

Mise en œuvre de IIF pour la reconnaissance automatique de documents

IIF - Biblissima

Christopher Kermorvant

T E K L I A

Traitements automatiques de documents

TEKLIA - Arkindex

- Traitement automatique de grandes collections de documents numérisés
- Basé sur des algorithmes de Machine/Deep Learning
- *Platform As A Service*
- Compatible IIIF (dès le début)

Arkindex versus eScriptorium ?



Traitements automatiques de documents

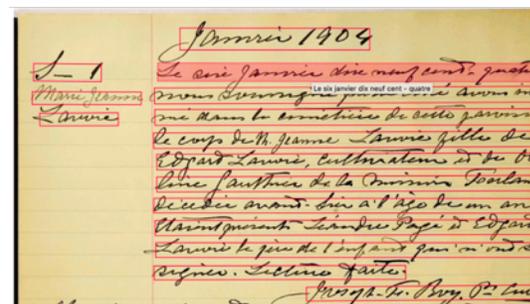
Page classification



Page segmentation



Text line detection and recognition



Named-entity and relation extraction, linking



Indexation

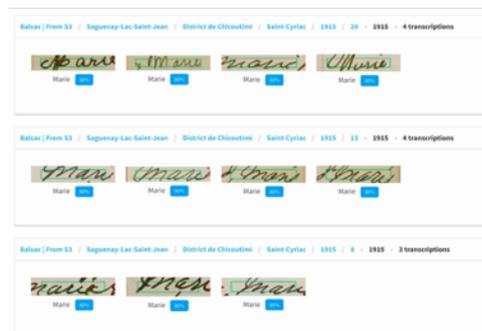
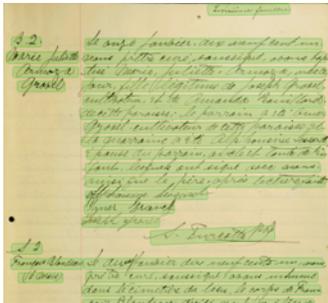


Illustration de l'usage de IIF sur 2 projets



HORAE (HOurs - Recognition, Analysis, Editions)
Etude des pratiques religieuses de la fin du Moyen Âge à travers
les livres d'heures
ANR-17-CE38-0008



BALSAC
Reconnaissance de 6 millions d'actes paroissiaux du Québec
1850-1920.
Université du Québec à Chicoutimi



Projet HORAE : études des livres d'heures

The screenshot displays the ArkIndex web application interface. At the top, the navigation bar includes 'ArkIndex', 'Projects', 'Process', and 'Search'. The user is logged in as 'kermorvant@tekli.com'. The main header shows the project name 'HORAE | Machine Learning', the volume 'Volume set set_1', and the specific page 'Page 58r'. The central area features a manuscript page with a digital overlay. The text is in Gothic script, with a large decorated initial 'D' in blue and gold. The text reads: 'concede. Per dñm nrm. Benedicamus domino Deo gratias sedetum anime p miãam dei requiescant in pa ce amen. ad primam. Deus in adiu torium meum intende. Omne ad adiuuandũ me testina. Gloria patri. Sicut erat. Hymnus. Memento salutis auc tor quod nrm quon dam corporis exultata u'. The left sidebar contains a list of elements on the page, such as 'Decorations 1-7', 'Initials 23-27', and 'Text line 3', which is currently selected. The right sidebar shows the 'Text line 3' details, including 'CLASSIFICATIONS' and 'TRANSCRIPTIONS'.

ArkIndex Projects Process Search

0 kermorvant@tekli.com

HORAE | Machine Learning / Volume set set_1 / Volume France, Paris, Bibliothèque nationale de France, Bibliothèque de l'Arsenal, 1194 / Page 58r

Page n°117 Display Actions

Filter by type... Display

Page 58r

- Decorations 1
- Decorations 2
- Decorations 3
- Decorations 5
- Decorations 7
- Initials 23
- Initials 24
- Initials 25
- Initials 26
- Initials 27
- Rubrications 14
- Rubrications 15
- Section Hymn | L3 | Section
- Section Invocation | L3 | Section
- Section Prime | L2 | Section
- Text Segment Amen | Amen | Acclamatio...
- Text Segment Deo gratias | Deo gratias | ...
- Text Segment Gloria Patri, et Filio: et Spi...
- Text Segment Memento salutis auctor. Q...
- Text Segment Deus in adiutorium meum ...
- Text Segment Domine ad adiuuandum m...
- Text Segment Fidelium anime per miseri...
- Text Segment Benedicamus Domino | Ca...
- Text line 2
- Text line 3
- Text line 4
- Text line 5
- Text line 6
- Text line 7
- Text line 8
- Text line 9
- Text line 10
- Text line 11
- Text line 12
- Text line 13
- Text line 14
- Text line 15
- Text line 16

Text line 3 Created by U-FCN Line Historical

CLASSIFICATIONS

Add a classification ADD

TRANSCRIPTIONS

Created by Kaldi Horae New OM

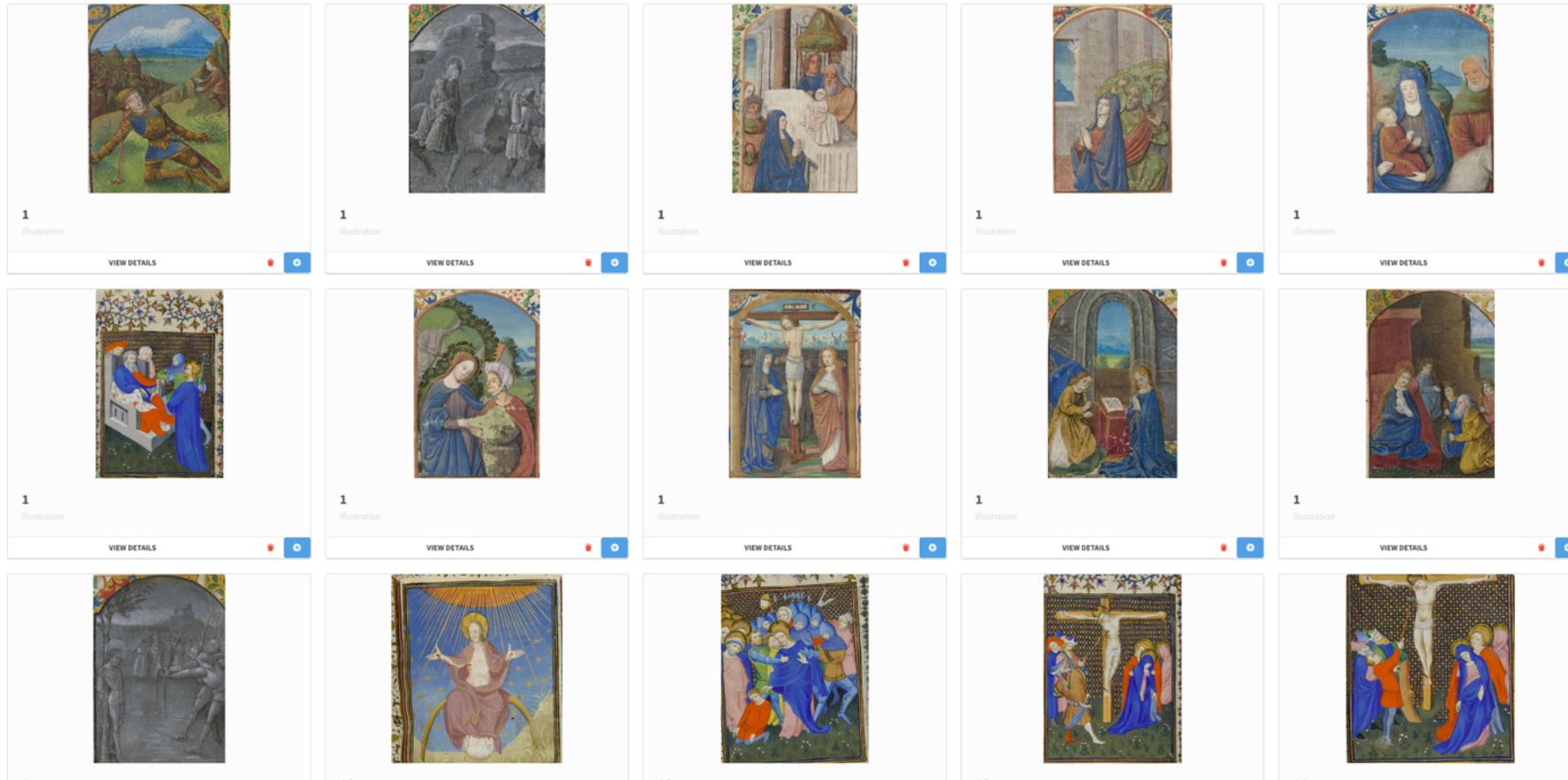
Benedicamus Domino Deo

METADATA

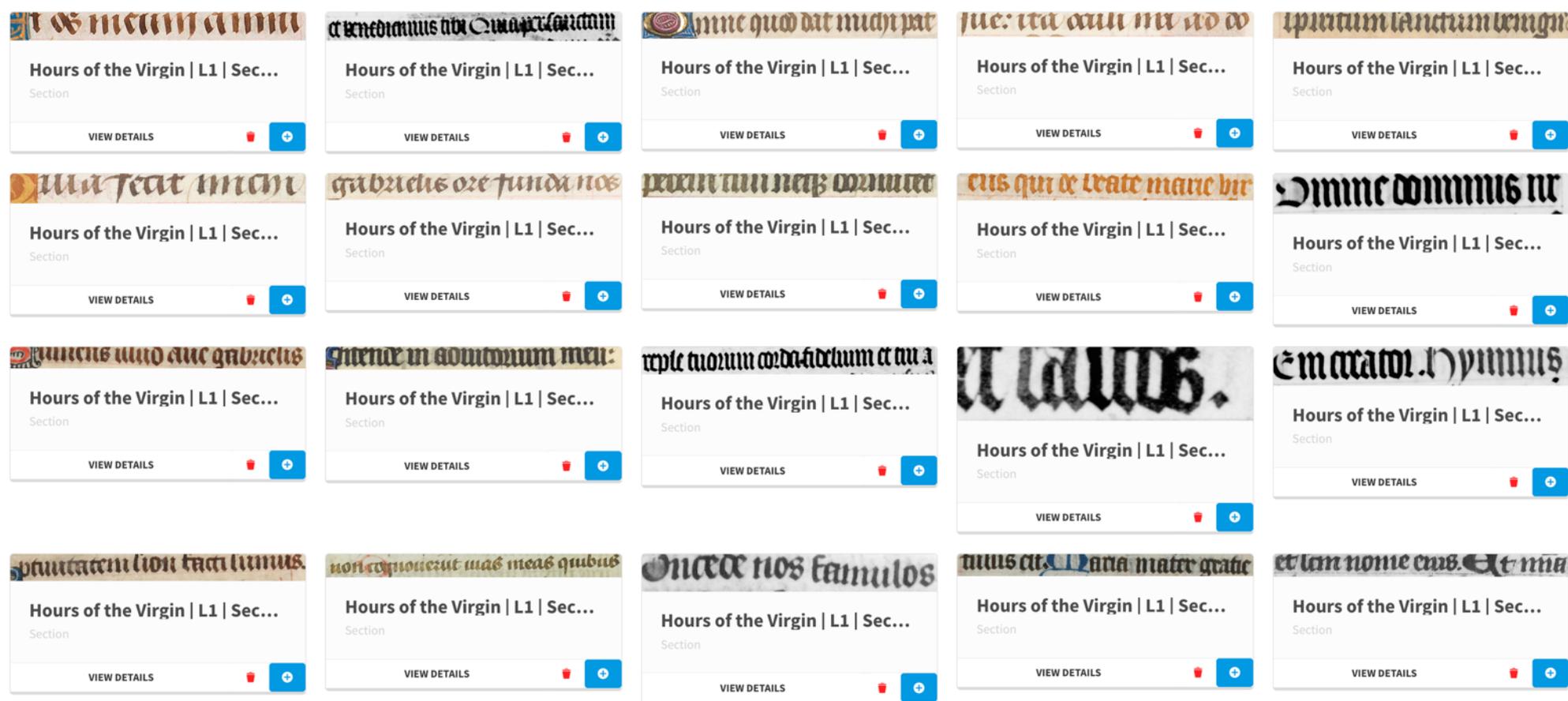
ALL ENTITIES

ROLES

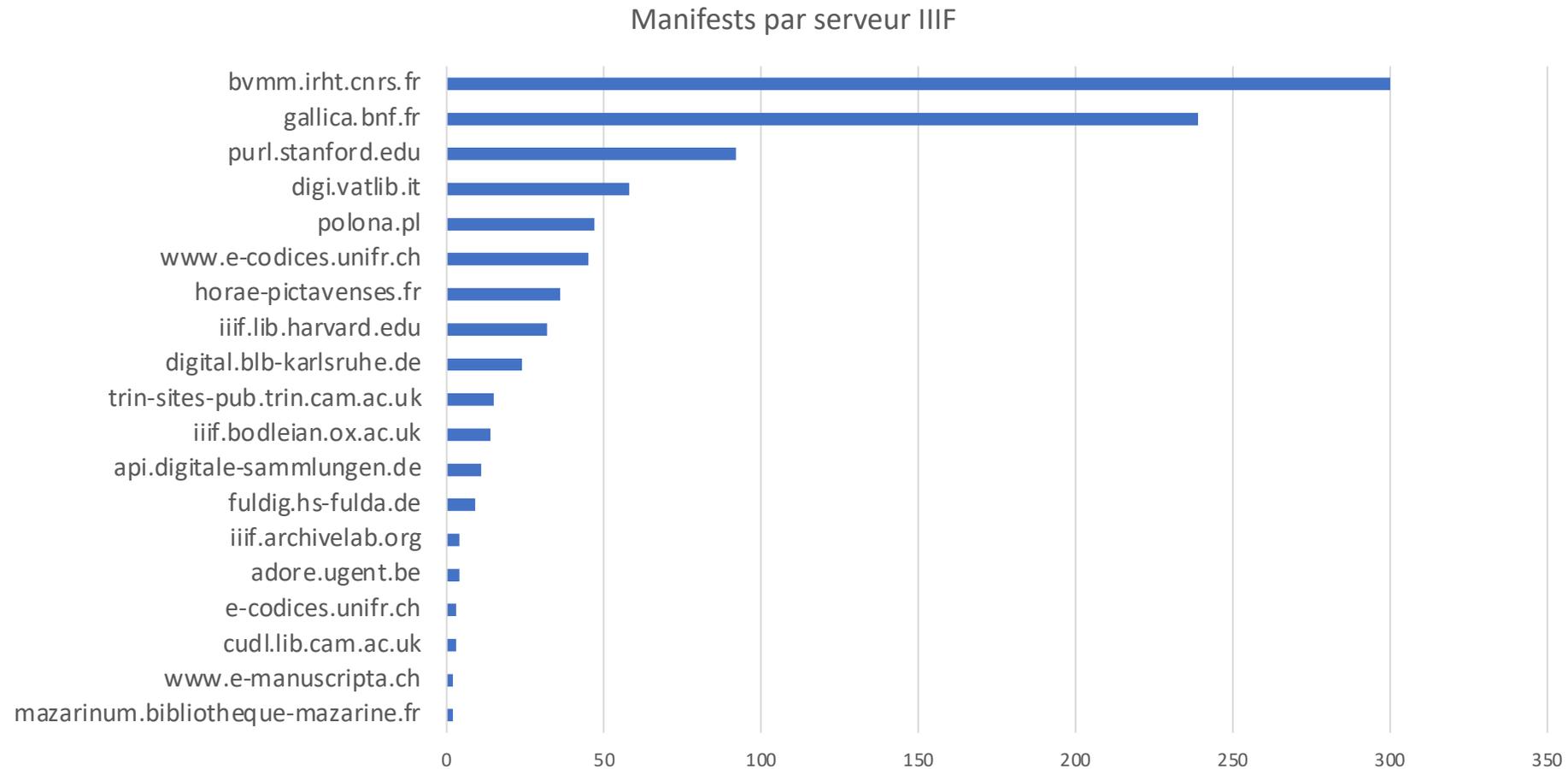
IIIF : Accès à toutes les miniatures



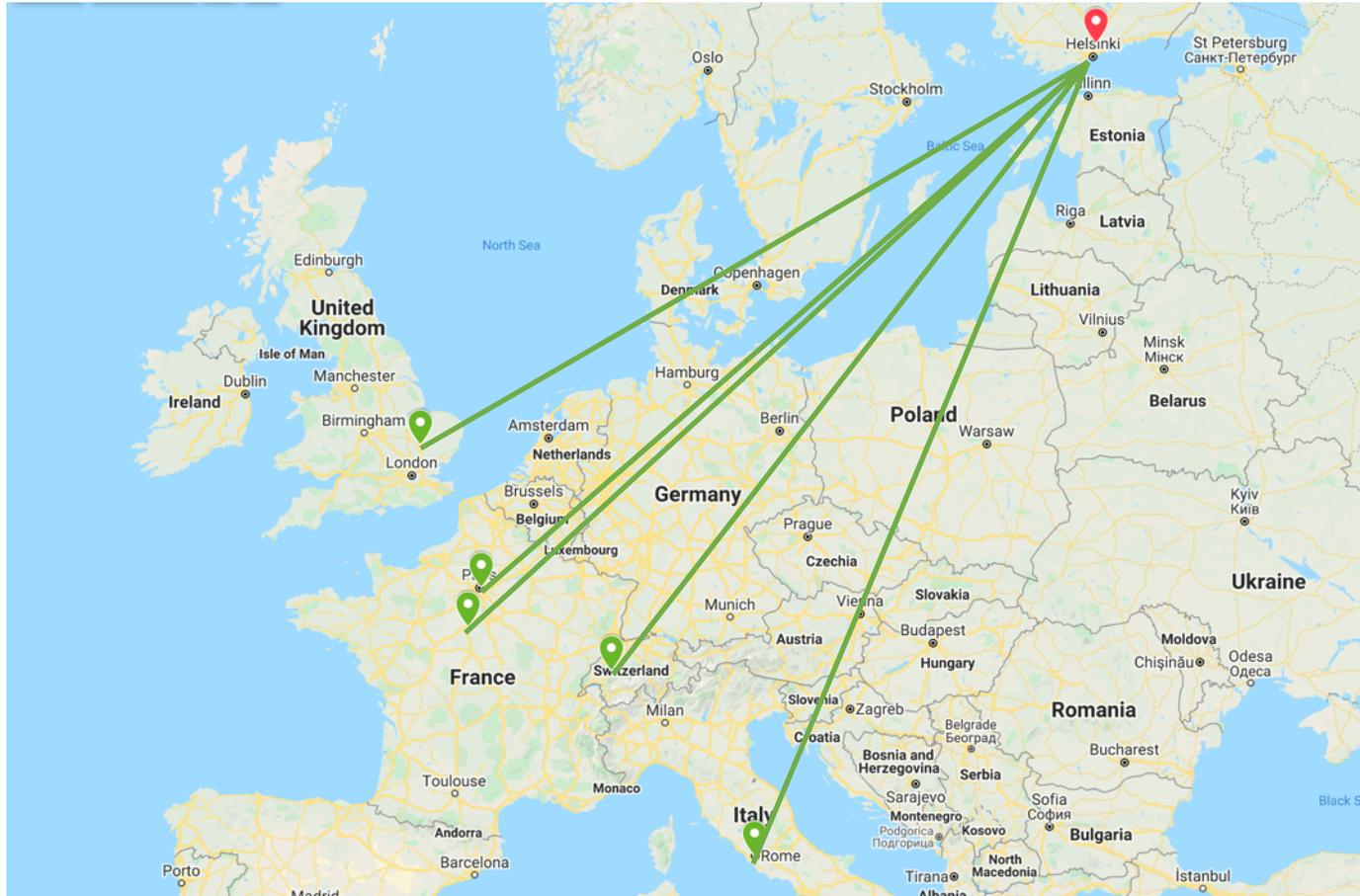
IIIF : Accès à la structure, les heures de la Vierge



Projet HORAE : 944 livres d'heures



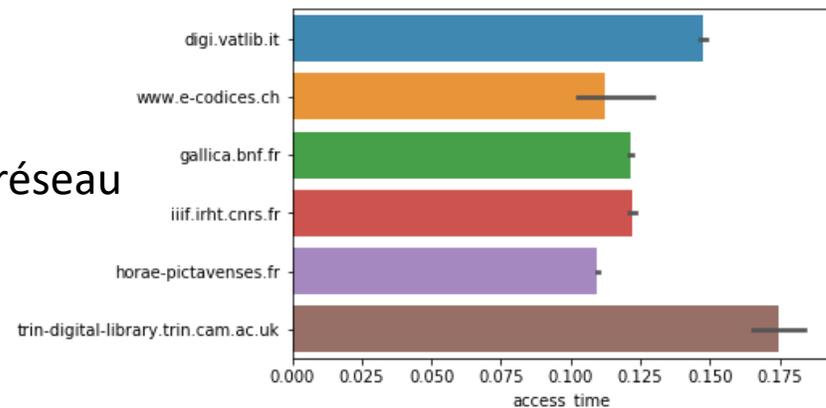
Projet HORAE : performances des serveurs



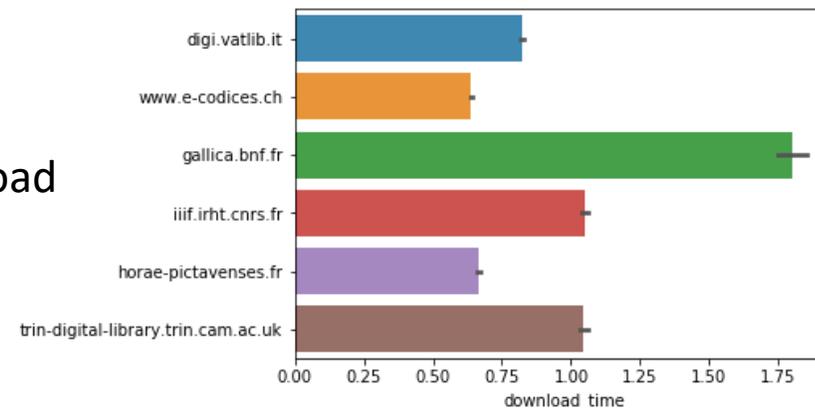
5 serveurs IIF en Europe
1 serveur de traitement à Helsinki

Projet HORAE : performances des serveurs

Temps réseau



Temps Download



62 529 requêtes

Projet HORAE : performances des serveurs

serveur	position	temps	nombre requêtes	logiciel
digi.vatlib.it	Vatican	0.82	21441	IIPImage
www.e-codices.ch	Fribourg, Suisse	0.62	14863	Loris
gallica.bnf.fr	Paris, France	1.80	11760	IIPImage
iiif.irht.cnrs.fr	Orléans, France	1.04	8806	OmekaS
horae-pictavenses.fr	Paris/Niors, France	0.66	2868	OmekaS
trin-digital-library.trin.cam.ac.uk	Cambridge, Angleterre	1.04	2791	Cantaloupe

Les performances dépendent de paramètres du serveur, de la charge, du cache, du format d'image...

Projet Balsac : Extraction d'information dans les registres paroissiaux du Québec (1850-1920)

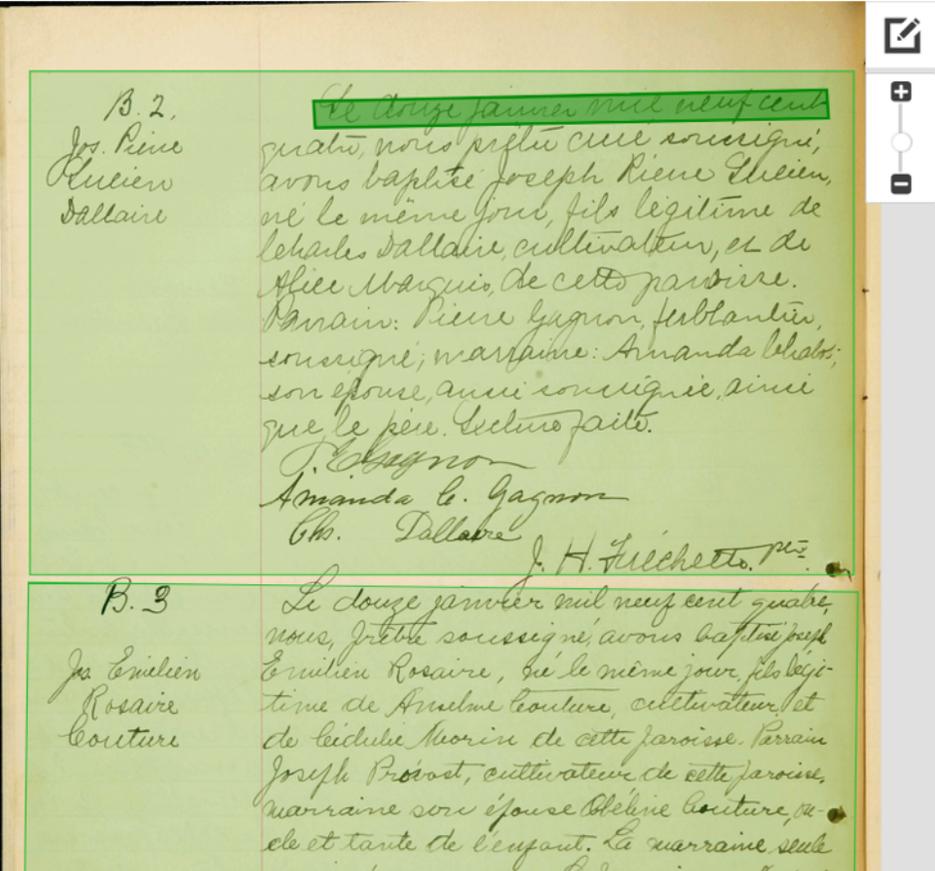
BALSAC | S3 / Center Québec / District District de Beauce / Parish Sainte-Claire / Register 1904 / Page 6

Page n°6 ← → Display ▾ Actions ▾

Acts (3/57) ▾ Hide

Page 6

- Act 1
- Act 2
- Act 3



Text line 2 Created by **U-FCN Line Historical**

U-FCN Balsac act

baptism

act

TRANSCRIPTIONS Created by **Kaldi Balsac**

Filter entities by worker version ▾

Le **_DATE** douze janvier mil neuf cent quatre ✓ ✕ , nous prêtre curé soussigné , avons baptisé **_PRENOM** Joseph Pierre Elibien ✓ ✕ , né le **_DATE** même jour ✓ ✕ , fils légitime de **_PERSONNE** Charles Dallaire ✓ ✕ , **_PROFESSION** cultivateur ✓ ✕ , et de **_PERSONNE** Alice Marquis ✓ ✕ , de **_LIEU_RESIDENCE** cette paroisse ✓ ✕ .

Parrain : Pierre Gagnon , ferblantier soussigné ; marraine : Amanda Chaloux son épouse , aussi soussignée , ainsi que le père . Lecture faite .

J . Gagnon

[View source image](#) [List elements on this image](#)

Projet Balsac : serveur IIF

Images

10 centres
36 districts
1 985 paroisses
44 742 registres
1 995 646 images

Format JPEG2000
Stockage AWS S3

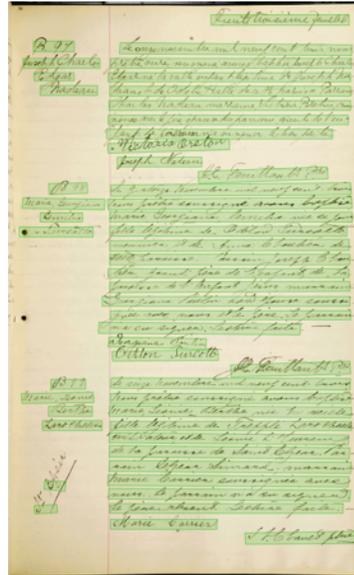
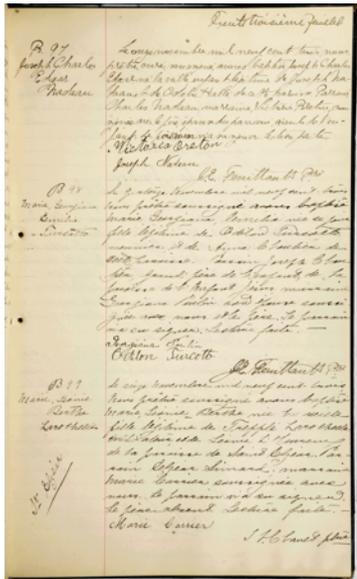
Choix de Cantaloupe en 2019

- support AWS S3 ✓
- support JP2000 ✓
- très versatile (cache, support de multiples formats d'images)
- assez performant

Maintenant ?

- Consommation de RAM importante
- Migration en stockage local car S3 trop lent
- Conversion des images en JPEG pour réduire le stockage et éviter les conversions par le serveur

Projet Balsac : Stratégies d'accès aux images



Détection des lignes de texte :

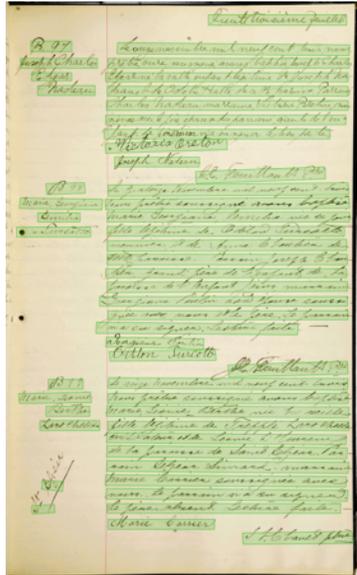
- Prédiction par réseaux de neurones profonds
- Besoin d'accéder à l'image de page complète

Mais pas besoin de travailler sur l'image en pleine taille car le réseau la réduit

- temps moyen de download en full size : 6.96 sec
- temps moyen de download en taille réduite : 0.70 sec

https://iiif.teklia.com/main/iiif/2/balsac-jpg%2FDN0753%2F03Q_CE306S24%2F1903%2F03Q_CE306S24_1903_067.jpg/full/470,768/0/default.jpg - size=470x768

Projet Balsac : Stratégies d'accès aux images



Le [onze novembre mil neuf cent trois](#), nous prêtre, curé soussigné, avons baptisé [Joseph Charles Edgar](#) né [la veille](#), enfant légitime de [Joseph Nadeau](#) et de [Odibe Hallé](#), de cette paroisse. Parrain [Charles Nadeau](#), marraine [Victoria Bleton](#), vu n'a signé avec le père époux du parrain, Etaients de l'en - Vant. Le parrain n'a su signer. Lechoire faite. J. Victoria Arston Joseph Nadeau J. E. Feuilteault Ptre Le [quatorze Novembre mil neuf cent trois](#) nous prêtre soussigné avons baptisé [Marie Georgiana](#), Emilia née [ce jour](#) fille légitime de [Odilon Turcotte meunier](#), et de [Anna Cloutier](#) de cette paroisse. Parrain [Joseph Cloutier](#), grand père de l'enfant de la paroisse de l'Enfant Jésus; marraine [Georgianna Pinaulin](#) son épouse, soussi gné avec nous. et le père. Le parrain n'a su signer. Lecture faite Georgiana Poulin [Odilon Turcotte](#) J. E. Feultant Ptre Le [seize novembre mil neuf cent trois](#) nous prêtre soussigné avons baptisé [Marie Léonie Berthe](#) née [la veille](#) fille légitime de [Trepplé Larochelle cultivateur](#), et de [Léonie L'Heureux](#) de la paroisse de [Saint Elzéar](#). Par rain [Elzéar Simard](#); marraine [Marie Carrier](#) soussignée avec nous. Le parrain n'a su signer Le père absent. Lecture faite. Marie Carrier J. E. Canuel ptre

Reconnaissance d'écriture:

- Prédiction par réseaux de neurones profonds
- Besoin d'accéder aux images de ligne

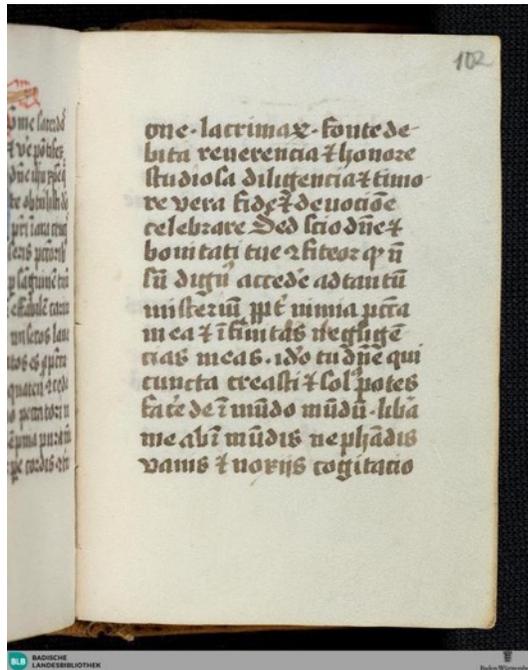
Mais l'utilisation des zones IIF

- est trop lente car une requête par ligne
- surcharge les serveurs

Téléchargement de l'image complète et découpage en ligne en local

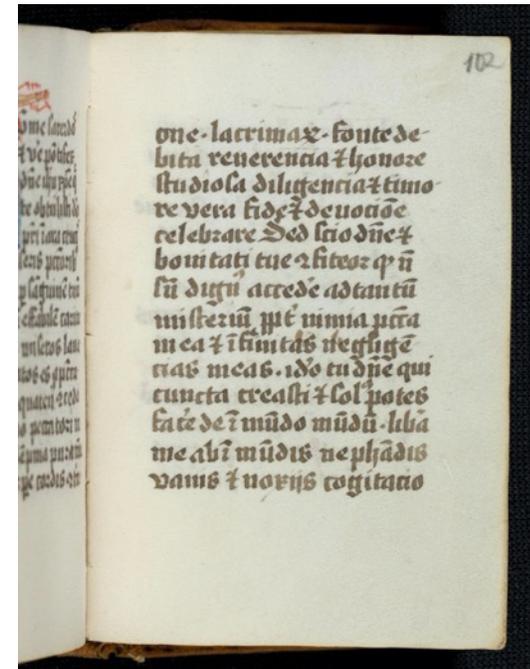
IIIF : quelques surprises

resized



<https://digital.blb-karlsruhe.de/blbhs/i3f/v20/1078658/full/604,767/0/default.jpg>

full



<https://digital.blb-karlsruhe.de/blbhs/i3f/v20/1078658/full/full/0/default.jpg>

IIIF : quelques surprises



L'image est de taille 5069×3616 selon info.json

Mais l'image téléchargée en full/full/0/default.jpg est de taille ... 4000×2853

Car le serveur limite la taille maximale des images en précisant `maxWidth` / `maxHeight`

Sauf quand le serveur ne le précise pas... (mauvaise configuration)

<https://adore.ugent.be/IIIF/images/archive.ugent.be%3A7F0C4994-C579-11E7-8646-155E6EE4309A%3ADS.50/full/full/0/default.jpg>

<https://adore.ugent.be/IIIF/images/archive.ugent.be%3A7F0C4994-C579-11E7-8646-155E6EE4309A%3ADS.50/info.json>

IIIF : quelques surprises

- Migration des serveurs IIIF vers HTTPS pas toujours effectuée : les navigateurs ne chargent pas les images
- Dans les manifests, les `strings` ne sont pas forcément des `strings`
`"label": [{"@value": "Titel"}, {"@value": "書名"}, {"@value": "Title"}]`
- Manifests non à jour (image en erreur, images supprimée)
- Mauvais arrondi au redimensionnement de certaines images : on demande 1024×1024, on obtient 1024×1025

SYNTHEsys+ : 30 institutions européennes

SYNTHEsys+ / Folder test / Page 1

The screenshot displays the SYNTHEsys+ interface. On the left, there is a 'Filter by type...' dropdown and a list of 'Text line' items from 1 to 22. The main area shows a specimen image of a yellow moth, *Euproctis aureoplaga*, with a yellow selection box. Labels include 'Type', 'Madagascar (Central 2500 ft) Jan-Mar. 1911 F.B. Pratt.', 'Kenrick Coll. Brit Mus. 1928-34.', 'Euproctis aureoplaga Kenrick TYPE.', and 'BMNH(E) 1465295'. A ruler is visible at the bottom.

	Gothenburg Global Biodiversity Centre		Global Biodiversity Information Facility		State Museum of Natural History Stuttgart
	Museum National d'Histoire Naturelle		Royal Museum of Central Africa		Zoologisches Forschungsmuseum Alexander Koenig
	Senckenberg Gesellschaft für Naturforschung		Picturae		Consortium of European Taxonomic Facilities
	Meise Botanic Garden		Finnish Museum of Natural History		Botanischer Garten und Botanisches Museum
	University of Copenhagen		University of Manchester		Naturhistoriska riksmuseet
	Greek Research and Technology Network		Biodiversity Information Standards		National Museum Prague
	Natural History Museum Vienna		Hellenic Centre for Marine Research		Royal Botanic Gardens, Kew
	Royal Botanic Garden Edinburgh		The Natural History Museum		Royal Belgian Institute of Natural Sciences
	Digirati		National Natural History Collections of the Hebrew University of Jerusalem		Smithsonian (hosting the Global Genome Biodiversity Network)
	Global Genome Biodiversity Network		Hungarian Natural History Museum		Museum für Naturkunde
	Naturalis Biodiversity Center		CSIC: Museo Nacional de Ciencias Naturales & Real Jardín Botánico		

Questions ?

kermorvant@tekliia.com

T E K L I A