L’intelligence artificielle au service d’Érudit : réflexions

Atelier ACFAS du 5 mai 2025

De la découvrabilité des contenus savants en français

Vincent Larivière, Marie-Jean Meurs et Émilie Paquin

Philippe Langlais

Mea culpa

- Présentation rose
- Vision d'un informaticien

Contexte

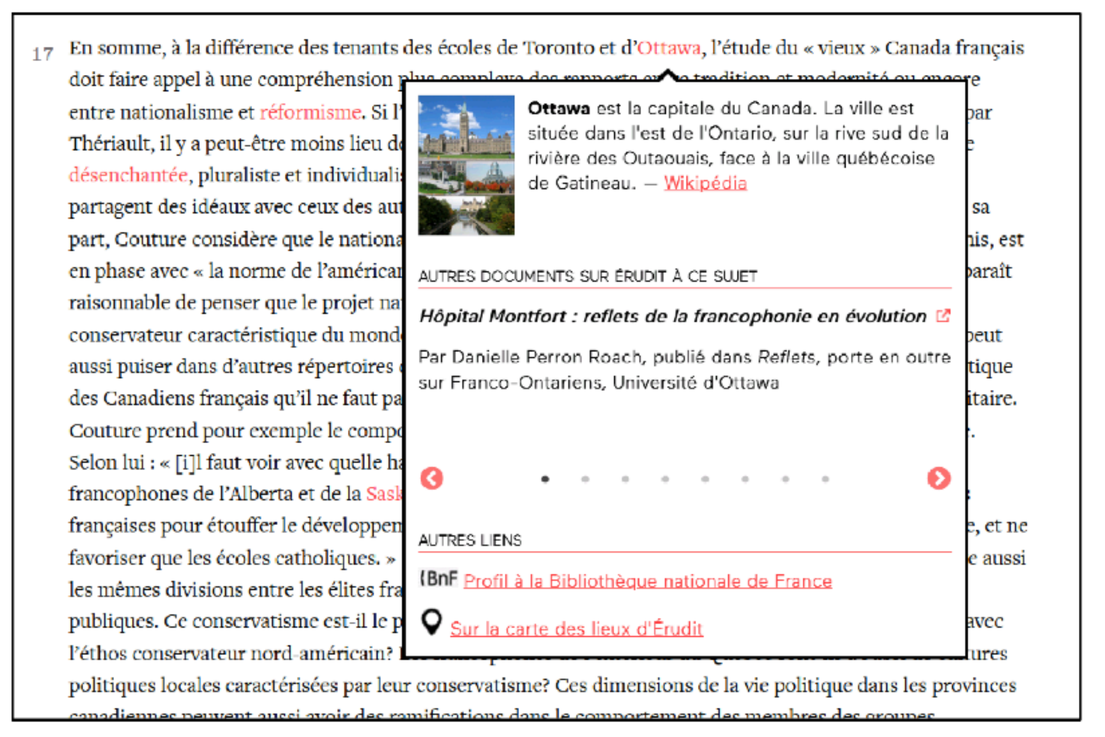
Demande FCI stratégique déposée au concours 2025

- fait suite à la demande FCI pilotée par V. Larivière (CO.SHS)

Contexte

Demande FCI stratégique déposée au concours 2025

- fait suite à la demande FCI pilotée par V. Larivière (CO.SHS)

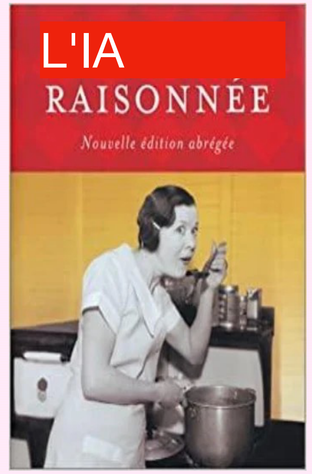


Large-scale Semantic Annotation for Improved Content Discovery in a Digital Library: A Case Study on Érudit, **Gotti et al.**, 2022

Contexte

Demande FCI stratégique déposée au concours 2025

- fait suite à la demande FCI pilotée par V. Larivière (CO.SHS)
- **objet**: intégration **raisonnée** de l'IA au sein d'Érudit



2 porteurs, une équipe

T. Niemann
Érudit

P. Langlais
UdeM/DIRO

S. Beth
Érudit

D. Baragiotta
Érudit

S. van Bellen
Érudit

L. Céspedes
Érudit

L. Bowker
ULaval

A. Rigouts Terryn
UdeM/Ling.

R. Bawden
INRIA

D. Tcheouali
UQAM

JY. Nie
UdeM/DIRO

P. Bellot
U. Aix-Marseille

V. Larivière
UdeM/EBSI

F. Saddat
UQAM

F. Yvon
U. Sorbonne

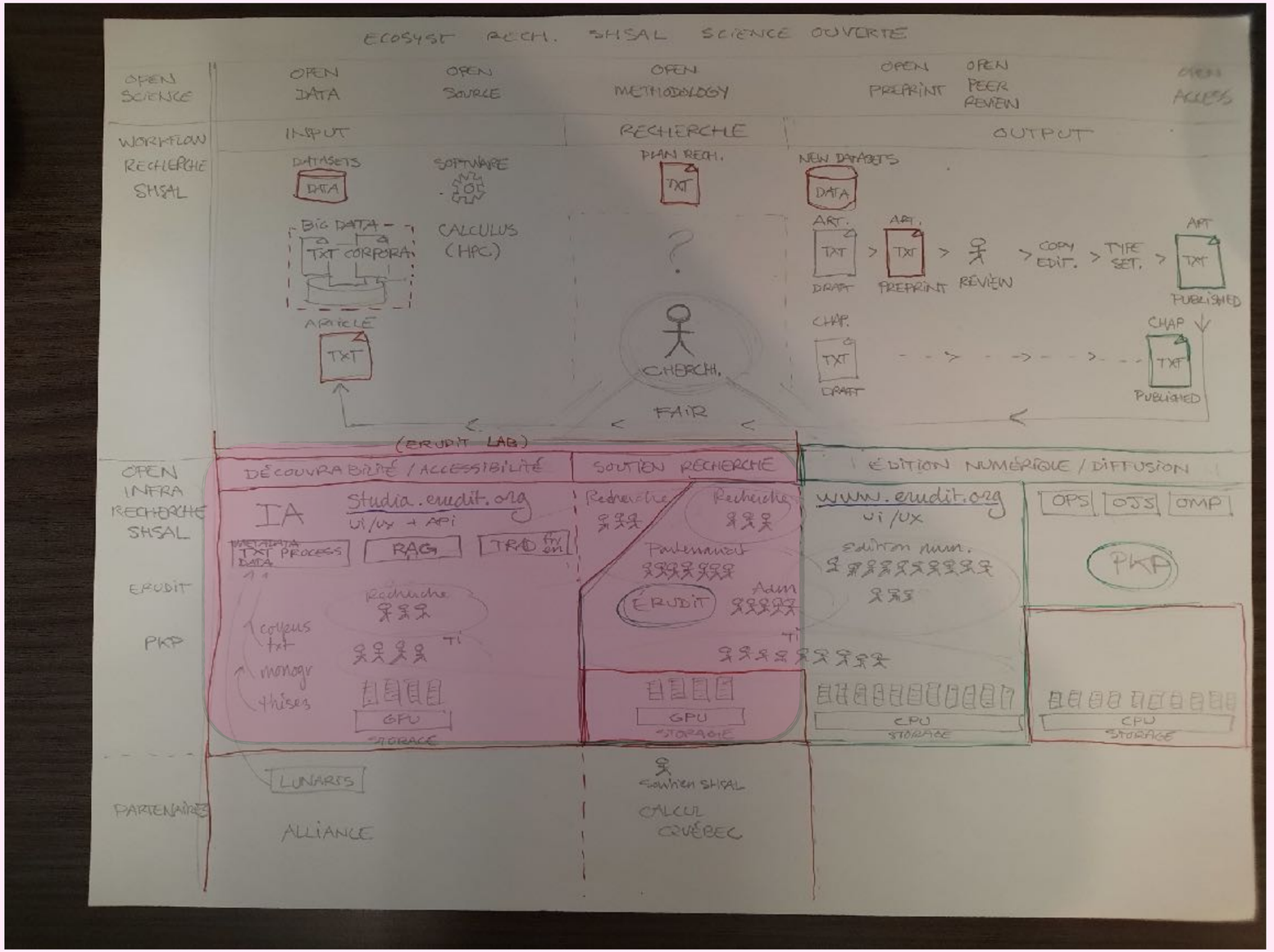
M. Simard
CNRC

B. Liu
UdeM/DIRO

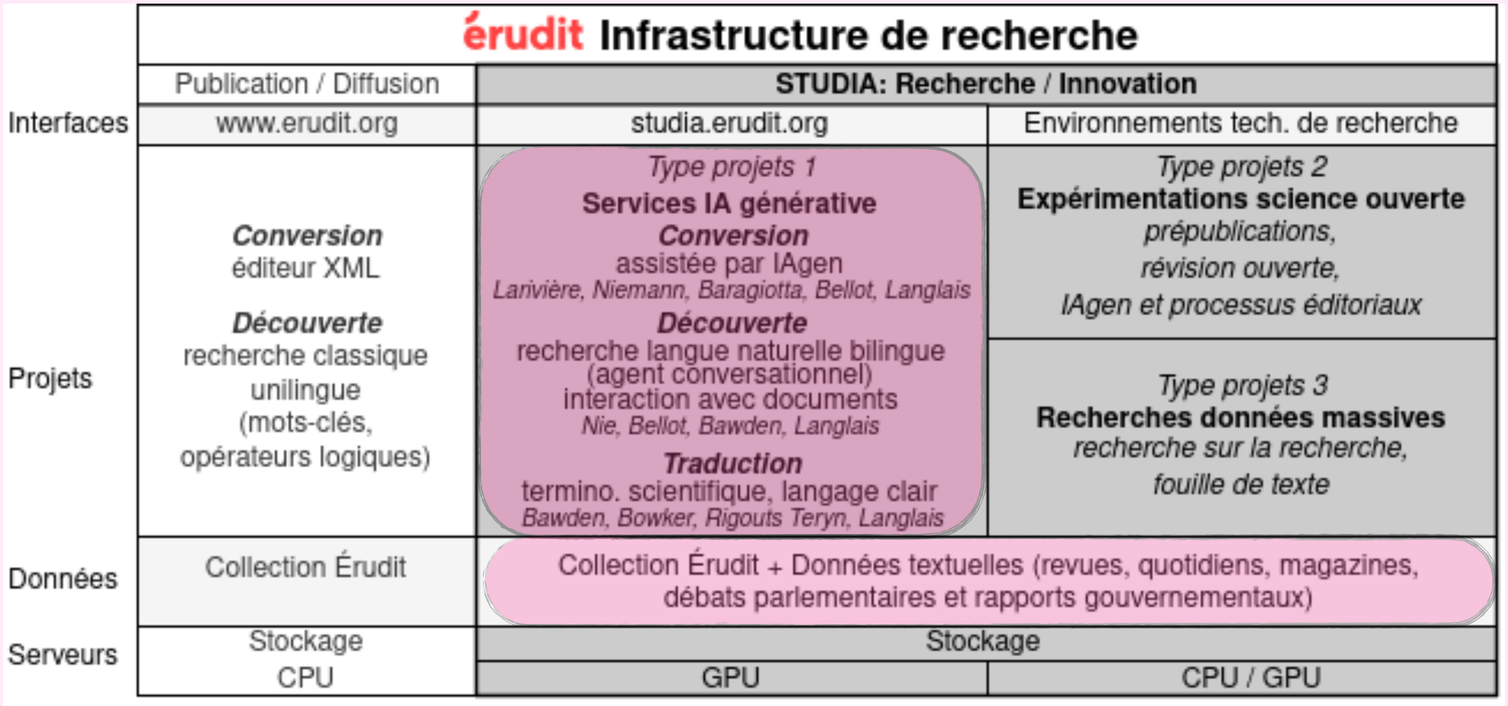
P. Drouin
UdeM/Ling.

Ivado **MILA** **Druide** **BTC** **LexRock AI**

En un schéma


ECOSYST RECH. SHSAL SCIENCE OUVERTE
OPEN SCIENCE
OPEN DATA
OPEN SOURCE
OPEN METHODOLOGY
OPEN PREPRINT
OPEN PEER REVIEW
OPEN ACCESS
WORKFLOW RECHERCHE SHSAL
INPUT
RECHERCHE
OUTPUT
DATASETS
DATA
SOFTWARE
PLAN RECH.
TXT
NEW DATASETS
DATA
BIG DATA
TXT CORPORA
CALCULUS (HPC)
ART.
TXT
DRAFT
API.
PREPRINT
REVIEW
COPY EDIT.
TYPE SET.
ART
PUBLISHED
ARTICLE
CHAP.
CHERCH.
FAIR
(ERUDIT LAB)
OPEN INFRA RECHERCHE SHSAL
ÉRUDIT
PKP
PARTENAIRES
DÉCOUVRABILITÉ / ACCESSIBILITÉ
SOUTIEN RECHERCHE
ÉDITION NUMÉRIQUE / DIFFUSION
IA
Studia.erudit.org
ui/ux + API
RAG
Recherche
Partenariats
ÉRUDIT
www.erudit.org
ui/UX
OPS
OJS
OMP
PKP
GPU
STORAGE
CPU
LUNARIS
ALLIANCE
CALCUL QUÉBEC

En un schéma


érudit Infrastructure de recherche
Publication / Diffusion
STUDIA: Recherche / Innovation
Interfaces
www.erudit.org
studia.erudit.org
Environnements tech. de recherche
Projets
Conversion
éditeur XML
Découverte
recherche classique
unilingue
(mots-clés,
opérateurs logiques)
Type projets 1
Services IA générative
Conversion
assistée par IAgen
Larivière, Niemann, Baragiotta, Bellot, Langlais
Découverte
recherche langue naturelle bilingue
(agent conversationnel)
interaction avec documents
Nie, Bellot, Bawden, Langlais
Traduction
termino. scientifique, langage clair
Bawden, Bowker, Rigouts Teryn, Langlais
Type projets 2
Expérimentations science ouverte
prépublications,
révision ouverte,
IAgen et processus éditoriaux
Type projets 3
Recherches données massives
recherche sur la recherche,
fouille de texte
Données
Collection Érudit
Collection Érudit + Données textuelles (revues, quotidiens, magazines,
débats parlementaires et rapports gouvernementaux)
Serveurs
Stockage
CPU
Stockage
GPU
CPU / GPU

4 volets

- extraction de **méta-données**
- **traduction**
- interrogation fluide de la collection (**chatbot**)
- **valorisation** des données

1- Extraction de méta-données

Métadonnées

tribune

titre

auteure

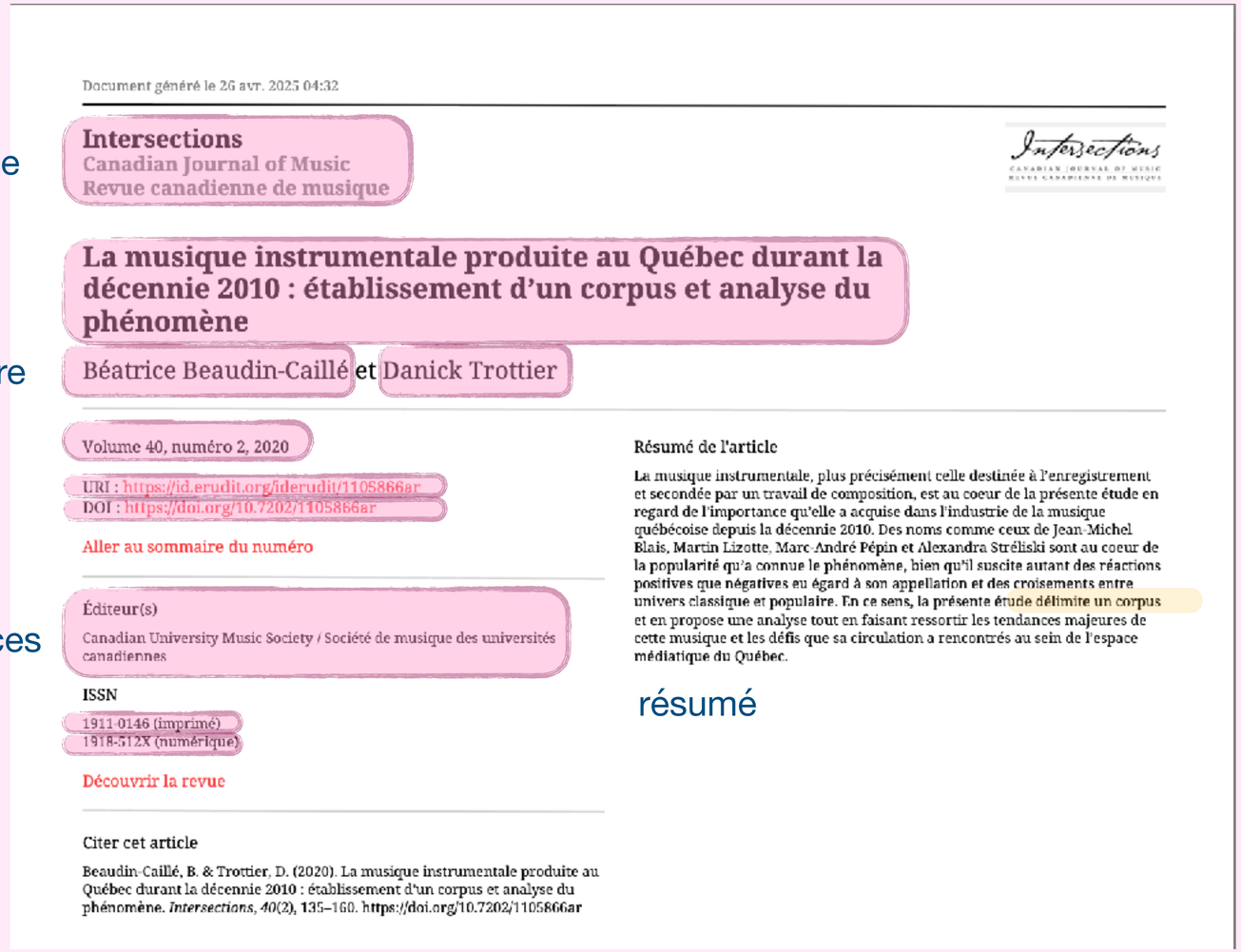
éditrices

entités

concepts

résumé

org. subv.
affiliations



Intersections
Canadian Journal of Music
Revue canadienne de musique

La musique instrumentale produite au Québec durant la décennie 2010 : établissement d'un corpus et analyse du phénomène

Béatrice Beaudin-Caillé et Danick Trottier

Volume 40, numéro 2, 2020

URI : https://id.erudit.org/iderudit/1105866ar
DOI : https://doi.org/10.7202/1105866ar

Éditeur(s)
Canadian University Music Society / Société de musique des universités canadiennes

ISSN
1911-0146 (imprimé)
1918-512X (numérique)

Résumé de l'article

La musique instrumentale, plus précisément celle destinée à l'enregistrement et secondée par un travail de composition, est au coeur de la présente étude en regard de l'importance qu'elle a acquise dans l'industrie de la musique québécoise depuis la décennie 2010. Des noms comme ceux de Jean-Michel Blais, Martin Lizotte, Marc-André Pépin et Alexandra Stréliski sont au coeur de la popularité qu'a connue le phénomène, bien qu'il suscite autant des réactions positives que négatives eu égard à son appellation et des croisements entre univers classique et populaire. En ce sens, la présente étude délimite un corpus et en propose une analyse tout en faisant ressortir les tendances majeures de cette musique et les défis que sa circulation a rencontrés au sein de l'espace médiatique du Québec.

LA MUSIQUE INSTRUMENTALE PRODUITE AU QUÉBEC DURANT LA DÉCENNIE 2010 : ÉTABLISSEMENT D'UN CORPUS ET ANALYSE DU PHÉNOMÈNE[1]

Béatrice Beaudin-Caillé et Danick Trottier

La présente étude se penche sur la musique instrumentale et son ascension depuis les années 2010, plus précisément celle enregistrée et découlant d'un travail original de composition. Pour mieux circonscrire le phénomène, notre analyse se penche sur la situation québécoise et son industrie de la musique. Des noms s'imposent d'entrée de jeu, comme ceux d'Alexandra Stréliski et Jean-Michel Blais, auxquels on peut en ajouter d'autres comme ceux de Martin Lizotte et Marc-André Pépin, puis d'autres encore comme Tambour et Flore Laurentienne. Si cette musique instrumentale que l'on nomme *néoclassique* dans les médias québécois — appellation qui ne va pas sans problème et nous y reviendrons plus loin — n'est pas nouvelle et que des musiciens occidentaux de réputation internationale peuvent y être associés, qu'on pense à Ludovico Einaudi, Yann Tiersen ou encore Chilly Gonzales, sa montée en popularité au Québec est un fait qui mérite une attention particulière pour en comprendre les tenants et aboutissants. Et bien que la musique sans paroles ait toujours existé dans les musiques populaires au Québec, que l'on pense aux danses folkloriques comme la gigue, aux œuvres pour piano d'un André Gagnon ou encore aux pièces d'un groupe comme Harmonium, il n'en reste pas moins que cette musique instrumentale forme un genre à part entière à travers le rôle joué par un instrument comme le piano, la diffusion médiatique qui en délimite la réception (i.e. l'idée de *néoclassique*) ainsi que le contexte particulier des années 2010 dans lequel elle s'inscrit.

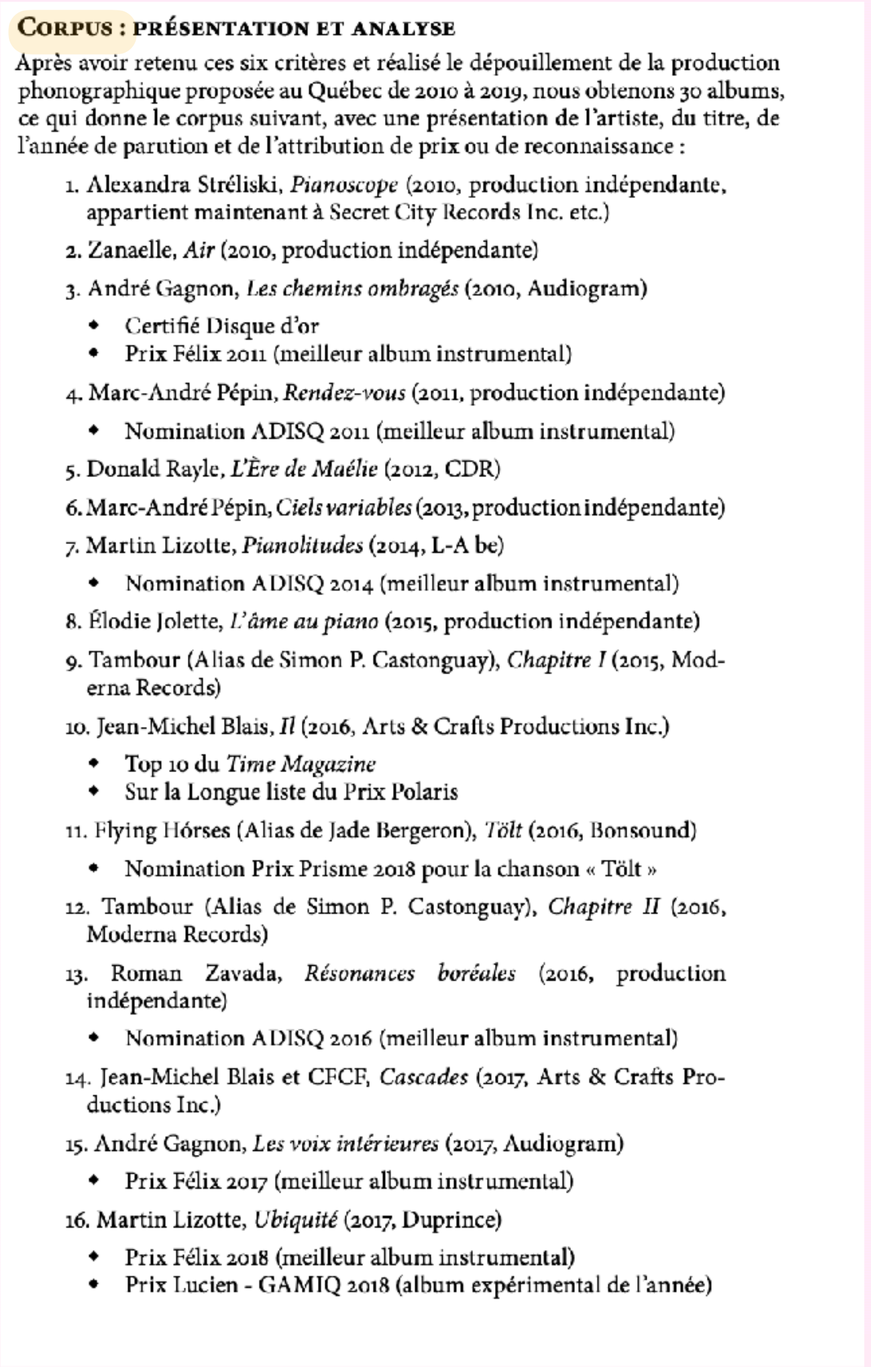
Les prochaines lignes s'attelleront à cette tâche en établissant un corpus d'albums en fonction de six critères qui seront étayés puis justifiés. Il s'ensuivra une analyse pour comprendre la portée de ce corpus dans l'espace culturel du Québec. Par la suite, la réflexion s'attardera autant à la couverture médiatique accordée au phénomène qu'aux faits venant confirmer sa popularité dans

[1] Le travail de Béatrice Beaudin-Caillé pour la production du présent article a été possible grâce à la subvention Soutien à la recherche pour la relève professorale du FRQ-SC accordée à Danick Trottier (UQAM) de 2018 à 2021.

Méta-données

Un corpus de 30 artistes musiciens a été colligé selon 6 critères:

- musique sans parole (ou en fond)
- affiliation aux musiques classiques (ex: néo-classique)
- diffusé autrement que sur tube
- nature de la proposition artistique (ex: pas de reprise)
- lieu à partir duquel l'oeuvre est produite et diffusée (au Québec)
- temporalité (2010-2019)



Corpus : présentation et analyse

Après avoir retenu ces six critères et réalisé le dépouillement de la production phonographique proposée au Québec de 2010 à 2019, nous obtenons 30 albums, ce qui donne le corpus suivant, avec une présentation de l'artiste, du titre, de l'année de parution et de l'attribution de prix ou de reconnaissance :

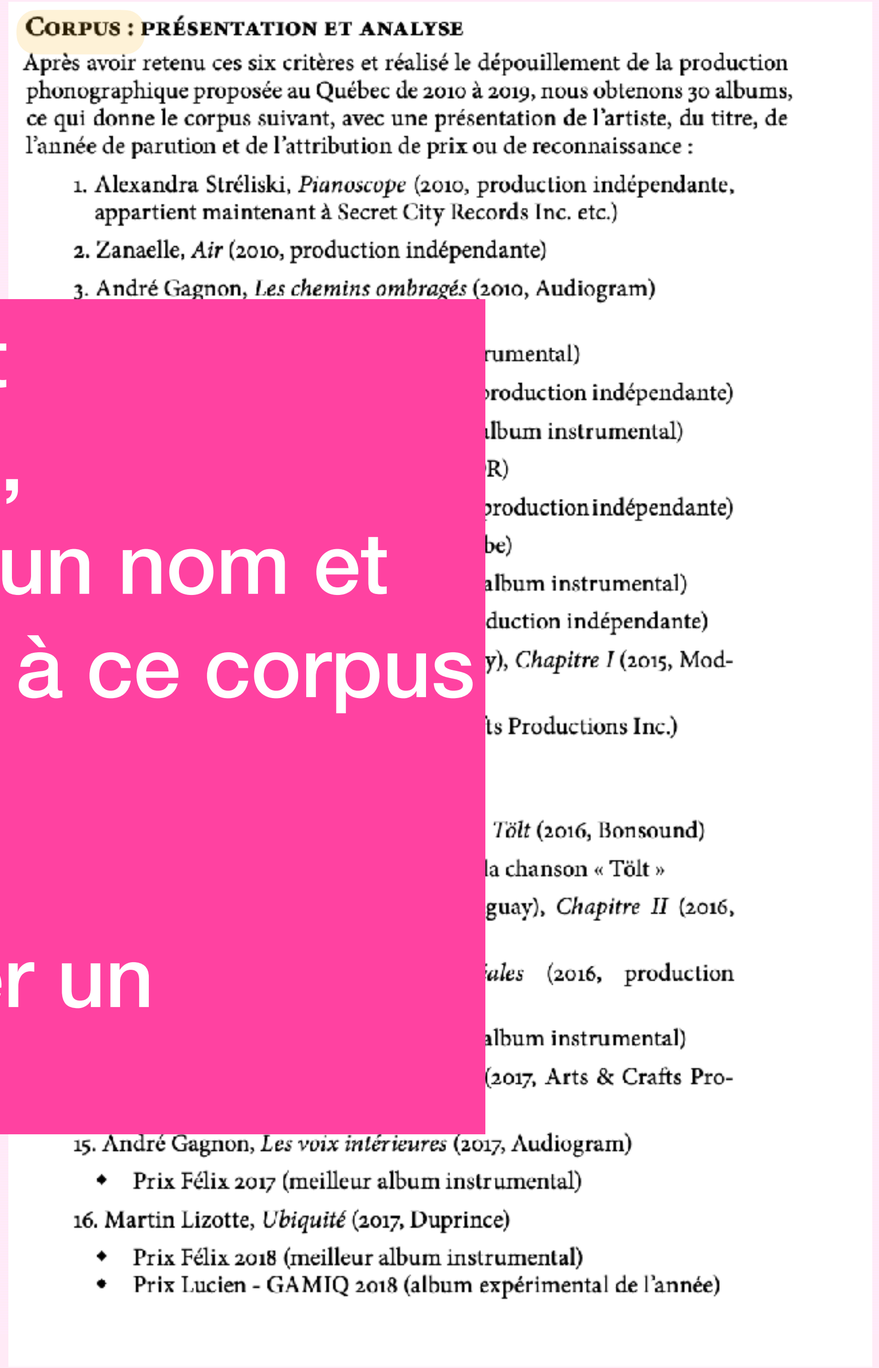
1. Alexandra Stréliski, *Pianoscope* (2010, production indépendante, appartient maintenant à Secret City Records Inc. etc.)
2. Zanaelle, *Air* (2010, production indépendante)
3. André Gagnon, *Les chemins ombragés* (2010, Audiogram)
   - Certifié Disque d'or
   - Prix Félix 2011 (meilleur album instrumental)
4. Marc-André Pépin, *Rendez-vous* (2011, production indépendante)
   - Nomination ADISQ 2011 (meilleur album instrumental)
5. Donald Rayle, *L'Ère de Maélie* (2012, CDR)
6. Marc-André Pépin, *Ciels variables* (2013, production indépendante)
7. Martin Lizotte, *Pianolitudes* (2014, L-A be)
   - Nomination ADISQ 2014 (meilleur album instrumental)
8. Élodie Jolette, *L'âme au piano* (2015, production indépendante)
9. Tambour (Alias de Simon P. Castonguay), *Chapitre I* (2015, Moderna Records)
10. Jean-Michel Blais, *Il* (2016, Arts & Crafts Productions Inc.)
    - Top 10 du *Time Magazine*
    - Sur la Longue liste du Prix Polaris
11. Flying Hórses (Alias de Jade Bergeron), *Tölt* (2016, Bonsound)
    - Nomination Prix Prisme 2018 pour la chanson « Tölt »
12. Tambour (Alias de Simon P. Castonguay), *Chapitre II* (2016, Moderna Records)
13. Roman Zavada, *Résonances boréales* (2016, production indépendante)
    - Nomination ADISQ 2016 (meilleur album instrumental)
14. Jean-Michel Blais et CFCF, *Cascades* (2017, Arts & Crafts Productions Inc.)
15. André Gagnon, *Les voix intérieures* (2017, Audiogram)
    - Prix Félix 2017 (meilleur album instrumental)
16. Martin Lizotte, *Ubiquité* (2017, Duprince)
    - Prix Félix 2018 (meilleur album instrumental)
    - Prix Lucien - GAMIQ 2018 (album expérimental de l'année)

Méta-données

Un corpus de 30 artistes musiciens a été colligé selon 6 c

- musiqu
- affiliati
  classic
- diffusé
- nature
  reprise
- lieu à partir duquel l'œuvre est produite et diffusée (au Québec)
- temporalité (2010-2019)

- Identifier ce corpus permettrait
  1) d'affiner l'indexation (en aval),
  2) de suggérer aux auteures qu'un nom et un espace pourrait être associé à ce corpus (en amont).
- L'IA pourrait également générer un "résumé" de ce corpus



Corpus : présentation et analyse

Après avoir retenu ces six critères et réalisé le dépouillement de la production phonographique proposée au Québec de 2010 à 2019, nous obtenons 30 albums, ce qui donne le corpus suivant, avec une présentation de l'artiste, du titre, de l'année de parution et de l'attribution de prix ou de reconnaissance :

1. Alexandra Stréliski, *Pianoscope* (2010, production indépendante, appartient maintenant à Secret City Records Inc. etc.)
2. Zanaelle, *Air* (2010, production indépendante)
3. André Gagnon, *Les chemins ombragés* (2010, Audiogram)

rumental)
roduction indépendante)
lbum instrumental)
R)
roduction indépendante)
be)
lbum instrumental)
duction indépendante)
y), *Chapitre I* (2015, Mod-
ts Productions Inc.)

Tölt (2016, Bonsound)
la chanson « Tölt »
guay), *Chapitre II* (2016,
ales (2016, production
lbum instrumental)
(2017, Arts & Crafts Pro-

15. André Gagnon, *Les voix intérieures* (2017, Audiogram)
    - Prix Félix 2017 (meilleur album instrumental)
16. Martin Lizotte, *Ubiquité* (2017, Duprince)
    - Prix Félix 2018 (meilleur album instrumental)
    - Prix Lucien - GAMIQ 2018 (album expérimental de l'année)

Où en est la science ?

- Un *benchmark* de 500K pages d'articles de arXiv (multi-domaines, STEM)

Publication[1]	Size	Ground-truth Labels
Fan [16]	147 documents	Metadata
Färber [17]	90K documents	References
Grennan [21]	1B references	References
Saier [51,50]	1M documents.	References
Ley [30,52]	6M documents	Metadata, references
Mccallum [39]	935 documents	Metadata, references
Kyle [34]	8.1M documents	Metadata, references
Ororbia [43]	6M documents.	Metadata, references
Bast [6]	12,098 documents	Body text, sections, title
Li [31]	500K pages	Captions, equations, figures, footers lists, metadata, paragraphs, references, sections, tables

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents

Norman Meuschke[1,ORCID], Apurva Jagdale[2], Timo Spinde[1,ORCID], Jelena Mitrović[2,3,ORCID], and Bela Gipp[1,ORCID]

[1] University of Göttingen, 37073 Göttingen, Germany
{meuschke, spinde, gipp}@uni-goettingen.de
[2] University of Passau, 94032 Passau, Germany
{apurva.jagdale, jelena.mitrovic}@uni-passau.de
[3] The Institute for Artificial Intelligence R&D of Serbia, 21000 Novi Sad, Serbia

Abstract. Extracting information from academic PDF documents is crucial for numerous indexing, retrieval, and analysis use cases. Choosing the best tool to extract specific content elements is difficult because many, technically diverse tools are available, but recent performance benchmarks are rare. Moreover, such benchmarks typically cover only a few content elements like header metadata or bibliographic references and use smaller datasets from specific academic disciplines. We provide a large and diverse evaluation framework that supports more extraction tasks than most related datasets. Our framework builds upon DocBank, a multi-domain dataset of 1.5M annotated content elements extracted from 500K pages of research papers on arXiv. Using the new framework, we benchmark ten freely available tools in extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. GROBID achieves the best metadata and reference extraction results, followed by CERMINE and Science Parse. For table extraction, Adobe Extract outperforms other tools, even though the performance is much lower than for other content elements. All tools struggle to extract lists, footers, and equations. We conclude that more research on improving and combining tools is necessary to achieve satisfactory extraction quality for most content elements. Evaluation datasets and frameworks like the one we present support this line of research. We make our data and code publicly available to contribute toward this goal.

Keywords: PDF · Information Extraction · Benchmark · Evaluation.

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents, **Meuschke et al.**, 2023 in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS

Où en est la science ?

- Un *benchmark* de 500K pages d'articles

benchmark = tâche(s) + métrique(s)

	Model	Zero-shot	Memory Usage (MB)	Number of Parameters	Embedding Dimensions	Max Tokens	Mean (Task)	Mean (TaskTyp
1	gemini-embedding-exp-03-07	99%	Unknown	Unknown	3072	8192	**68.32**	**59.64**
2	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.21
3	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	56.00
4	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.23	55.17

Ororbia [43]	6M documents.	Metadata, references
Bast [6]	12,098 documents	Body text, sections, title
Li [31]	500K pages	Captions, equations, figures, footers lists, metadata, paragraphs, references, sections, tables

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents, **Meuschke et al.**, 2023 in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents

Norman Meuschke[1,ORCID], Apurva Jagdale[2], Timo Spinde[1,ORCID], Jelena Mitrović[2,3,ORCID], and Bela Gipp[1,ORCID]

[1] University of Göttingen, 37073 Göttingen, Germany
{meuschke, spinde, gipp}@uni-goettingen.de
[2] University of Passau, 94032 Passau, Germany
{apurva.jagdale, jelena.mitrovic}@uni-passau.de
[3] The Institute for Artificial Intelligence R&D of Serbia, 21000 Novi Sad, Serbia

Abstract. Extracting information from academic PDF documents is crucial for numerous indexing, retrieval, and analysis use cases. Choosing the best tool to extract specific content elements is difficult because many, technically diverse tools are available, but recent performance benchmarks are rare. Moreover, such benchmarks typically cover only a few content elements like header metadata or bibliographic references and use smaller datasets from specific academic disciplines. We provide a large and diverse evaluation framework that supports more extraction tasks than most related datasets. Our framework builds upon DocBank, a multi-domain dataset of 1.5M annotated content elements extracted from 500K pages of research papers on arXiv. Using the new framework, we benchmark ten freely available tools in extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. GROBID achieves the best metadata and reference extraction results, followed by CERMINE and Science Parse. For table extraction, Adobe Extract outperforms other tools, even though the performance is much lower than for other content elements. All tools struggle to extract lists, footers, and equations. We conclude that more research on improving and combining tools is necessary to achieve satisfactory extraction quality for most content elements. Evaluation datasets and frameworks like the one we present support this line of research. We make our data and code publicly available to contribute toward this goal.

Keywords: PDF · Information Extraction · Benchmark · Evaluation.

Où en est la science ?

- Un *benchmark* de 500K pages d'articles de arXiv (multi-domaines, STEM)

Publication[1]	Size	Ground-truth Labels
Fan [16]	147 documents	Metadata
Färber [17]	90K documents	References
Grennan [21]	1B references	References
Saier [51,50]	1M documents.	References
Ley [30,52]	6M documents	Metadata, references
Mccallum [39]	935 documents	Metadata, references
Kyle [34]	8.1M documents	Metadata, references
Ororbia [43]	6M documents.	Metadata, references
Bast [6]	12,098 documents	Body text, sections, title
Li [31]	500K pages	Captions, equations, figures, footers lists, metadata, paragraphs, references, sections, tables

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents

Norman Meuschke[1,ORCID], Apurva Jagdale[2], Timo Spinde[1,ORCID], Jelena Mitrović[2,3,ORCID], and Bela Gipp[1,ORCID]

[1] University of Göttingen, 37073 Göttingen, Germany
{meuschke, spinde, gipp}@uni-goettingen.de
[2] University of Passau, 94032 Passau, Germany
{apurva.jagdale, jelena.mitrovic}@uni-passau.de
[3] The Institute for Artificial Intelligence R&D of Serbia, 21000 Novi Sad, Serbia

Abstract. Extracting information from academic PDF documents is crucial for numerous indexing, retrieval, and analysis use cases. Choosing the best tool to extract specific content elements is difficult because many, technically diverse tools are available, but recent performance benchmarks are rare. Moreover, such benchmarks typically cover only a few content elements like header metadata or bibliographic references and use smaller datasets from specific academic disciplines. We provide a large and diverse evaluation framework that supports more extraction tasks than most related datasets. Our framework builds upon DocBank, a multi-domain dataset of 1.5M annotated content elements extracted from 500K pages of research papers on arXiv. Using the new framework, we benchmark ten freely available tools in extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. GROBID achieves the best metadata and reference extraction results, followed by CERMINE and Science Parse. For table extraction, Adobe Extract outperforms other tools, even though the performance is much lower than for other content elements. All tools struggle to extract lists, footers, and equations. We conclude that more research on improving and combining tools is necessary to achieve satisfactory extraction quality for most content elements. Evaluation datasets and frameworks like the one we present support this line of research. We make our data and code publicly available to contribute toward this goal.

Keywords: PDF · Information Extraction · Benchmark · Evaluation.

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents, **Meuschke et al.**, 2023 in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS

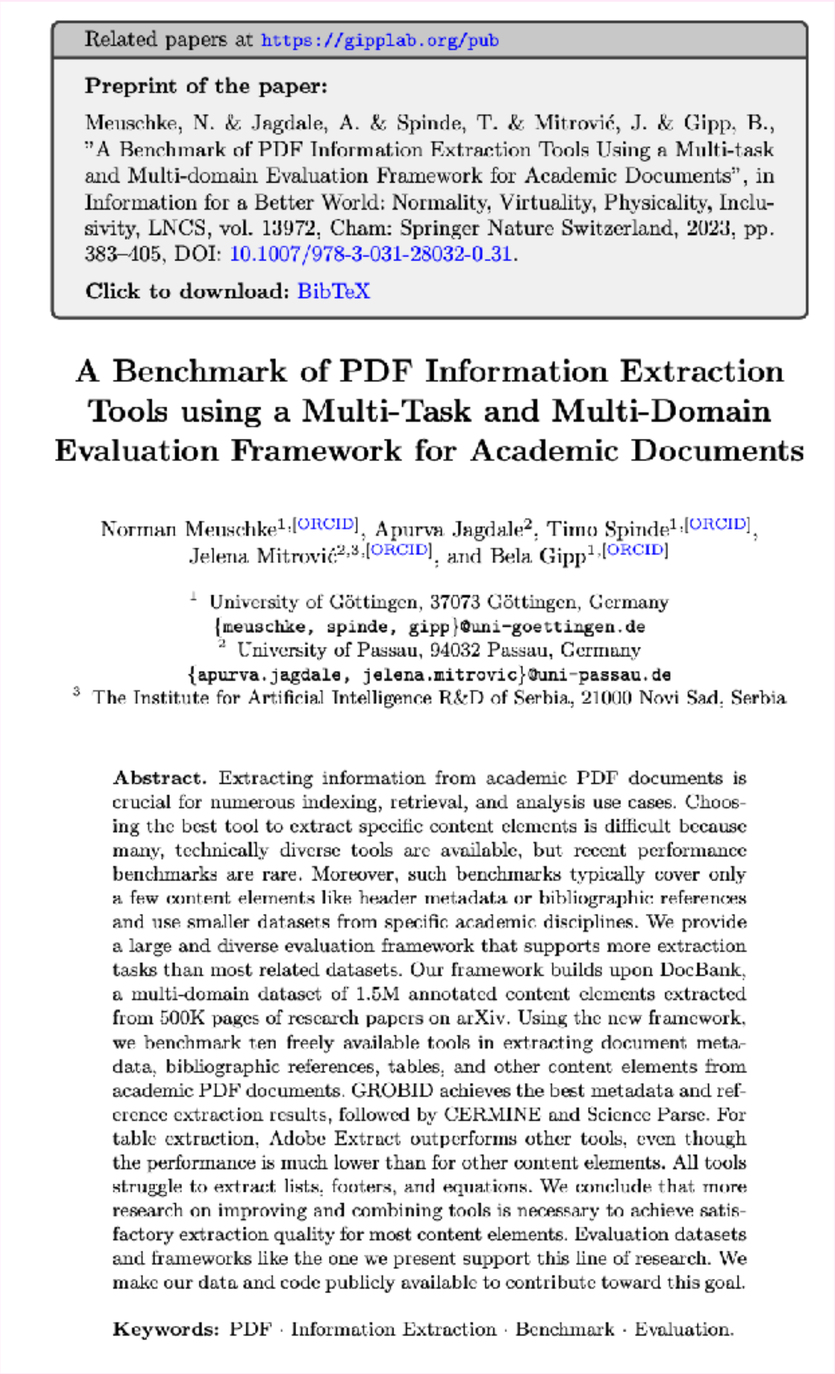
Où en est la science ?

- Évaluation de 10 outils non commerciaux à l'état de l'art capables d'extraire des informations de pdf.

Tool	Version	Task[1]	Technology	Output
Adobe Extract	1.0	G, T	Adobe Sensei AI Framework	JSON, XLSX
Apache Tika	2.0.0	G	Apache PDFBox	TXT
Camelot	0.10.1	T	OpenCV, PDFMiner	CSV, Dataframe
CERMINE	1.13	G, M, R	CRF, iText, Rules, SVM	JATS
GROBID	0.7.0	G, M, R, T	CRF, Deep Learning, Pdfalto	TEI XML
PdfAct	n/a	G, M, R, T	pdftotext, rules	JSON, TXT, XML
PyMuPDF	1.19.1	G	OCR, tesseract	TXT
RefExtract	0.2.5	R	pdftotext, rules	TXT
ScienceParse	1.0	G, M, R,	CRF, pdffigures2, rules	JSON
Tabula	1.2.1	T	PDFBox, rules	CSV, Dataframe

[1] (G) General, (M) Metadata, (R) References, (T) Table

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents, **Meuschke et al.**, 2023 in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS



Related papers at https://gipplab.org/pub

Preprint of the paper:

Meuschke, N. & Jagdale, A. & Spinde, T. & Mitrović, J. & Gipp, B., "A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents", in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS, vol. 13972, Cham: Springer Nature Switzerland, 2023, pp. 383–405, DOI: 10.1007/978-3-031-28032-0_31.

Click to download: BibTeX

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents

Norman Meuschke[1,][ORCID], Apurva Jagdale[2], Timo Spinde[1,][ORCID], Jelena Mitrović[2,3,][ORCID], and Bela Gipp[1,][ORCID]

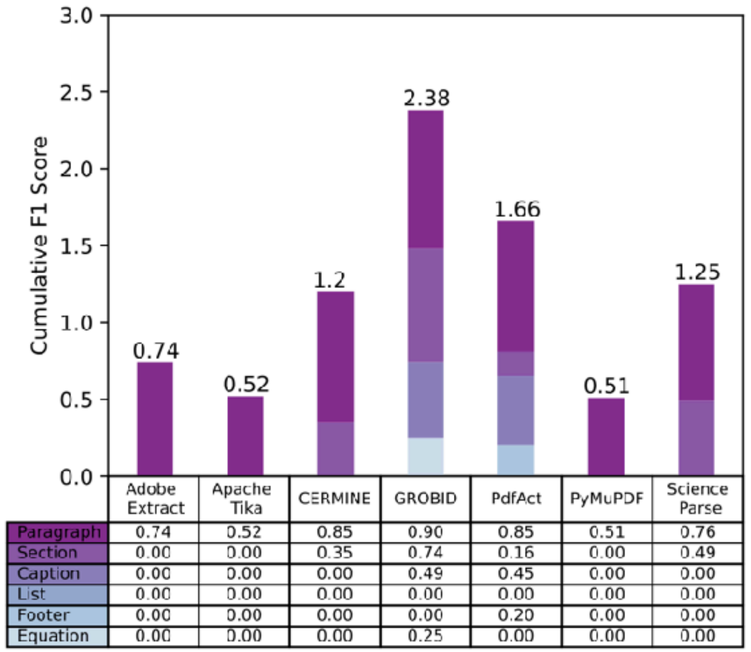
[1] University of Göttingen, 37073 Göttingen, Germany
{meuschke, spinde, gipp}@uni-goettingen.de
[2] University of Passau, 94032 Passau, Germany
{apurva.jagdale, jelena.mitrovic}@uni-passau.de
[3] The Institute for Artificial Intelligence R&D of Serbia, 21000 Novi Sad, Serbia

Abstract. Extracting information from academic PDF documents is crucial for numerous indexing, retrieval, and analysis use cases. Choosing the best tool to extract specific content elements is difficult because many, technically diverse tools are available, but recent performance benchmarks are rare. Moreover, such benchmarks typically cover only a few content elements like header metadata or bibliographic references and use smaller datasets from specific academic disciplines. We provide a large and diverse evaluation framework that supports more extraction tasks than most related datasets. Our framework builds upon DocBank, a multi-domain dataset of 1.5M annotated content elements extracted from 500K pages of research papers on arXiv. Using the new framework, we benchmark ten freely available tools in extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. GROBID achieves the best metadata and reference extraction results, followed by CERMINE and Science Parse. For table extraction, Adobe Extract outperforms other tools, even though the performance is much lower than for other content elements. All tools struggle to extract lists, footers, and equations. We conclude that more research on improving and combining tools is necessary to achieve satisfactory extraction quality for most content elements. Evaluation datasets and frameworks like the one we present support this line of research. We make our data and code publicly available to contribute toward this goal.

Keywords: PDF · Information Extraction · Benchmark · Evaluation.

Où en est la science ?

- un protocole d'évaluation (métriques) et des résultats



	Adobe Extract	Apache Tika	CERMINE	GROBID	PdfAct	PyMuPDF	Science Parse
Paragraph	0.74	0.52	0.85	0.90	0.85	0.51	0.76
Section	0.00	0.00	0.35	0.74	0.16	0.00	0.49
Caption	0.00	0.00	0.00	0.49	0.45	0.00	0.00
List	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Footer	0.00	0.00	0.00	0.00	0.20	0.00	0.00
Equation	0.00	0.00	0.00	0.25	0.00	0.00	0.00

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents, **Meuschke et al.**, 2023 in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS

Related papers at https://gipplab.org/pub

Preprint of the paper:

Meuschke, N. & Jagdale, A. & Spinde, T. & Mitrović, J. & Gipp, B., "A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents", in Information for a Better World: Normality, Virtuality, Physicality, Inclusivity, LNCS, vol. 13972, Cham: Springer Nature Switzerland, 2023, pp. 383–405, DOI: 10.1007/978-3-031-28032-0_31.

Click to download: BibTeX

A Benchmark of PDF Information Extraction Tools using a Multi-Task and Multi-Domain Evaluation Framework for Academic Documents

Norman Meuschke[1,ORCID], Apurva Jagdale[2], Timo Spinde[1,ORCID], Jelena Mitrović[2,3,ORCID], and Bela Gipp[1,ORCID]

[1] University of Göttingen, 37073 Göttingen, Germany
{meuschke, spinde, gipp}@uni-goettingen.de
[2] University of Passau, 94032 Passau, Germany
{apurva.jagdale, jelena.mitrovic}@uni-passau.de
[3] The Institute for Artificial Intelligence R&D of Serbia, 21000 Novi Sad, Serbia

Abstract. Extracting information from academic PDF documents is crucial for numerous indexing, retrieval, and analysis use cases. Choosing the best tool to extract specific content elements is difficult because many, technically diverse tools are available, but recent performance benchmarks are rare. Moreover, such benchmarks typically cover only a few content elements like header metadata or bibliographic references and use smaller datasets from specific academic disciplines. We provide a large and diverse evaluation framework that supports more extraction tasks than most related datasets. Our framework builds upon DocBank, a multi-domain dataset of 1.5M annotated content elements extracted from 500K pages of research papers on arXiv. Using the new framework, we benchmark ten freely available tools in extracting document metadata, bibliographic references, tables, and other content elements from academic PDF documents. GROBID achieves the best metadata and reference extraction results, followed by CERMINE and Science Parse. For table extraction, Adobe Extract outperforms other tools, even though the performance is much lower than for other content elements. All tools struggle to extract lists, footers, and equations. We conclude that more research on improving and combining tools is necessary to achieve satisfactory extraction quality for most content elements. Evaluation datasets and frameworks like the one we present support this line of research. We make our data and code publicly available to contribute toward this goal.

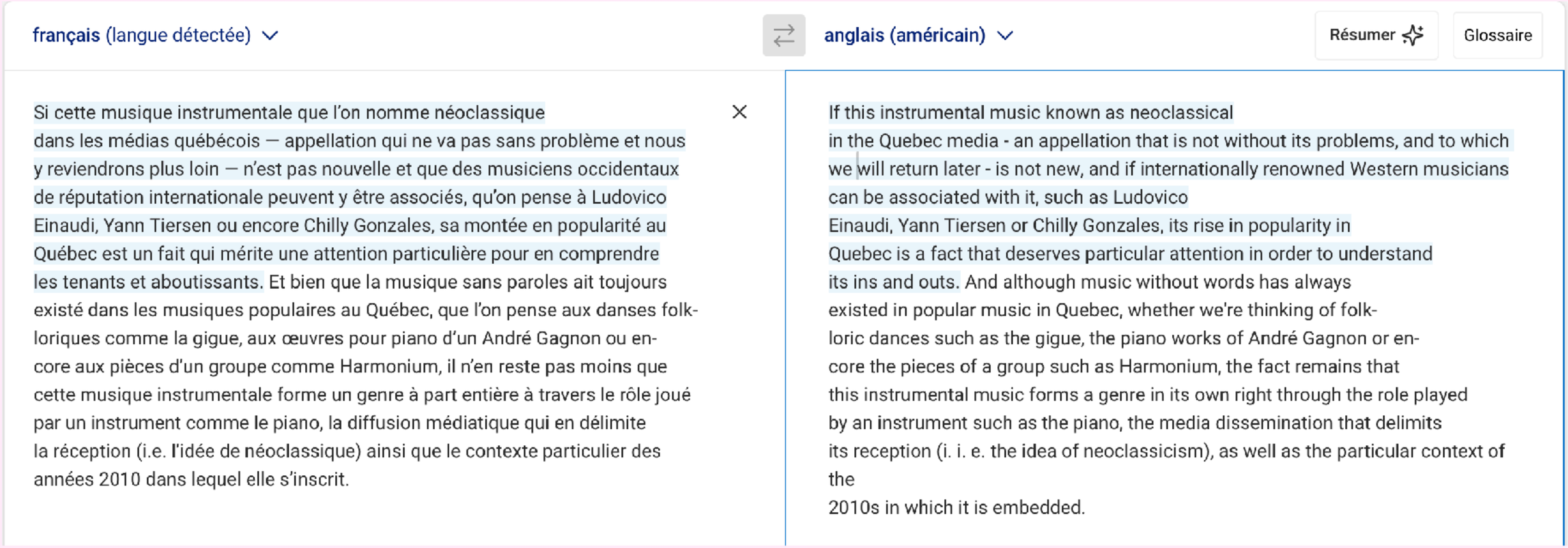
Keywords: PDF · Information Extraction · Benchmark · Evaluation.

Objectifs

- développer un *benchmark* en SHS
  - capitalisant sur les annotations fines déjà apposées sur les articles par l'équipe d'Érudit
- test de différentes technologies
  - GROBID, GML, etc.
- élaboration d'une chaîne d'extraction de métadonnées
  - utile aux annotatrices d'Érudit

2- Traduction automatique

a) Traduction fr<->en



produite par DeepL.com (version libre) le 26 avril 2025

Traduction fr-en

Erreurs potentielles

If this instrumental music known as neoclassical in the Quebec media - an appellation that is not without its problems, and to which we will return later - is not new, and if internationally renowned Western musicians can be associated with it, such as Ludovico Einaudi, Yann Tiersen or Chilly Gonzales, its rise in popularity in Quebec is a fact that deserves particular attention in order to understand its ins and outs. And although music without words has always existed in popular music in Quebec, whether we're thinking of folkloric dances such as the gigue, the piano works of André Gagnon or encore the pieces of a group such as Harmonium, the fact remains that this instrumental music forms a genre in its own right through the role played by an instrument such as the piano, the media dissemination that delimits its reception (i. i. e. the idea of neoclassicism), as well as the particular context of the 2010s in which it is embedded.

Enjeux

Traduire automatiquement peut être aussi simple que:

```
en_fr_translator = pipeline("translation_en_to_fr")
en_fr_translator("How old are you?")
```

librairie *transformers* de HugginFace

Mais:

- **Acceptabilité** des différentes actrices de l'édition
  - éditrices, auteures, utilisatrices d'Érudit
- Grand Modèles de Langue (GML) versus modèles dédiés (ex: DeepL)
- Conservation du format
- **Évaluation** ?
  - sciences humaines et sociales = enfant pauvre du traitement des langues

Objectifs

- Promouvoir/prévenir la transformation de l'IA générative auprès des éditrices et autrices
  - par consultation des différentes actrices
  - période de probation autorisant le retrait d'un article
- Réaliser un *benchmark* d'articles de SHS et leur traduction
  - façon MATOS (Machine Translation for Open Science)
  - avec l'aide du Bureau de la Traduction du Canada pour:
    - traduire des articles
    - identifier des erreurs fines dans des traductions automatiques
- La traduction augmente t-elle la découvrabilité ?
- La "traduction" en language clair

3- Agent conversationnel

RAG is in the air

Librairie EYROLLES

Rechercher un livre, un auteur, une collection

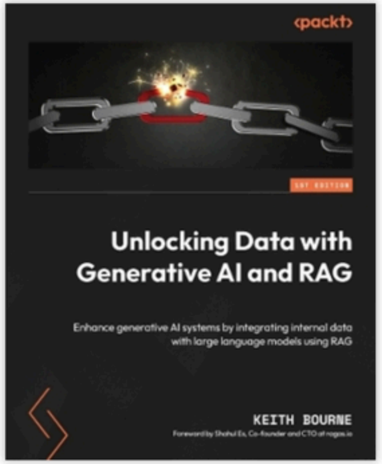
MON COMPTE
Connexion

MON PANIER
0 article

Vous avez entre 15 et 18 ans ? **Profitez du Pass Culture à la librairie Eyrolles !**

TOUS NOS RAYONS

Informatique › Unlocking Data with Generative A...



AJOUTER À UNE LISTE

Librairie Eyrolles - Paris 5e
Disponible en magasin

Unlocking Data with Generative AI and RAG: Enhance generative AI systems by integrating internal dat

Keith / Es Bourne

(0 avis) Donner votre avis

346 pages, parution le 26/09/2024

Livre papier
53,32 €

✓ **Expédié sous 24h**
Livraison à partir de 0,01€ dès 35€ d'achats
Pour une livraison en France métropolitaine

✓ **Retrait à la librairie - Paris 5e**
Disponible dès le 28/04

QUANTITÉ 1

Ajouter au panier

Avantages Eyrolles.com
- **Livraison à partir de 0,01 €** en France métropolitaine
- Paiement en ligne **SÉCURISÉ**
- **Livraison** dans le monde
- **Retour** sous 15 jours
- **+ d'un million et demi** de livres disponibles

Caractéristiques techniques

	PAPIER
Éditeur(s)	Packt publishing
Auteur(s)	Keith / Es Bourne
Parution	26/09/2024
Nb. de pages	346
EAN13	9781835887905

CARACTÉRISTIQUES TECHNIQUES

Service Client Virtuel

DCL
Data Conversion Laboratory Inc.

LEARNING SERIES

Genius Without the Gibberish: How RAG and Structured Content Boost Generative AI Reliability

Wed, Jan 15, 2025 12:00 PM - 1:00 PM EST

We hope you enjoyed our webinar. You can view it on demand by visiting https://www.dataconversionlaboratory.com/courses and selecting "Artificial Intelligence."

RAG is in the air

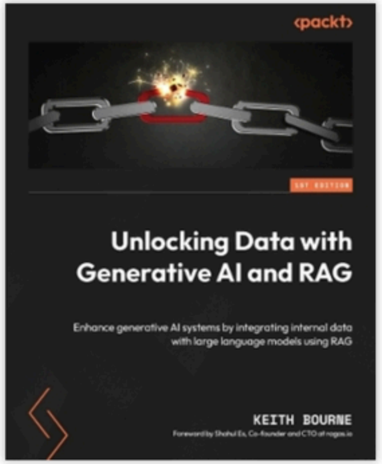
Librairie EYROLLES

Rechercher un livre, un auteur, une collection

MON COMPTE
Connexion

MON PANIER
0 article

TOUS NOS RAYONS



AJOUTER À UNE LISTE

Librairie Eyrolles - Paris 5e
Disponible en magasin

Medium

Search

Level Up Coding

Coding tutorials and news. The developer homepage gitconnected.com && skilled.dev && levelup.dev

Follow publication

Member-only story

Testing 18 RAG Techniques to Find the Best

crag, HyDE, fusion and more!

Fareed Khan · Follow · 36 min read · Mar 11, 2025

1.8K · 27

Read this story for free: link

We are going to start with a simple RAG approach, which we all know, and then test more advanced techniques like CRAG, Fusion, HyDE, and more!

DCL
Data Conversion Laboratory Inc.

ARNING SERIES

ut the Gibberish: How RAG
d Content Boost Generative

PM - 1:00 PM EST

ır webinar. You can view it on demand by visiting https://www.dataconversionlaboratory.com/courses and selecting "Artificial Intelligence."

Service Client Virtuel

RAG

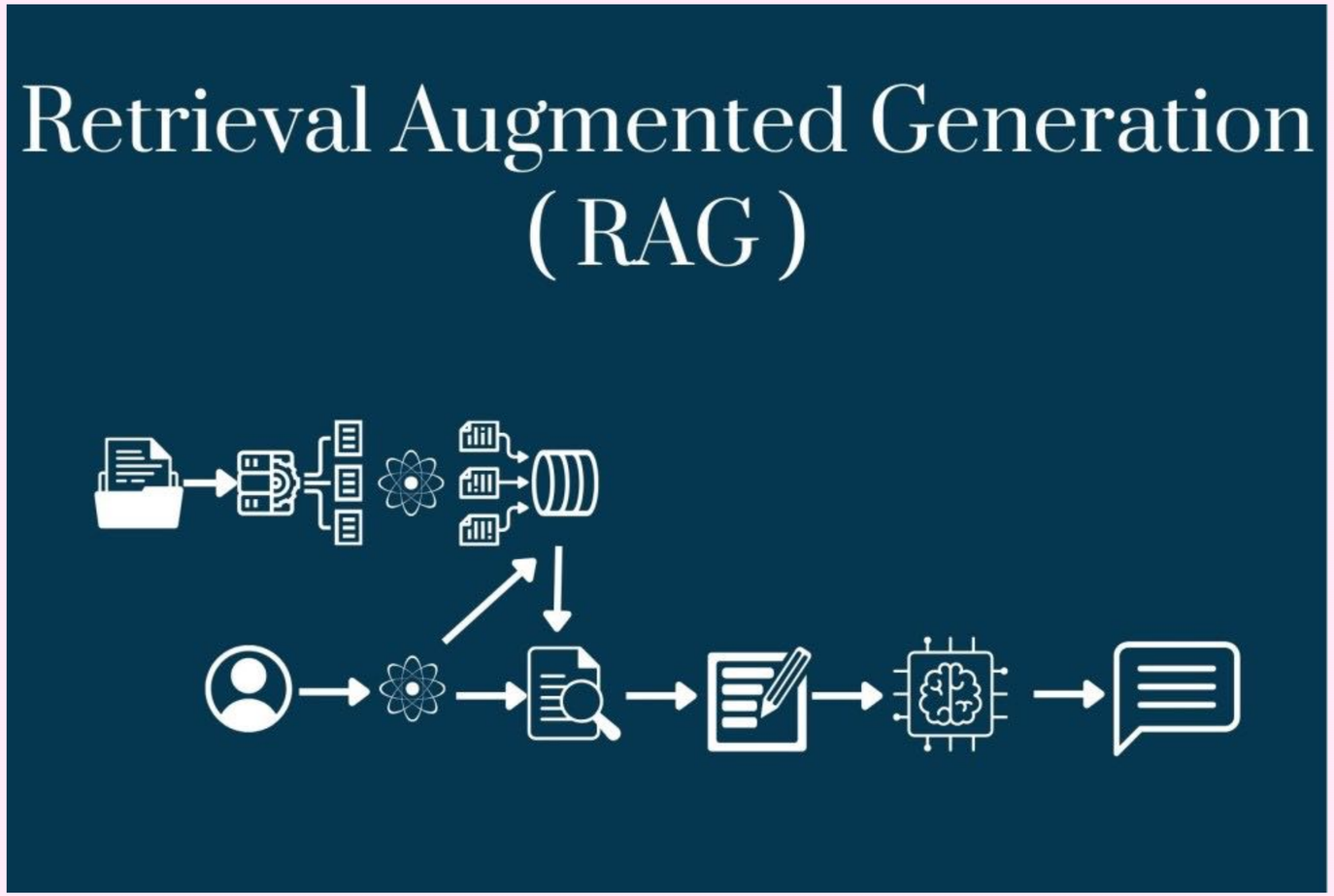


Image prise ici

Au moins 3 composants (différentiables ou pas):

- (Reformulateur de requêtes)
- Retriever
- (Sélecteur)
- Prompt Builder
- Generator

Pas forcément un problème technologique

```python
from langchain_core.prompts import PromptTemplate

prompt_template = PromptTemplate.from_template("""You are an assistant for question-answer
Question: {question}
Context: {context}
Answer: """)

def rag_pipeline(query):
    # Tout d'abord, on recherche les documents
    retrieved_docs = vector_store.similarity_search(query)
    # Ensuite, on injecte les documents dans le prompt
    prompt = prompt_template.invoke({
        "question": query,
        "context": "\n\n".join(doc.page_content for doc in retrieved_docs)
    })
    # Enfin, on envoit l'intégralité du prompt au LLM
    return llm.invoke(prompt).strip()
```

Il ne nous reste plus qu'à exécuter notre fonction pour réaliser une inférence de notre pipeline RAG.

```python
query = """
In 2009, how much loans and guarantees were proposed by the bank? Please give details abou
"""

# Effectuer une requête
response = rag_pipeline(query)
print(response)
```

note: clairement un problème éthique (surconsommation des GMLs)

Pris du blog: Créer un RAG avec LangChain et Mistral à partir de zéro, Maxime Jumelle, 13 fév. 2025

Objectifs

- Mettre au point un *benchmark* RAG ouvert propre à Érudit (train/test)
  - requiert la collaboration de chercheurs faisant usage d'Érudit
  - possibilité d'utiliser des GLMs pour générer des questions/réponses
- Tester différentes technologies de RAG
- Identifier les problèmes:
  - hallucinations, contenus biaisés (ex: écartant des textes de certaines communautés)
  - utilisation détournée d'un agent conversationnel (ex: comment faire une bombe ?)

4- Valorisation des données

Données textuelles de la recherche d'Érudit

Journals, newspapers, magazines

- Érudit, 342 journals (1905 to 2024) -- 568,478 files, **203 GB**
- Bibliothèques et Archives nationales du Québec (17th century) -- 4,627,040 files, **18 TB**
- Canadiana/CRKN (18th century to 1930) -- 80,085 files, **405 GB**
- Library and Archives Canada (1820 to 1917) -- 789 files, **5 GB**

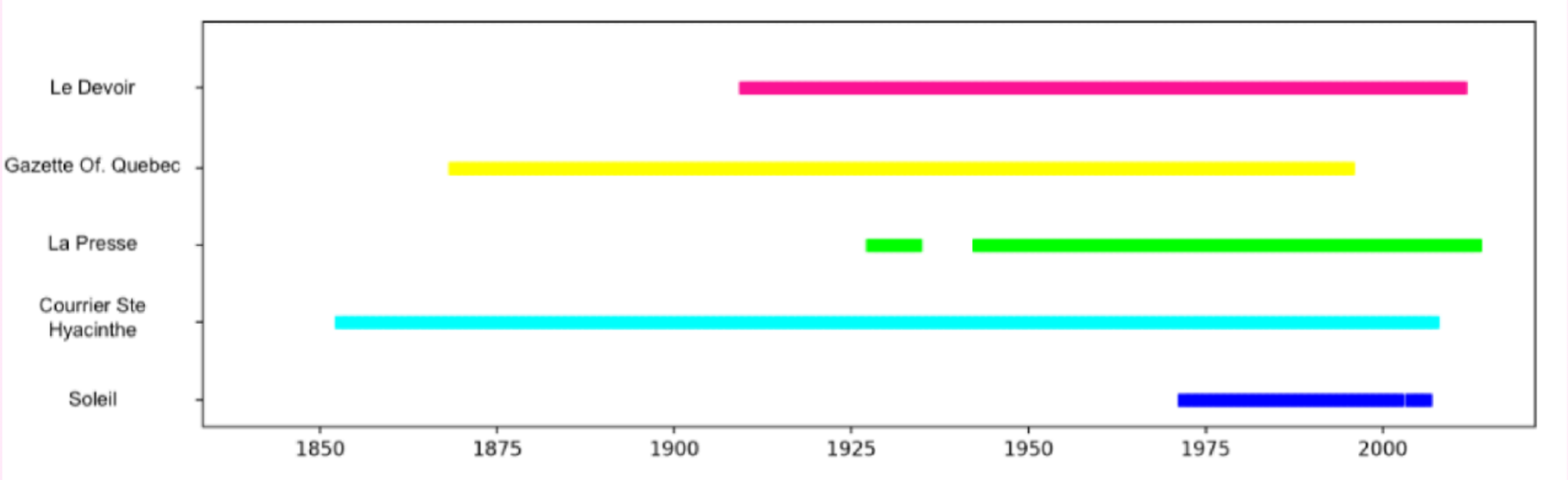
Parliamentary debates

- Library and Archives Canada
  - **Canada Gazette** (1842 to 1997) -- 14,560 files, **206 GB**
  - Cabinet Conclusions (1944 to 1979) -- 41,249 files, **10 GB**
- Library of the National Assembly of Québec
  - Journal des débats de l'Assemblée nationale du Québec (1908 to 2019) -- 33,339 files, **31 GB**

Government Reports

- National Centre for Truth and Reconciliation (2002 to 2021) -- 15,391 files, **535 MB**

Journaux 1/3



Figures prises du mémoire de maîtrise de **David Kletz**, Méthodologies pour la détection de diachronies sémantiques et leurs impacts, 2021

Journaux 2/3

Recapitalisation probable de la Lake Superior Corp.

D'après le Wall Street Journal, le groupe canadien qui a acquis le contrôle de la Lake Superior Corporation est en train d'établir un plan pour modifier le capital ainsi que celui de quelques filiales. Il est possible que la charte actuelle du New-Jersey soit remise et que la

actionnaires à une assemblée prochaine. Un avis de convocation ainsi qu'une copie du plan de recapitalisation sont envoyés immédiatement aux actionnaires.

"Vos directeurs sont porteurs d'un nombre considérable d'actions des deux catégories et approuvent unanimement le plan comme étant dans le meilleur intérêt de tous les actionnaires et de la compagnie. Si vous ne pouvez pas assister à l'assemblée, faites une procuration à l'ordre de vos directeurs, si le plan est adopté ,les actions seront placées sur une base de dividende immédiatement."

Version numérisée d'un article de La Presse du 22/06/1928

RBcapitallsatlon probable de la Lake Superior Corp.
avait ivnîreprinee aux usines de bt, Ca- j faveur d#5 pori-llr.-» d'actioru? ordirlftlr#» therina»; il rn» fait plaisir de air# que 1 actueBé», soit au taux d'un* pour une. Ofis ligne», censista.nl surtout d# la- ; ^ tmU,^ s<*t 86,996 aettan» "A" «t
biew'.x d# command# et inMaflatlon gé« j io,606 «otioni» "B", demeurera dans le nfirai# électrique, dont le marolifi s é-I trÀw .pour le» besoin» fulUîi» de la oom-ter:d oontlnuellement au Panada, ont
été tri># favoraidement accueülies pafi; si le plan est adopté, les dfirecte<9c>-» le» ollent». î ont Vin trillion de reoommandvr le pale-
<93>Vert directeur» fittidH-nt depuis quel- | nw-r.t d'«n dividende de *3 ipar action . qua iemt» un plan pour modifier !»i -A" et espèrent pouvoir «n dont muer raht-gfinfiraî, en sera ie pnésrident eijpn»! d< la oompagnl#: ce plan est, paiemaxit.
gérant-général, iinair, tenant terrotné et Son-, soumis aux! .....- ¦¦ .... actionnaire» S une aasemblfi» prochaine. I ~r"
ÏTn avl» de convocation ainsi qu'un# cd- ; pie du plan de rficapitAlieatk>n sont en- ; vpyéé' Immédiatement aux actionnaire» !
¦<95>Vos directeurs sont porteur» d'un! nombre considérable (l'action* de» deux ' catégories et approuvent unanimement I le plan comme étant dan* le meilleur [intfirft' rie tous le» actionnaire» et de la D'après le Wall Street Journal, le cobipsurnle. Si vtws ne pouvez P»" groupe canadien qui a acquis la contrôle iasolaier ft rassemblée, fafte» une propu-de la Lake Superior CoritoraÉlon est en [ration A l'ordre de vos directeurs, b, le
train d'établir un plan pour modifier leiplan est adopté .le»
action» «fWit plft-i-apltai iilnst que celui de quelque» ftîlft-
|eéev* sur une btu» de dividende Immé-le«. U «si iKisalbl» que la
chart# actuelle I dlatement."
du New-Jersey soir remis# et que l

OCR

Journaux 3/3

Version numérisée d'un article de Télé RADIO MONDE de la semaine du 12 au 19/01/1980

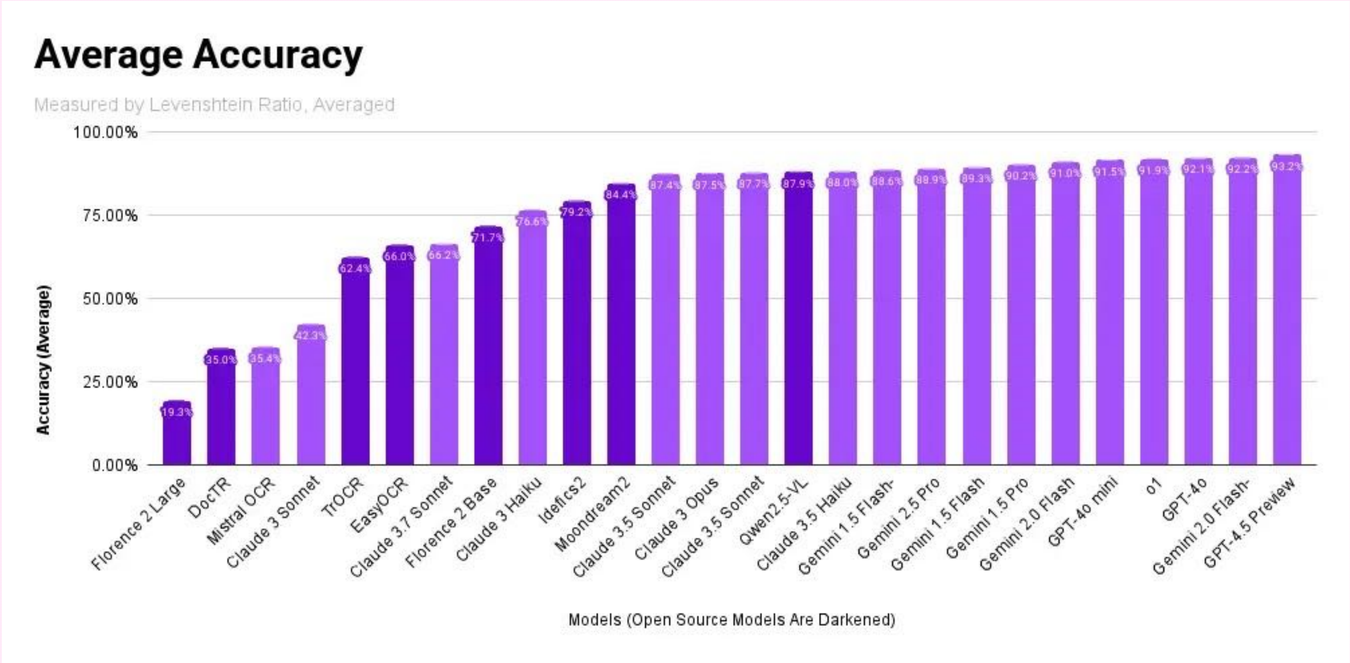
```
###PAGE###1###
Mention incorrecte : Date

2 ao


.


A
VOLUME 42 NUM…RO 18 DU 12 AU 19 JANVIER 80 PRIX- 50e ó ….U. & N.B.: 5

¶-.Ô*1 i
I


N'AVAIT PAS
```

OCR

Objectifs

pour un corpus plus serviable

- Utiliser des modèles d'OCR plus récents



pris du blog:

Best OCR Models for Text Recognition in Images, Leo Ueno

- Classer les documents à l'aide de critères de qualité de l'OCR

(Brève) conclusion

En résumé

- Le passage à l'IA au sein d'une plateforme comme Érudit requiert plus que de l'intégration de technologies
  - *benchmarks* bien pensés
  - sensibilisation ciblée des actrices de l'édition (éditrices, auteures)
- Le domaine des SHS peut nourrir le traitement automatique des langues (TAL)
  - domaine (SHS) largement sous-étudié en TAL
  - volet français: un plus
- Une équipe forte


Questions / commentaires ?

Métadonnées

Dépôt de thèse du Canada

Collecte effectuée en 2023 (alors que Thèses Canada n'était qu'une coquille vide): 57 dépôts pour un total de 728,754 documents

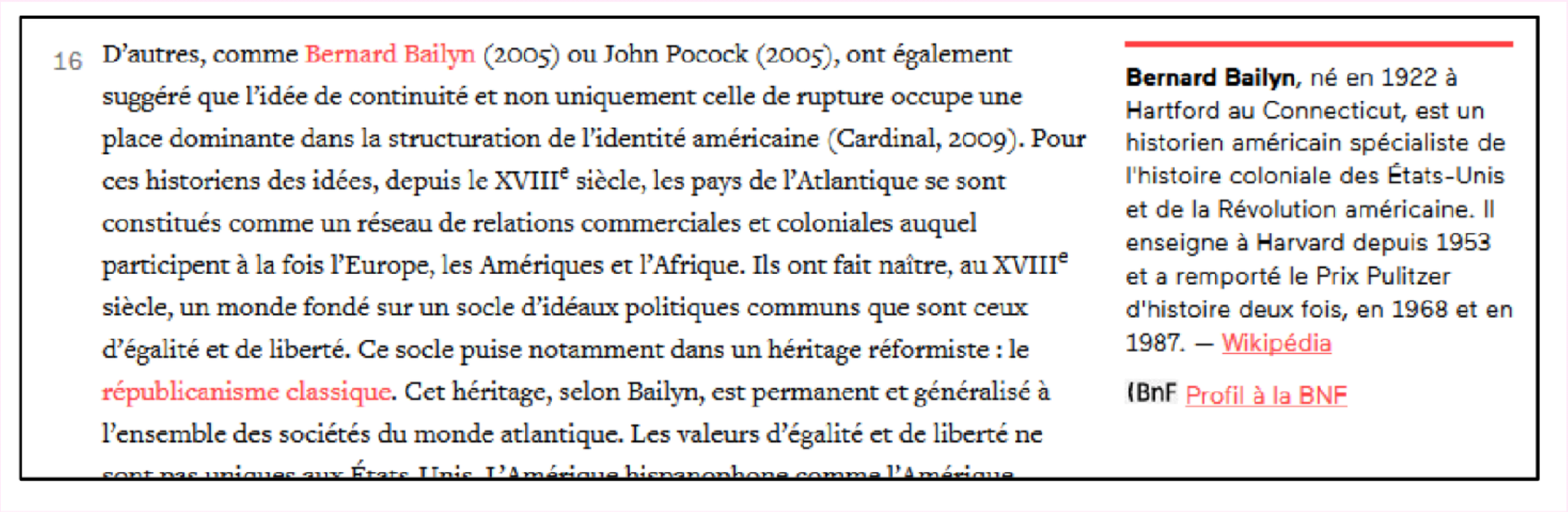
Tag	Frequency of absence
Source	99.4%
Coverage	95.0%
Relation	74.4%
Contributor	60.4%
Rights	55.4%
Publisher	49.2%
Subject	32.9%
Format	29.5%
Language	23.1%
Abstract	17.7%
Creator	12.6%
Type	7.0%
Date	3.0%
Identifier	0.3%
Title	0.0005%

% champs vides (OAI_DC format)

The state of OAI-PMH repositories in Canadian Universities, **Piedboeuf et al.**, DCMI 2023

```
<metadata>
<oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd
http://www.w3.org/1999/02/22-rdf-syntax-ns# http://www.openarchives.org/OAI/2.0/rdf.xsd ">
<dc:title>Personality extraction through LinkedIn</dc:title>
<dc:creator>Piedboeuf, Frédéric</dc:creator>
<dc:contributor>Langlais, Philippe</dc:contributor>
<dc:contributor>Lapalme, Guy</dc:contributor>
<dc:subject>Extraction de personalité</dc:subject>
<dc:subject>MBTI</dc:subject>
<dc:subject>DiSC</dc:subject>
<dc:subject>LinkedIn</dc:subject>
<dc:subject>Réseau sociaux</dc:subject>
<dc:subject>Profilage d’auteur</dc:subject>
<dc:subject>Personality Extraction</dc:subject>
<dc:subject>Social Network</dc:subject>
<dc:subject>Author Profiling</dc:subject>
<dc:subject>
Applied Sciences - Computer Science / Sciences appliqués et technologie - Informatique (UMI : 0984)
</dc:subject>
<dc:description>
L’extraction de personnalité sur les réseaux sociaux est un domaine qui n’a que récemment commencé à capturer l’attention
des chercheurs. La tâche consiste à, en partant d’un corpus de profils d’utilisateurs de réseaux sociaux, être capable
de classifier leur personnalité correctement, selon un modèle de personnalité tel que défini en psychologie. Ce
mémoire apporte trois innovations au domaine. Premièrement, la collecte d’un corpus d’utilisateurs LinkedIn.
Deuxièmement, l’extraction sur deux modèles de personnalités, MBTI et DiSC, l’extraction sur DiSC n’ayant pas encore
été faite dans le domaine, et finalement, la possibilité de passer d’un modèle de personnalité à l’autre est explorée,
dans l’idée qu’il serait ainsi possible d’obtenir les résultats de multiples modèles de personnalités en partant d’un
seul test.
</dc:description>
<dc:date>2019-11-19T19:23:16Z</dc:date>
<dc:date>NO_RESTRICTION</dc:date>
<dc:date>2019-11-19T19:23:16Z</dc:date>
<dc:date>2019-10-30</dc:date>
<dc:date>2019-05</dc:date>
<dc:type rdf:resource="http://purl.org/coar/resource_type/c_46ec" xml:lang="en">thesis</dc:type>
<dc:type rdf:resource="http://purl.org/coar/resource_type/c_46ec" xml:lang="fr">thèse</dc:type>
<dc:identifier>http://hdl.handle.net/1866/22536</dc:identifier>
<dc:language>eng</dc:language>
<dc:format>application/pdf</dc:format>
</oai_dc:dc>
</metadata>
```

Allium



16 D'autres, comme Bernard Bailyn (2005) ou John Pocock (2005), ont également suggéré que l'idée de continuité et non uniquement celle de rupture occupe une place dominante dans la structuration de l'identité américaine (Cardinal, 2009). Pour ces historiens des idées, depuis le XVIII[e] siècle, les pays de l'Atlantique se sont constitués comme un réseau de relations commerciales et coloniales auquel participent à la fois l'Europe, les Amériques et l'Afrique. Ils ont fait naître, au XVIII[e] siècle, un monde fondé sur un socle d'idéaux politiques communs que sont ceux d'égalité et de liberté. Ce socle puise notamment dans un héritage réformiste : le républicanisme classique. Cet héritage, selon Bailyn, est permanent et généralisé à l'ensemble des sociétés du monde atlantique. Les valeurs d'égalité et de liberté ne sont pas uniques aux États-Unis. L'Amérique hispanophone comme l'Amérique

Bernard Bailyn, né en 1922 à Hartford au Connecticut, est un historien américain spécialiste de l'histoire coloniale des États-Unis et de la Révolution américaine. Il enseigne à Harvard depuis 1953 et a remporté le Prix Pulitzer d'histoire deux fois, en 1968 et en 1987. — Wikipédia

(BnF Profil à la BNF

Large-scale Semantic Annotation for Improved Content Discovery in a Digital Library: A Case Study on Érudit, **Gotti et al.**, 2022