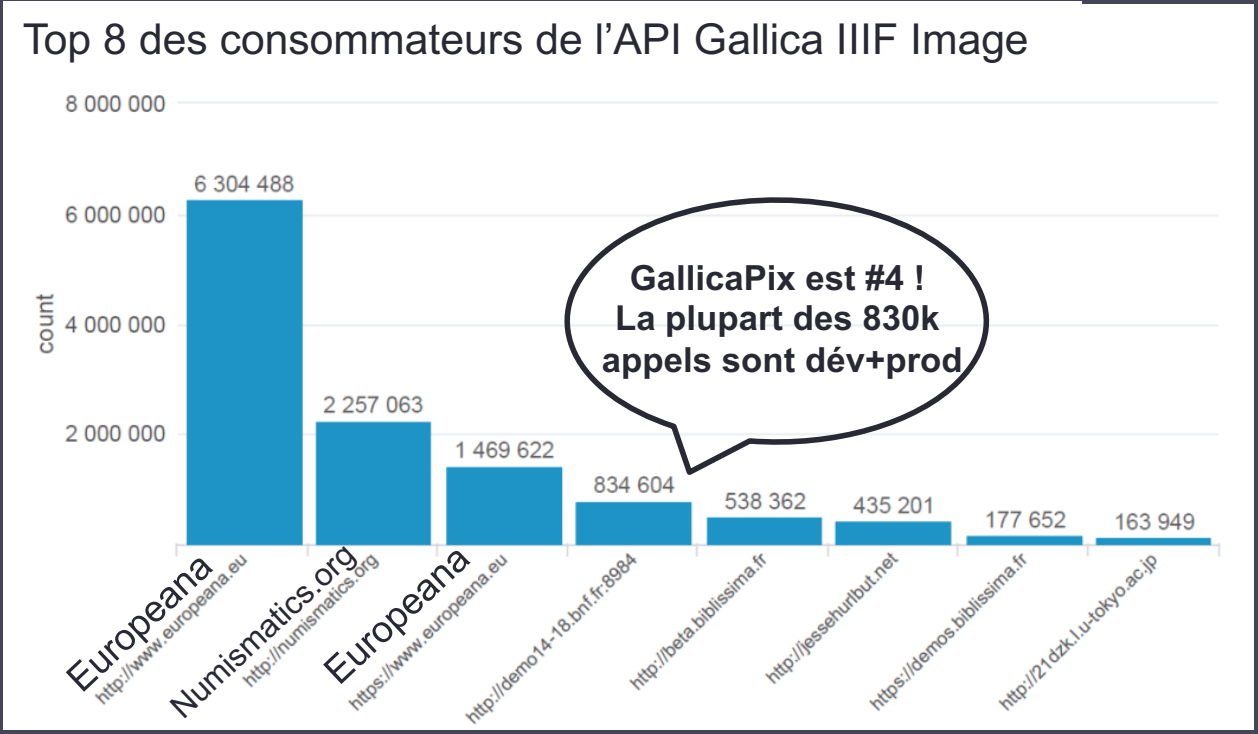
Focus IIIF : humanités numériques

- Les chercheurs sollicitent de plus en plus les institutions patrimoniales pour la mise à disposition de contenus iconographiques numérisés. Ils s'essayent désormais à l'extraction des illustrations de documents textuels (presse, magazines, encyclopédie et dictionnaires…)
- Un écosystème IIIF peut être mis en place directement au-dessus des collections numériques des institutions, à des fins de recherche, grâce aux API IIIF Presentation et IIIF Search.
- IIIF facilite aussi R&D et expérimentation avec les techniques d'intelligence artificielle (appliquées aux images et aux textes).

Focus IIIF

Points d'attention :

- Le téléchargement peut être lent par rapport à du local
- Peut mettre sous pression les serveurs IIIF locaux… L'usage de IIIF pour le prototypage et la production peut altérer la qualité de service des vrais « utilisateurs »



Service IIIF Image à la BnF :

- 5 serveurs
- Proliant BL460c G7 (2 Xeon CPU E5649, 2,53 Ghz, 24 coeurs)
- 10 M appels/mois

Focus IIIF

L'analyse d'image est plus facile avec IIIF

- Traitements basiques avec les paramètres IIIF (region, rotation)
- Adapter la qualité de l'image aux attentes du modèle (size, quality)
- Pas besoin de gérer des fichiers, pas de stockage local
- La plupart des API acceptent les URL en entrée

```
curl -X POST -u "apikey:****" --
form
"url=https://gallica.bnf.fr/iiif/ar
k:/12148/bpt6k9604090x/f1/22,781,43
34,4751/,700/0/native.jpg"
"https://gateway.watsonplatform.net
/visual-
recognition/api/v3/classify?version
=2018-03-19"
```

Remarques finales sur IIIF

- **Je n'aurais jamais eu l'idée de démarrer ce projet sans IIIF**. Il m'aurait d'abord fallu mettre en place un serveur d'images !
- **IIIF est tellement efficace que nous avons besoin de plus d'implémentations IIIF…**
- **Nous avons besoin de l'IA pour tenir le rythme de l'accroissement de nos collections numériques** (numérisées ET nées numérique) : classifier, filtrer, annoter, enrichir. Heureusement, **IA et IIIF forme un beau tandem** (de la création de jeux d'entrainement à la médiation numérique)

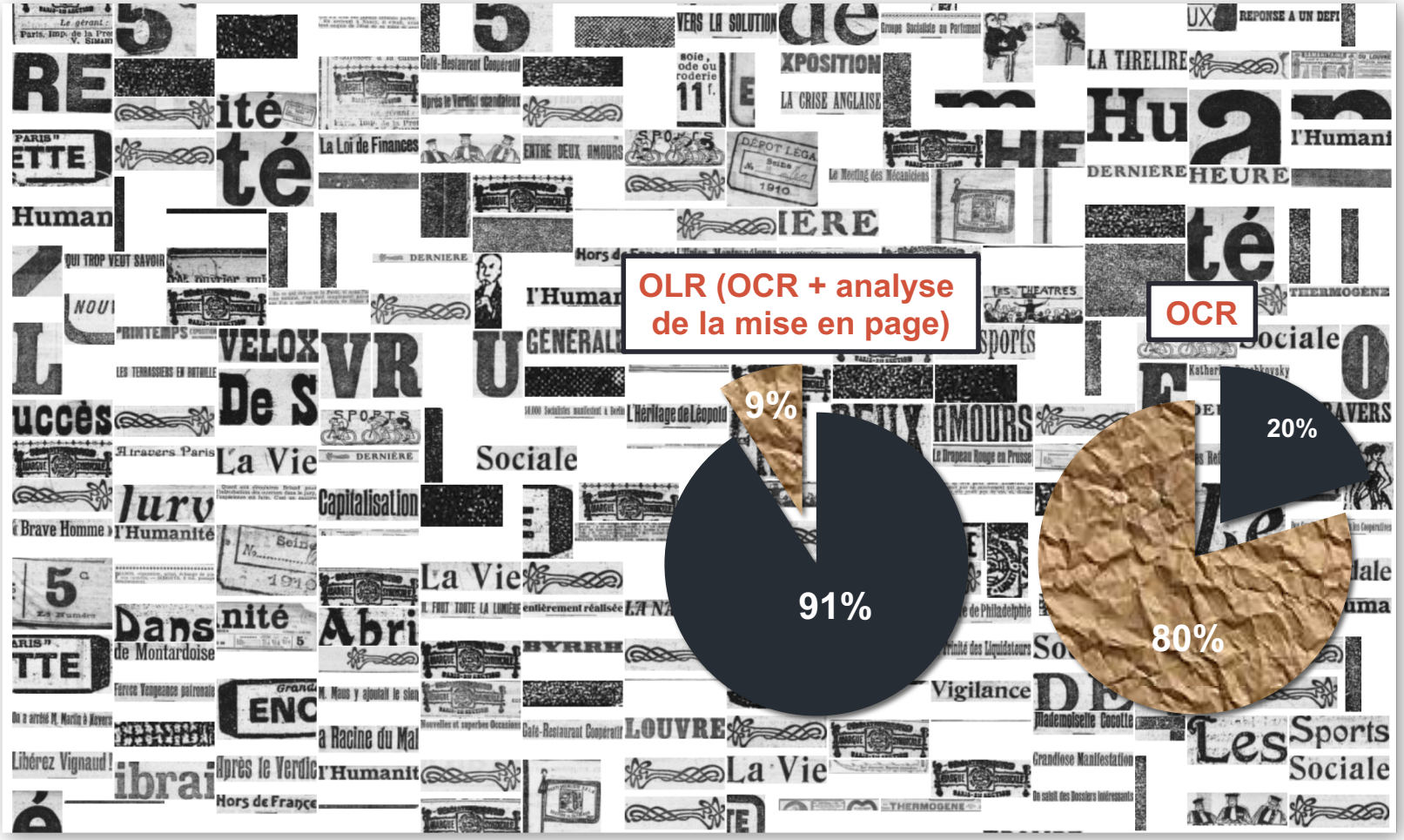
Focus IIIF

La création de jeux d'entrainement est facilitée par IIIF

1. Amorçage du jeu (bootstrap) grâce aux métadonnées
2. Sélection de nouveaux éléments dans l'application web
3. Export du jeu (liste d'URL IIIF)
4. Téléchargement des images
5. Entrainement du modèle

Le partage des jeux de données est aussi facilité !

Où sont les illustrations ?



Presse : 80 % des illustrations indiquées par l'OCR n'en sont pas !