



Nouvelles approches pour l'extraction d'informations dans les documents (manuscrits) numérisés

Mélodie Boillet · Solène Tarride · Christopher Kermorvant

TEKLIA, Paris, France



New approaches?

- Standard sequential workflow
- Integration of the different steps into a single model
- End-to-end goal oriented models
- Continuous improvement



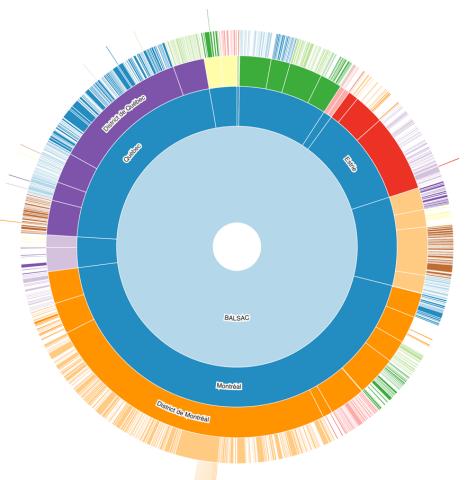
The e-Balsac project



Genealogical database manually extracted
from parish and civil registers for more than
50 years

Goal : Automatic processing of the Québec parish registers 1850-1920

10 Centers
36 Districts
1,985 Parishes
44,742 Registers
1,995,646 images

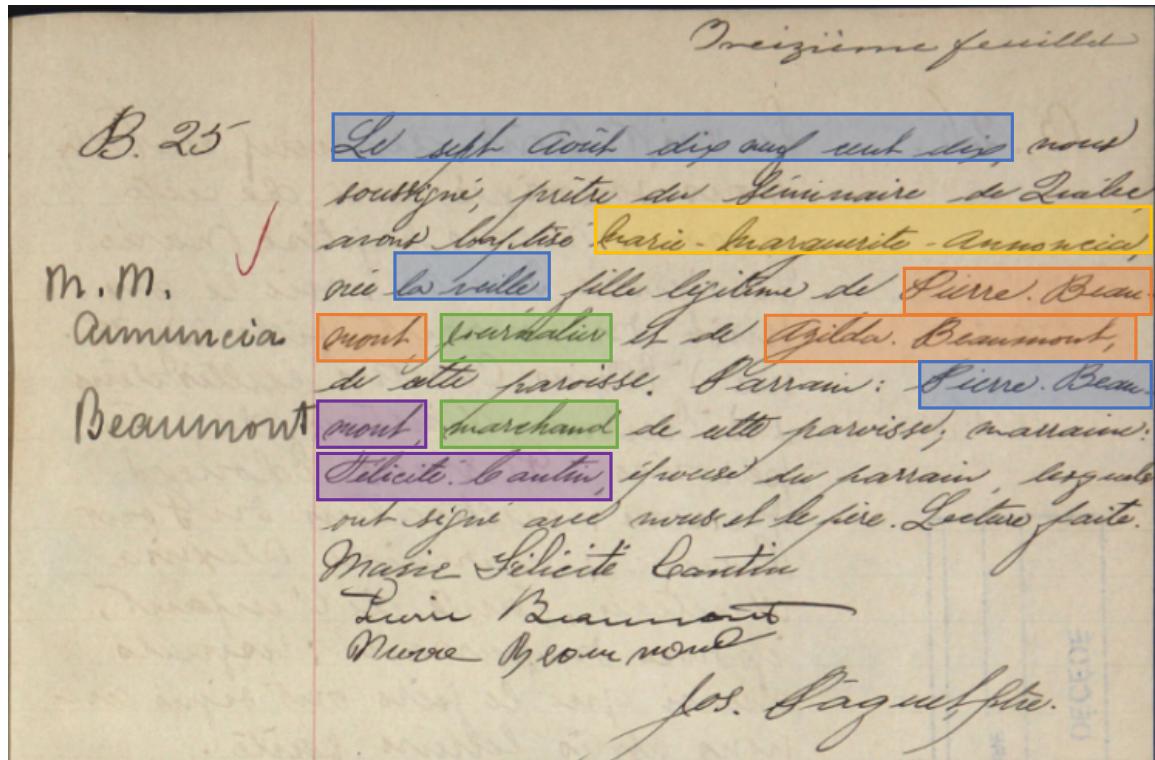


M. 41 Le septembre mil neuf cent quatre-vingt, ou la publication de
Pierre Chénier Attois bans de mariage, faite au greffe de nos Inscopacio[n]es
et Entre Pierre Chénier, fils unique de Jeanne Chénier et d'Alphonse
Bernadette St Louis Roy de cette paroisse d'autre part, et Bernadette St Louis, fille unique
de feu Louis St Louis et de Josephine Vaudier, aussi de cette
paroisse d'autre part; n[on] s'étant découvert aucun empêche-
ment de mariage et du consentement des parents des dits
mariages, nous, prêtre consacré, avons reçu leur mutuel
consentement de mariage et leur avons donné la bén-
dition mystique en présence de Jeanne Chénier, prie de
l'époux et de Louis Corriveau, beau-frère de l'épouse. Les époux
Seuls et seuls avec nous. Lecture faite.

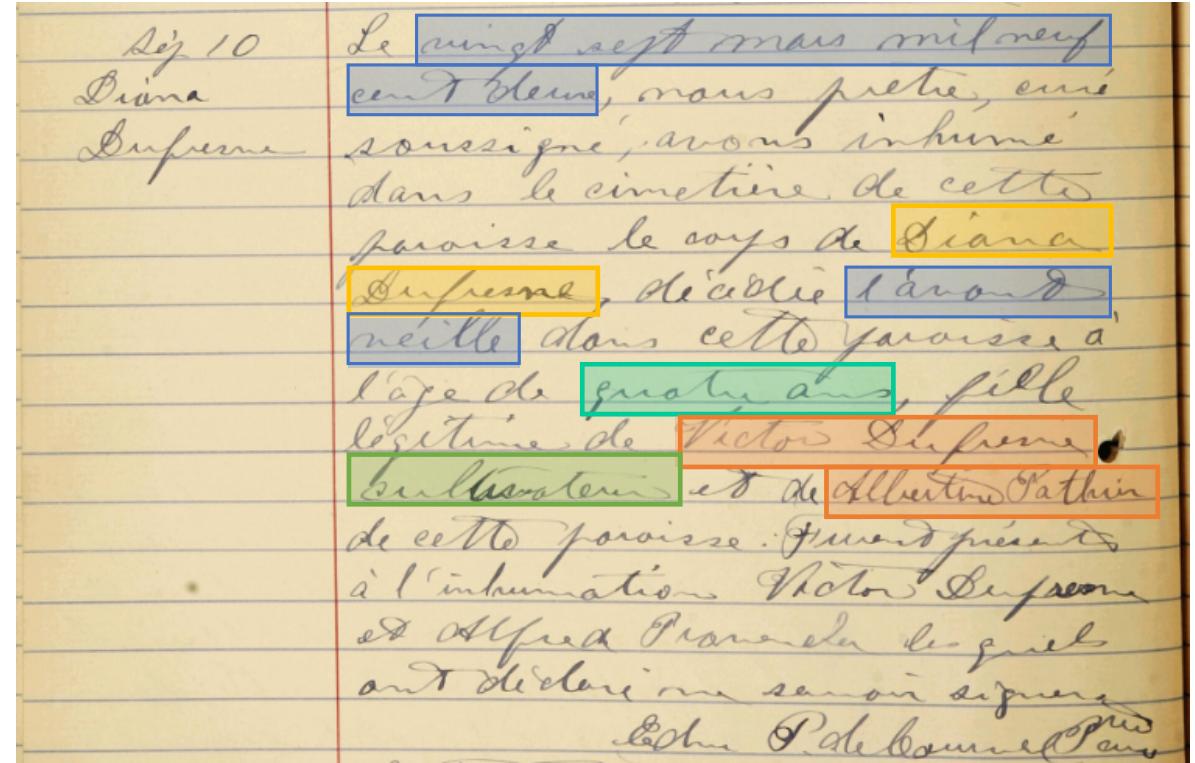
M. 12 Le septembre dix-neuf cent quarante, vu la dispense
Marie Renaud de deux baus de mariage, par nous accordée en vertu de pouvoirs
et à nos confis par sa Grandeur Monseigneur Charles
Aurore Paquette, Mgr. Gauthier, archevêque d'Ottawa, vu aussi la puden-
tation d'un baau fait au prieur de notre mense paroissiale
ainsi que à celui de l'Assomption, entre Marie Renaud
fille majeure d'Alarie Renaud et d'Abdala Latremouille de
l'Assomption d'une part, et Aurore Paquette, fille
unique d'Olivier Paquette et d'Anna Morin de cette paroisse
d'autrepart; ne s'étant rencontré avec une empêchement
de mariage et du conseillement des parents de la dite
masure, nous, prêtre ouvrier, avons reçu leur mutuel
consentement de mariage et leur avons donné la bén-
édiction nuptiale en présence d'Alarie Renaud, prie de l'ép-
ouse et d'Olivier Paquette, prie de l'épouse. Tousont, Lépine et
mons. Lecture faite.

M. 73 Le sept Septembre mil neuf cent quatre-
Oscar Anderson, via la dépense de deux lans de mariage
par nous accordé en vertu de pouvoirs à
Florida Trumblay nous conférés par sa Grandeur Augr Charles

Québec civil and parish registers



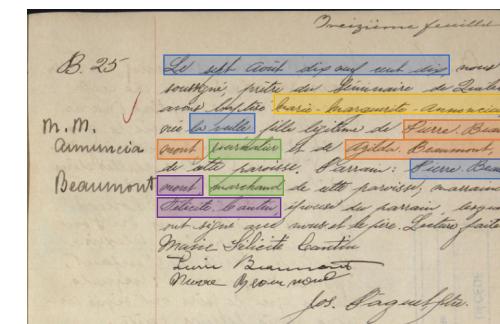
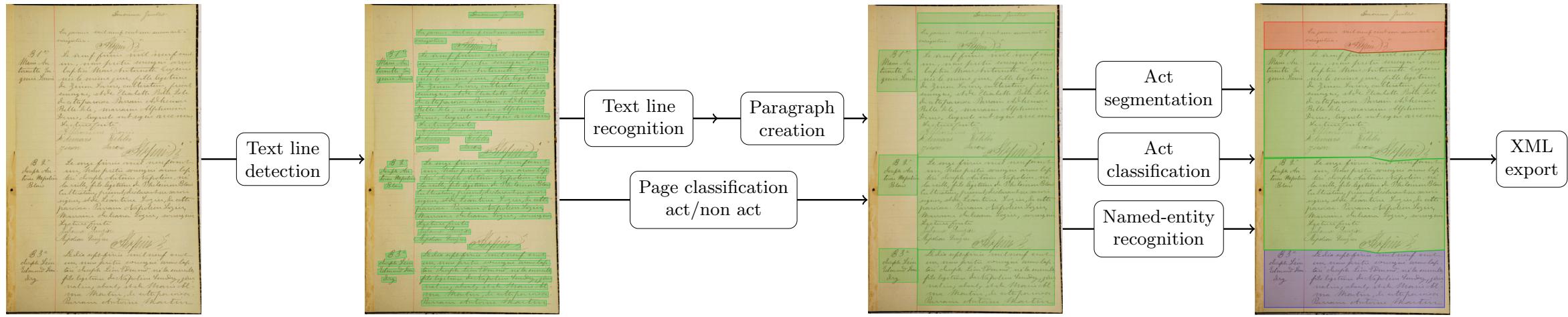
Birth act



Death of a single person act

■ Date; ■ Subject of the act; ■ Parents; ■ Spouse; ■ Age; ■ Occupation; ■ Godfather/godmother

Overview of the document processing workflow

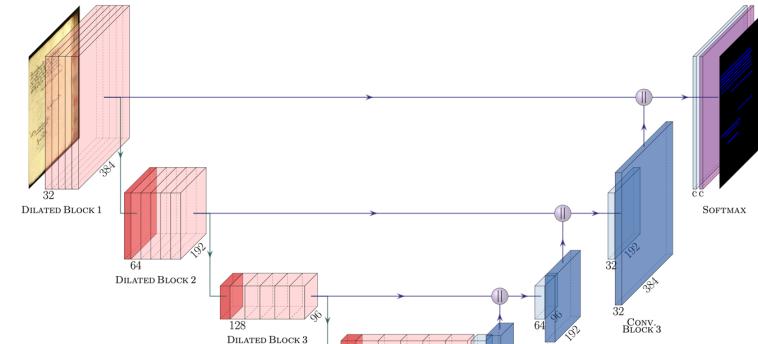


NER

Data & Technologies

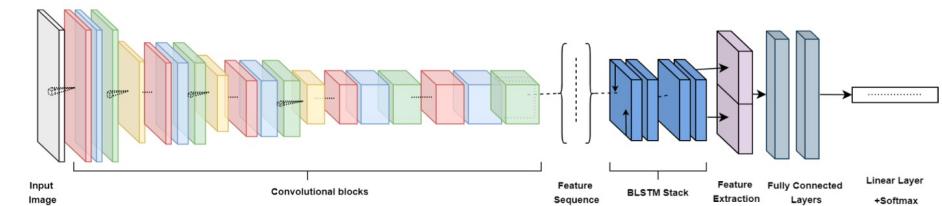
- Automatic line detection + correction
- Manual transcription
- Named-entity annotation + relations
- Annotation : 0,05% of the full corpus

	pages	acts	lines	words	entities
count	896	2,661	45,479	205,165	25,564
	entities				
	PER	LOC	DATE	OCC	
count	15,810	2,823	4,551	2,380	



DLA: Doc-UFCN model

<https://gitlab.teklia.com/dla/doc-ufcn>



HTR : PyLaia CNN+biLSTM

<https://gitlab.teklia.com/atr/pylaia>

NER

flair

Classification

scikit
learn

Results: Automatic data validation with act template

Quantitative evaluation

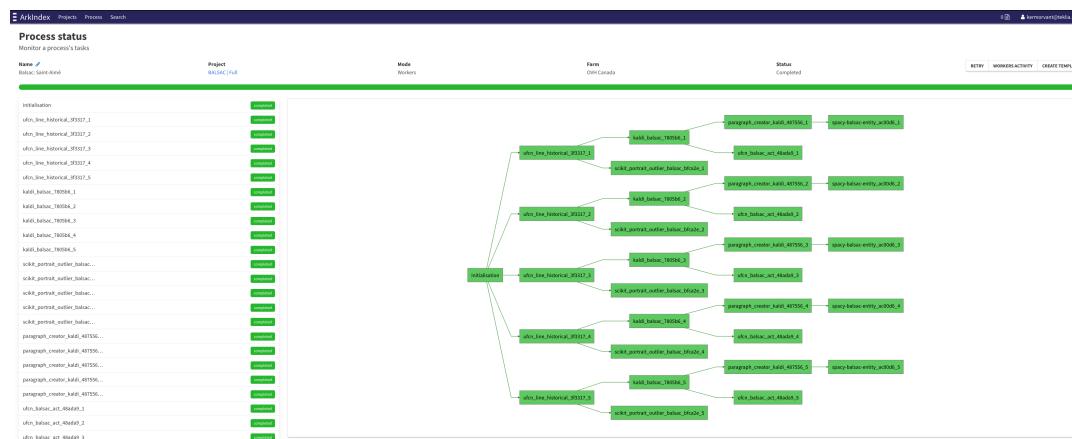
Entities	Count
Occupations	4 624 291
Locations	5 429 807
Dates	8 655 893
Persons	30 986 429

Qualitative evaluation

type	Birth act	Death act
Valid	75.5%	69.0%
Fusion	19.9%	18.2%
Invalid	7.1%	8.9%
Special	0.4%	3.2%

Lessons learned

- Complex workflow with one model per task
- Unsupervised metrics of the full process are needed to improve the quality
- Automatic processing but under human monitoring
- The workflow would be simpler now using end-to-end transformer models



Final goal : key-value information extraction

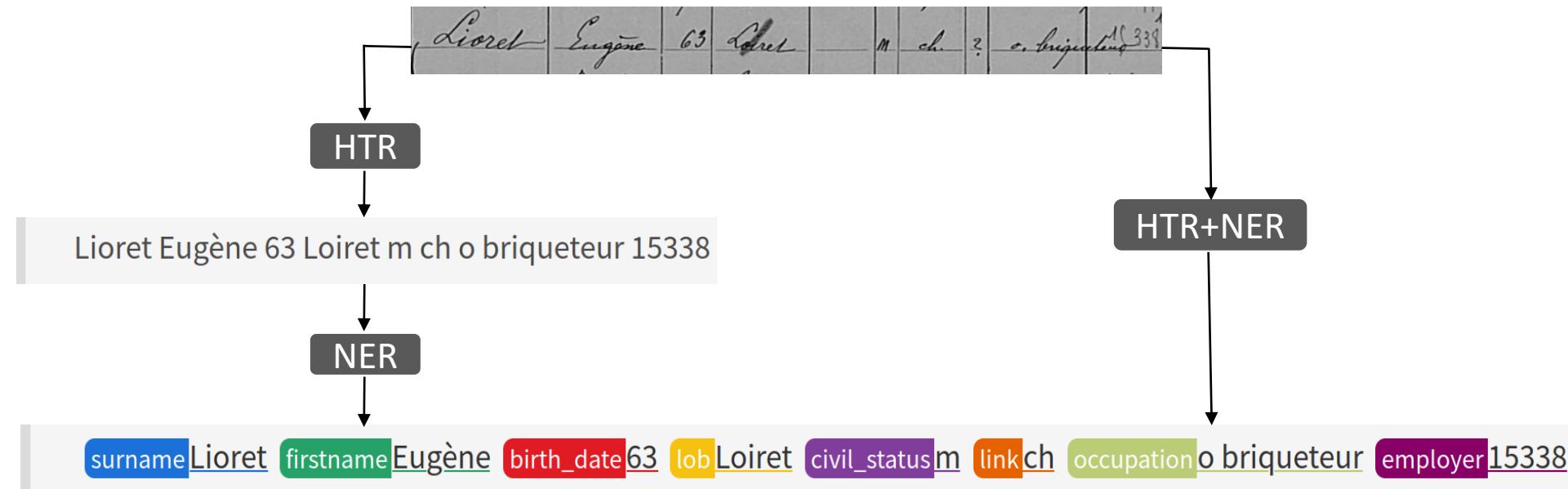
B.4.
Marie
Alice
Cora
Charlotte
Dufault.

Le vingt-sept février mil neuf
cent un, sous ma signature, prêtre curé
de cette paroisse, avons baptisé Marie
Alice Cora Charlotte, née la veille fille
légitime de Napoléon Dufault,
et de Alice Mercier de cette paroisse.
Le parrain a été Pierre Dufault,
grand-père de l'enfant et la marraine
son épouse Marie Lavallée
qui ainsi que le père ont signé
avec nous. Le parrain a déclaré
ne paroir signer lecture faite.
Marie Lavallée
Napoléon Dufault
J. D. Benoit (signature)



Key	Value
Child name	Marie Alice Cora Charlotte
Birth date	26/02/1901
Father name	Napoléon
Father last name	Dufault
Mother name	Alice
Mother last name	Mercier

Integration of HTR+NER

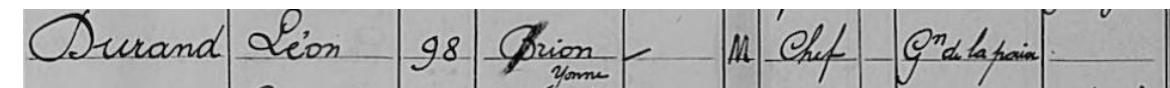


Sequential approach
DAN or PyLaia for HTR
SpaCy for NER

End-to-end approach
DAN [4] for HTR+NER
Special tokens are used to tag entities

Integration of HTR+NER

Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
GT + SpaCy	0.0	96.4	Line
PyLaia + SpaCy	17.19	76.3	Line
DAN + SpaCy	8.18	84.0	Page
DAN end-to-end	7.83	85.9	Page



surname Durand firstname Léon birth_date 98 lob Brion Yonne civil_status M
link Chef occupation Gardien de la paix

- End-to-end models are competitive
- End-to-end models are easier to train, improve and deploy

One step beyond : Integration DLA+HTR+NER

DÉSIGNATION	NOMBRE des familles dans les villes et villages	NOMBRE des familles dans les villages	NOMBRE des familles dans les villes	NOMS DE FAMILLE	PRÉNOMS	ANNÉE DE NAISSANCE	LIEU DE NAISSANCE	NATIONALITÉ	STATUTS DU MARIAGE	PROFESSION	Pour les personnes nées à l'étranger	
											11	12
Par Thonon				Lecatard	François	1819	Boulemette	France	Chef	agriculteur		
	171			Cours de Lecatard	Marie	1842	Castelotay		Espouse	ménageant		
	191			Lecatard	Bernice	1876	Boulemette		enfant	je		
	172			Lecatard	Georges	1875	Castelotay					
	173			Laufrid	Céleste	1872	Boulemette					
	174			Perroud	François	1812	Boulemette		Chef	magasinier		
	175			Perroud	Eva	1844			Espouse	je		
	176			Perroud	Yvonne	1869			enfant	je		
	177			Ludolfier	Julie	1872	Boulemette		Chef	je		
	178			Derbez	Césarine	1872	Perol		je	cuisinière		
	179			Touraud	Joseph	1815	Castelotay		je	capitaine		
	180			Fourcier-Lamant	Marie	1861	St-Prix		Espouse	ménageant		
	181			Toubert	Jan Bapt	1816	Yveret					
	182			Velard & Miller	Josephine	1830	Boulemette		Chef	maisonnière		
	183			Bellot	Elie	1866			enfant	journaliste		
	184			Bellot	Thérèse	1876						
	185			Bellot	Lulu	1871						
	186			Abel	Gerard	1866	Fourney		Chef	tailleur		
	187			Abel	oile	1866	Castelotay		Espouse	je		
	188			Abel	Suzie	1871	Boulemette		enfant	tailleur		
	189			Abel	Marie	1879				je		
	190			Abel	Alfred	1870				je		
	191			Abel	Fantine	1870				je		
	192			Abel	Bernice	1866						
	193			Garnier	Marie	1818	Fourney		Chef	boulanger		
	194			Garnier	Auguste	1870	Boulemette		enfant	ménageant		
	195			Garnier	Adeline	1870	Fourney		Chef	butcher		
	196			Garnier	Zézay	1873	Boulemette		Espouse	je		
	197			Garnier	Yvonne	1873	Boulemette		enfant	je		
	198			Garnier	Albert	1874	Fourney					
	199			Garnier	Paul	1874	Fourney					
	200			Garnier	Yves	1874	Boulemette					



surname	Léautaud	firstname	François	occupation	négociant	linkchef	birth_date	1839	nationality	française	lob	Barcelonnette	
surname	Caire ep	Léautaud	firstname	Marie	occupation	néant	linképouse	birth_date	1849	nationality	idem	lob	Encharteaues
surname	Léautaud	firstname	Berthe	occupation	idem	linkenfant	birth_date	1876	nationality	idem	lob	Barcelomette	
surname	Jauffred	firstname	Aglâé	occupation	idem	linkdomestique	birth_date	1875	nationality	idem	lob	Gnetrartrayes	
surname	Brunet	firstname	François	occupation	brig ferrier	linkchef	birth_date	1852	nationality	idem	lob	Bayons	
surname	Margaillan ep	Brunet	firstname	Euphanie	occupation	néant	linképouse	birth_date	1865	nationality	idem	lob	Selonnet
surname	Brunet	firstname	Fernand	occupation	idem	linkenfant	birth_date	1892	nationality	idem	lob	Bayons	
surname	Brunet	firstname	Eva Marie	occupation	idem	linkidem	birth_date	1894	nationality	idem	lob	idem	
surname	Brunet	firstname	Yvonne	occupation	idem	linkidem	birth_date	1899	nationality	idem	lob	idem	
surname	Audiffred	firstname	Julie	occupation	néant	linkchef	birth_date	1825	nationality	idem	lob	Barcelonnette	
surname	Derbez	firstname	Césarine	occupation	culurière	linkidem	birth_date	1872	nationality	idem	lob	Revel	
surname	Arnaud	firstname	Joseph	occupation	cafetier	linkidem	birth_date	1865	nationality	idem	lob	Enchartayes	
surname	Risoulép Arnaud	firstname	Marie	occupation	néant	linképouse	birth_date	1865	nationality	idem	lob	St Pons	
surname	Jaubert	firstname	Jean Baptiste	occupation	domestique	linkidem	birth_date	1876	nationality	idem	lob	Nvernet	
surname	Allard vve Bellon	firstname	Joséphine	occupation	blanchisseuse	linkchef	birth_date	1834	nationality	idem	lob	Barcelomette	
surname	Bellon	firstname	Félix	occupation	journalier	linkenfant	birth_date	1864	nationality	idem	lob	idem	
surname	Bellon	firstname	Clémentine	occupation	blanchisseuse	linkidem	birth_date	1876	nationality	idem	lob	idem	
surname	Bellon	firstname	Julie	occupation	repasseuse	linkidem	birth_date	1875	nationality	idem	lob	idem	
surname	Abel	firstname	Edouard	occupation	tailleur	linkchef	birth_date	1849	nationality	idem	lob	Faucon	
surname	Jean ép Abel	firstname	Odile	occupation	néant	linképouse	birth_date	1856	nationality	idem	lob	Condarnine	
surname	Abel	firstname	Paul	occupation	tailleur	linkenfant	birth_date	1878	nationality	idem	lob	Barcelonnette	
surname	Abel	firstname	Marie	occupation	néant	linkidem	birth_date	1879	nationality	idem	lob	idem	
surname	Abel	firstname	Alfred	occupation	idem	linkidem	birth_date	1890	nationality	idem	lob	idem	
surname	Abel	firstname	Fortuné	occupation	idem	linkidem	birth_date	1892	nationality	idem	lob	idem	
surname	Abel	firstname	Berthe	occupation	idem	linkidem	birth_date	1900	nationality	idem	lob	idem	
surname	Eymé	firstname	Marius	occupation	boulanger	linkchef	birth_date	1842	nationality	idem	lob	Puy de Euzèbe	
surname	Lyme	firstname	Marguerite	occupation	néant	linkenfant	birth_date	1880	nationality	idem	lob	Barcelonnette	
surname	Maurel	firstname	Adrien	occupation	cant	linkchef	birth_date	1870	nationality	idem	lob	Fours	
surname	Signout ep Maurel	firstname	Constantine	occupation	néant	linképouse	birth_date	1868	nationality	idem	lob	S. Pon	
surname	Maurel	firstname	Jean	occupation	idem	linkenfant	birth_date	1904	nationality	idem	lob	Barcelomarle	

One step beyond : DLA+HTR+NER

Method	Socface/POPP		
	CER (%) ↓	F1 (%) ↑	Level
DAN end-to-end	7.83	85.9	Line*
	-	-	-
	11.74	85.3	Page

- Full page end-to-end models are competitive
- They are even easier to train, improve and deploy
- Ground truth generation is easier (can be trained from page transcription)

<https://socface.site.ined.fr/fr/collaboration/progressions-des-collectes/>



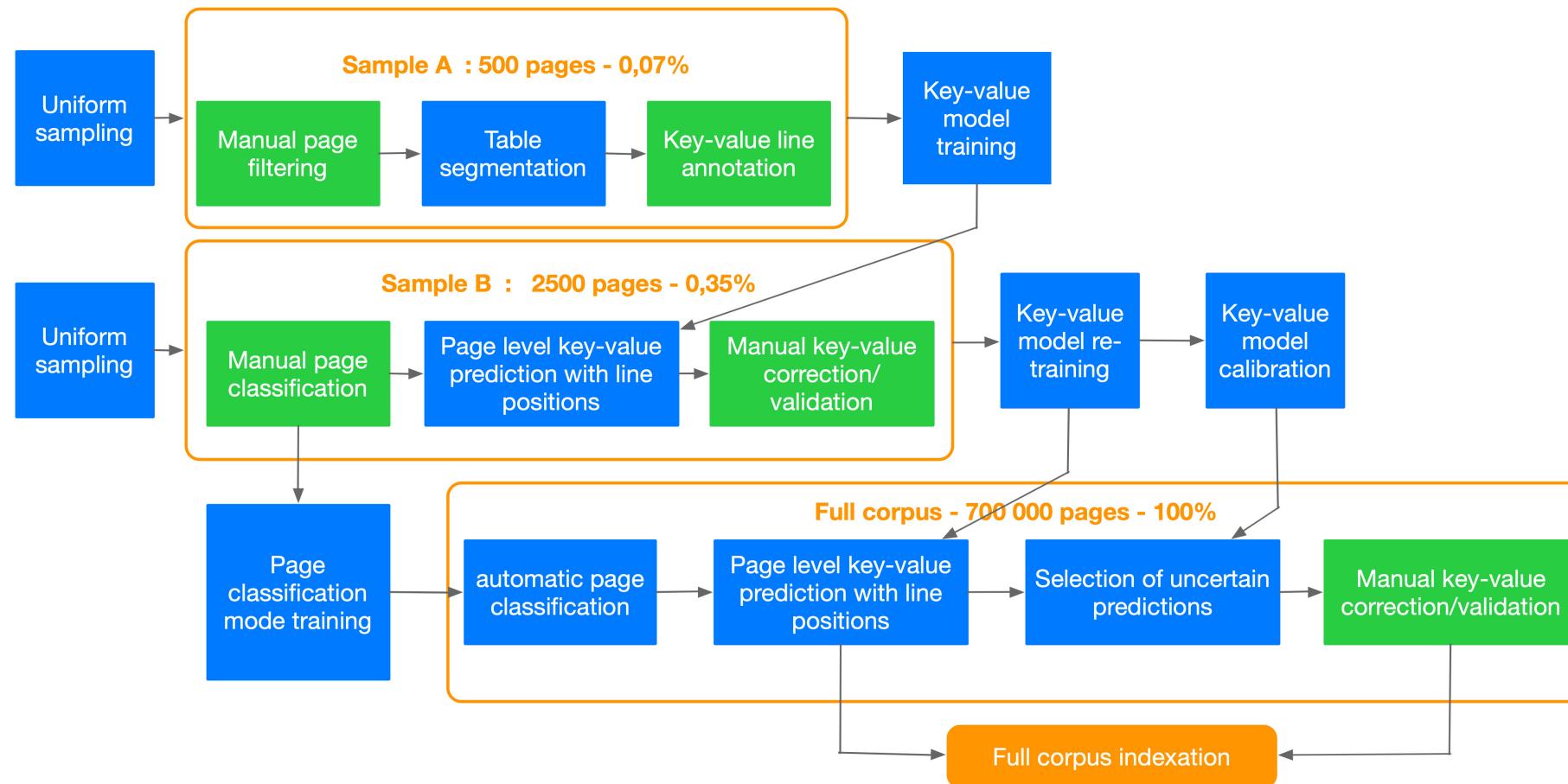
ICRC

In practice : the French WWII POW lists

- 36 million names of POWs from WWII in the ICRC Tracing Archives
- 3000 requests can be processed each year
- Automatic indexing would considerably increase the number of beneficiaries

Staats- an- gehörig- keit	Nr. der Er- fassungs- matrie	Name	Vorname	Geburts- tag	Geburts- ort	Vor- name des Vaters	Familien- name der Mutter	Name und Anschrift der zu benachrichtigenden Person	Dienstgrad	Truppenteil (m. Rg. u.fw.)	Matrikel- Nr.	Ort und Tag der Gefangenennahme	Verwundungen, Verlebungen, Tod (Bereitungsplatz)	Bemerkungen (z.B. Zugänge von anderen Wagnen, bei Blutsverlust oder Waffen erschossen)
RANZ	38565	VANDEENSTEEN	Ado	12-5	Totyghem	Rene	LÉPIN	Mme Jeanne VANDEENSTEEN à la Gare Verte SE Paul d'Epinay Hte Vincennes	SOLDAT	121 Infanterie Lyon	1101	25-5-40	BOULOGNE	
RANZ	38566	MAURY	Fernand	7-3	Montluç	Jeanne	VAIT- TIÈRES	Mme MAURY Julia Limonze par Puylauroque Tarn et Garonne	SOLDAT	417 pompiers	142	24-5-40	BOULOGNE	3190
RANZ	38567	TERRASSE	Martin	16-10- 1906	Espaly	Henri	LAZERT	Mme Louise TERRASSE Pau Bassa Puy Laroque Tarn et Garonne	SOLDAT	F. A. R. C.	2000	BOULOGNE		
RANZ	38568	MOHAMED BEN SOUGLATI	Mohamed	1921	Kenifra	Ahmed Ben Souglati	ITTO BENT ALI OU HAFEM	ITTO BENT ALI OU HACEM Djouza Ouladji Kenifra Maroc	SOLDAT	264 R.A.D.	1939	T 306 Maroc	Belgique 25-5-40	
RANZ	38569	MOHAMED BEN LAHSEN	Mohamed	1918	douar oued Jlimane	Rhego nani	BOU- CHA'IB	LAHSEN BEN RHEZONANI douar oued Jlimane Tunisie Mackra	SOLDAT	264 RAD	T 299		Belgique	
RANZ	38570	DJILLALI HAMED BEN ALI	Djillali Li	17-4	Adadi	AKHIED BEN ALI	FATHA BENT LEM- RADAM	AKHIED BEN ALI Lemradam Bouaz Semicha Tunisie Kairouan Tunisie Sousse Tunisie Hammam Maroc	SOLDAT	5e grenaie RAD	15-5-40		Belgique 19-3-40	
RANZ	38571	PERINET	Georges Adrien	22-1	Epens	Emile	GRILLE -MIN	PERINET Georges Chez M. Octave LARGENTIER Somme Verle sur Courtisols Marne	SOLDAT	106 R. I.	118 Chalon CA 3		Seclin 28-5-40	
RANZ	38572	BOURQUIN	Pierre	10-8	PARIS	Pierre	LLUCAS	Mme Pierre BOURQUIN 72 rue de Colombes Asnières	SOLDAT	106 R. I.	1015		Seclin 28-5-40	

In practice : the French WWII POW lists



Sample A : Automatic line detection + Annotation

Callico Projects / ICRC | List annotation / Campaign Line transcription / Tasks / Annotation table_row 3

The screenshot shows a historical document table on the left and its corresponding digital annotation form on the right.

Document Table (Left):

Soldaten-gehörig-keit	Nr. der Er- kennungs-marke	Name	Geburts-namne	Geburts- tag	Geburts-ort	Vor-name des Vaters	Familien-namne des Vaters	Name und Unterschrift der zu kennzeichnenden Person	Dienstgrad	Zugehörigkeit (m. Ap. u/w.)	Matrikel-Nr.	Ort und Tag der Gefangenennahme	Befreiungen, Verlebungen, Tod (Beerdigungsplatz)	Bemerkungen z. B. Zugabe von anderen Angaben bei Gleichheit darf der Matrikel verändert werden
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F	72892	Harris	Marcel	19/01/1885	Amiens	Marcel	Hippolyte	Marie-Catherine Harris Hippolyte 19/01/1885 Paris 10 ^e	Uffz	Pioneer		Européen 23/11/40	Verwandlung 10/6/90	Européen 10/6/90 Européen 10/6/90 Européen 10/6/90
	72893	Bindel	Louis	19/01/1885	Avranches	Albert	Emile	Alphonse Bindel Louis Emile Alphonse (Asturales d'Anvers)	Sold.	519 Regional		Monte Carlo 23/11/40	"	"
	72894	Zwintzen	Elois	17/10/1885	Wambach	Albert	Juliac	Alphonse Zwintzen Elois Alphonse (Nord)	u	Inf.		Cannobri 19/11/40	"	"
	72895	Leroux	Paul	10/11/1885	Bozonnet	Antoine	Antoine	Alphonse Leroux Paul Bozonnet Antoine (Pole E)	u	173 Ref.		Monte Carlo 23/11/40	"	"
	72896	Brionne	Jacques	10/03/1880	Camier	Jacques	Lamant	Alphonse Brionne Jacques Camier Jacques Lamant (Chaine d'Or) (Bordeaux)	u	Marine		Boulogne 25/11/40	"	"
	72897	Romain	Yves	11/01/1885	Rognon	Gustave	Arlette	Alphonse Romain Yves Gustave Rognon Arlette (van)	Uffz	SM Panc		Boulogne 25/11/40	"	"
	72898	Selcrose	André	19/01/1885	Bapaume	François	Tiquet	Alphonse Selcrose André Bapaume François Tiquet (Pole C)	Oberfeldwebel	Art.		Boulogne 25/11/40	"	"

Digital Form (Right):

Pending

Numéro de référence / Reference number

Nom / Last name / Name

Prénom / First name(s) / Vorname

Date de naissance / Date of birth / Geburtstag



Callico Maintainer

Annotation platform, companion of Arkindex

<https://gitlab.teklia.com/callico/>

Sample B : Prediction validation/correction

Prediction

Name	Jardoin	firstname	Maurice	Birth date	4.8.09	RF number	4963
Name	Faliani	firstname	Joseph	Birth date	19.5.09	RF number	4963
Name	Sermoud	firstname	Léon	Birth date	6.10.18	RF number	4963
Name	Savouillon	firstname	Félix	Birth date	18.1.08	RF number	4963
Name	Méric	firstname	Roger	Birth date	19.1.11	RF number	4963
Name	Besson	firstname	Jean	Birth date	7.3.08	RF number	4963
Name	Haon	firstname	Marcel	Birth date	20.7.13	RF number	4963
Name	Carrasset	firstname	Fernand	Birth date	16.2.18	RF number	4963

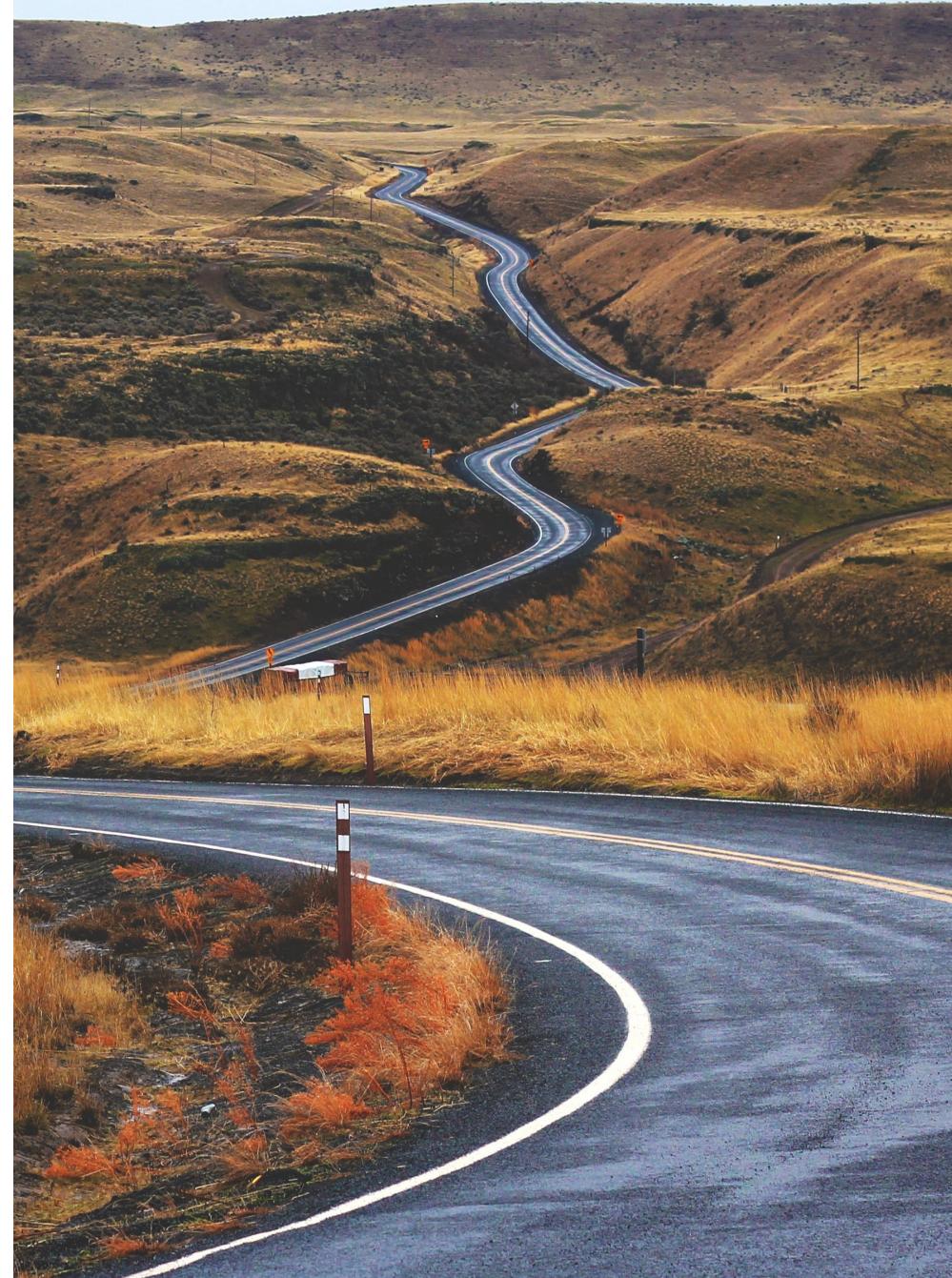
Validation/correction

Re-training

	Time second	Name F1	Firstname F1	Date F1	RF number F1	All F1
Sample A (500)	32	0.91	0.86	0.86	0.97	0.90
Sample B (2500)	16	0.96	0.93	0.96	0.97	0.95

New methods for information extraction

- Full page end-to-end methods are effective and easy to deploy
- In standard cases, little experimentation and fine-tuning of models
- The entire pipeline is annotation-dependent: data cleansing is the critical point
- Does it apply to printed documents ?
- The next frontier: continuous training ?



New methods for information extraction

#non-answerable

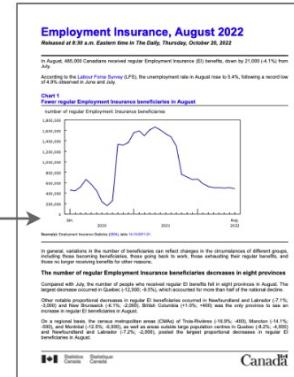
Q: In which year does the Net Requirement exceed 25,000?

A: None

#extractive #list

Q: What are the Years mentioned in Chart 1?

A: [2020, 2021, 2022]

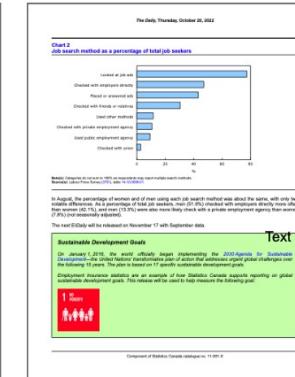


Page 1

#abstractive #counting

Q: How many attorneys are listed for the plaintiffs?

A: Two



Page 2

#layout-navigating #graphic-intensive

Q: Are the margins of the page uniform on all pages?

A: Yes

The table provides data on beneficiaries receiving regular income benefits by occupation. The data points are approximately:

Occupation	Percentage (%)
Health care and social assistance	10.10
Manufacturing, construction, mining, quarrying, agriculture, forestry, fishing and trapping	9.80
Trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing	9.50
Information, culture, recreation, accommodation and food service industries, and repair and maintenance services, except automotive repair and maintenance	9.20
Manufacturing, mining and quarrying	8.90
Manufacturing, construction, mining, quarrying, agriculture, forestry, fishing and trapping, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing, information, culture, recreation, accommodation and food service industries, and repair and maintenance services, except automotive repair and maintenance	8.60
Manufacturing, construction, mining, quarrying, agriculture, forestry, fishing and trapping, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing, information, culture, recreation, accommodation and food service industries, and repair and maintenance services, except automotive repair and maintenance, manufacturing, mining and quarrying	8.30
Manufacturing, construction, mining, quarrying, agriculture, forestry, fishing and trapping, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing, information, culture, recreation, accommodation and food service industries, and repair and maintenance services, except automotive repair and maintenance, manufacturing, mining and quarrying, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing	8.00
Manufacturing, construction, mining, quarrying, agriculture, forestry, fishing and trapping, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing, information, culture, recreation, accommodation and food service industries, and repair and maintenance services, except automotive repair and maintenance, manufacturing, mining and quarrying, trade, transportation, warehousing, postsecondary education, government, public administration, defence, health care and social assistance, finance, insurance and real estate, rental, hiring and leasing, manufacturing, mining and quarrying	7.70

Page N

#multi-hop #layout-navigating

Q: From the list of Top 10 Key Recovery Components, which is the last component listed on the second page?

A: Hope

#abstractive #graphic-intensive

Q: Does this document contain any checkboxes?

A: No

Document Visual Question Answering

- Foundation models ?
- It is adapted to large batches of documents ?



Thank you !

kermorvant@teklia.com

T E K L I A