

# Infrastructures pour corpus vivants

de l'OCR à l'édition collaborative des télégrammes de Vichy

Vincent Martin-Schreiber   Florian Mathieu   Jasmin Macarios

## La découverte du corpus

# Découverte du corpus

- ▶ Recherche à **Bibliothèque et Archives Canada (BAC)**
- ▶ Découverte du fonds **RG 24 C 22**
  - ▶ Bobines T-17425 à T-17429
- ▶ Archives peu exploitées computationnellement

**Une question immédiate :** peut-on rendre ces archives computationnellement exploitables ?

Reel/Bobine T-17425 - T-17429						24-171
RG	Reel / Bobine	Titre / Title	Dates	Date 1	Date 2	
RG24-C-22	T-17425	Signal intercepted from - Free French Diplomatic traffic - Message number FG3861 to FG7117	1944/0 - 1945/7	1944.13	1945-07-31	
RG24-C-22	T-17426	Signal intercepted from - Free French Diplomatic traffic - Message number FG21001 to FG 3965	1944/01 - 1944/10	1944-01-01	1944-10-31	
RG24-C-22	T-17427	Signal intercepted from - Free French Diplomatic traffic - Message number FG1 to FG1980	1943/04 - 1944/01	1943-04-01	1944-01-31	
RG24-C-22	T-17427	Signal intercepted from - Vichy French Diplomatic traffic - Message number D3747 to D5400	1943/01 - 1945/01	1943-01-01	1945-01-31	
RG24-C-22	T-17428	Signal intercepted from - Vichy French Diplomatic traffic - Message number D1 995 to D3746	1943/04 - 1943/01	1943-04-01	1943-01-31	
RG24-C-22	T-17429	Signal intercepted from - Vichy French Diplomatic traffic - Message number D11 to D3100	1941/09 - 1943/04	1941-09-01	1943-04-30	

SECRET

From: Buenos Aires  
To: Vichy

Dated: Nov. 28, 1942  
Rec'd: Nov. 30, 1942

830

Le sabordage de la flotte française par les équipages a provoqué hier soir une émotion particulière dans la population de Buenos Aires.

Ce matin, les journaux rendent un hommage unanime à l'héroïsme de nos marins; par de tels sacrifices, écrivent-ils, la France immortalisera les épreuves (effroyables) qui se sont (abattues) sur elle.

Le comité France-Amérique organisera au début de la semaine prochaine un service religieux pour les marins morts à Toulon.

LATOURNELLE

R-3557-9

From: Buenos Aires  
To: Vichy

Dated: Nov. 28, 1942  
Rec'd: Nov. 30, 1942

832

Votre tél. 665X.

M. EUGENE COLSON inconnu en Argentine.

x-D-3464

LATOURNELLE

R-3561

File D-3502

Examination Unit,  
National Research Council  
December 1, 1942

SECRET

Ce qui a déclenché le projet :

- ▶ Structure **claire et homogène**
- ▶ Texte **tapé à la machine**
- ▶ Métadonnées **standardisées** : numéro, date, expéditeur, destinataire, etc.
- ▶ Corpus complet



## Le corpus en chiffres

---

**13 848** pages numérisées  
~**11 000** télégrammes individuels  
**5** bobines de microfilm  
**4 ans** sept. 1941 – juil. 1945

---

Accessibles via le protocole **IIIF** — [heritage.canadiana.ca](https://heritage.canadiana.ca)

## Contexte historique

# Vichy et la France Libre

- ▶ **Juin 1940** : armistice franco-allemand, naissance du régime de Vichy
- ▶ Deux légitimités françaises en conflit :
  - ▶ **Vichy** (Pétain, gouvernement de fait)
  - ▶ **France Libre** (de Gaulle, Londres puis Alger)
- ▶ Réseaux diplomatiques actifs malgré l'Occupation
- ▶ Communications chiffrées vers ambassades et consulats



Figure 1: Bundesarchiv, Bild 183-H25217 / CC-BY-SA 3.0

# L'Examination Unit canadienne

- ▶ Agence canadienne de **cryptanalyse et de renseignement sur les signaux (SIGINT)**
- ▶ Active de **1941 à 1945**
- ▶ Déchiffre les communications diplomatiques de Vichy et de la France Libre
- ▶ Rattachée au **Conseil national de recherches du Canada (CNRC)**
- ▶ Coordination avec les Alliés (Bletchley Park)

## CANADA'S BLETCHLEY PARK

The Examination Unit  
in Ottawa's Sandy Hill  
1941–1945



REVISED EDITION

## Intérêt de ces archives

*Accessible depuis les années 1990, mais inutilisable computationnellement*

- ▶ Source primaire pour l'histoire de la **politique étrangère de Vichy**
- ▶ Archives **peu exploitées** par les historiens

## Problème et objectifs

# Des archives inaccessibles computationnellement

## Avant le projet

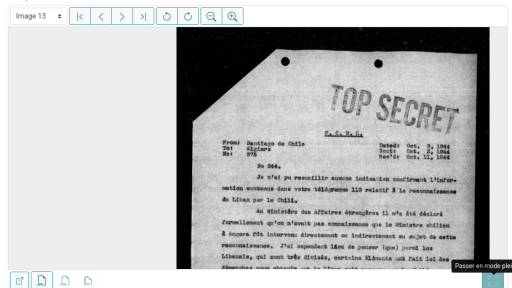
- ▶ Format image uniquement (JPEG)
- ▶ Aucune recherche possible
- ▶ Aucune extraction de données
- ▶ Transcription manuelle : ~2 300 heures



Conditions d'utilisation English Français

À propos Canadiana RCDR Quoi de neuf Guide de citation Nous joindre

Department of National Defence : Examination Unit : T-17425



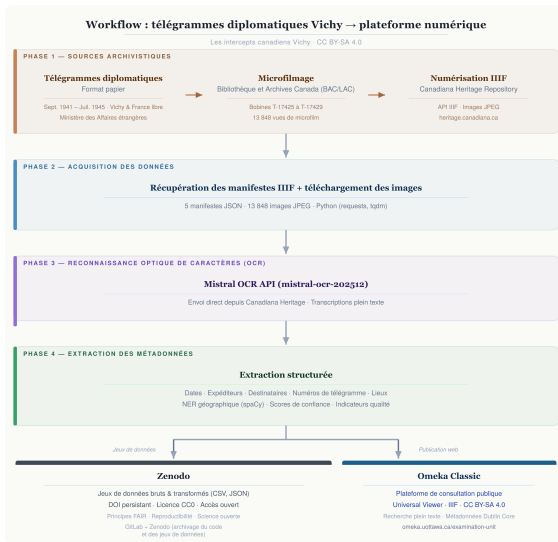
# Objectifs

- ▶ Rendre le corpus accessible
- ▶ Obtenir des données aisément utilisables (trouver = une priorité)



## Le pipeline

# Vue d'ensemble



# Phase 1 — Sources archivistiques

## PHASE 1 — SOURCES ARCHIVISTIQUES

### **Télégrammes diplomatiques**

Format papier

Sept. 1941 – Juil. 1945 · Vichy & France libre  
Ministère des Affaires étrangères



### **Microfilmage**

Bibliothèque et Archives Canada (BAC/LAC)

Bobines T-17425 à T-17429  
13 848 vues de microfilm



### **Numérisation IIIF**

Canadiana Heritage Repository

API IIIF · Images JPEG  
[heritage.canadiana.ca](https://heritage.canadiana.ca)

# Phase 2 — Acquisition des données

## PHASE 2 — ACQUISITION DES DONNÉES

### **Récupération des manifestes IIIF + téléchargement des images**

5 manifestes JSON · 13 848 images JPEG · Python (requests, tqdm)

# Phase 3 — Transcription

PHASE 3 — RECONNAISSANCE OPTIQUE DE CARACTÈRES (OCR)

## **Mistral OCR API (mistral-ocr-202512)**

Envoi direct depuis Canadiana Heritage · Transcriptions plein texte

From: Buenos Aires

To: Vichy

Dated: Nov. 28, 1942

Rec'd: Nov. 30, 1942

830

Le sabordage de la flotte française par les équipages a provoqué hier soir une émotion particulière dans la population de Buenos Aires.

Ce matin, les journaux rendent un hommage unanime à l'héroïsme de nos marins; par de tels sacrifices, écrivent-ils, la France immortalisera les épreuves (effroyables) qui se sont (abattues) sur elle.

Le comité France-Amérique organisera au début de la semaine prochaine un service religieux pour les marins morts à Toulon.

LATOURNELLE

R-3559-9

# Phase 4 — Extraction des métadonnées

## PHASE 4 — EXTRACTION DES MÉTADONNÉES

### Extraction structurée

Dates · Expéditeurs · Destinataires · Numéros de télégramme · Lieux  
NER géographique (spaCy) · Scores de confiance · Indicateurs qualité

# Défis spécifiques

## Images multi-télégrammes

**SECRET**

From: Bogota  
To: Washington  
Dated: September 15, 1942  
Rec'd: September 16, 1942

7

M. VANSE, ancien secrétaire (gaulliste) à Mexico City, part pour le (Chili) en qualité de "représentant personnel du Général DE GAULLE". Il a essayé de fonder une véritable légation à Bogota mais, officiellement du moins, les autorités colombiennes lui en ont refusé le droit. D'autre part je serais heureux de savoir si M. VANIN a bien remis le pli que je lui ai confié.

RELOUIS

Dated: September 15, 1942  
Rec'd: September 16, 1942

From: Habana  
To: Washington  
46

Pour l'Attaché naval. Les travaux de remise en état sommaire de l'INDIANA s'effectuent normalement. Dépenses prévues pour cette opération sont environ \$11,000. Elles seront payées intégralement à Colon par l'agent Compagnie Général Trans-atlantique qui dispose des fonds nécessaires. Les autorités américaines prêtent un concours très cordial et très utile à notre équipage.

LEONISSEL

Dated: September 16, 1942  
Rec'd: September 19, 1942

From: Washington  
To: Vichy  
3405

Mss tils. 3541<sup>x</sup> et 3502<sup>xx</sup>.  
Autorisation sortir des E.-U. Mss. UPDENRAPP expirant avant fin du mois, je serais reconnaissant à V.E. de bien vouloir signaler d'urgence sa décision que Miss ANNE (MORGAN) espère ...  
HENRY-HAYE  
X D-2887  
XX D-3746

Examination Unit,  
National Research Council,  
September 19, 1942.

**SECRET**

## Télégrammes multi-pages

**TOP SECRET**

From: Paris  
To: Ottawa  
Nos: 132-137  
Dated: Oct. 6, 1944  
Sent: Oct. 7, 1944  
Rec'd: Oct. 11, 1944

Circulaire no 2.

La Fédération de la Presse Française s'est étonnée des ordonnances sur la presse publiées les 1<sup>er</sup> et 2 octobre 1944 et qui prévoient essentiellement:

- I. Interdiction de tous les journaux et périodiques qui, existant avant le 22 juin 1940, ont continué à paraître plus de 15 jours après l'armistice (133) en zone nord et plus de 15 jours après le 11 novembre 1942 en zone sud.
- II. Interdiction des journaux donnant lieu à poursuite. L'interdiction est maintenue jusqu'à jugement ou à décision de non-lieu. A défaut de poursuite l'interdiction prendra fin au bout de six mois.
- III. A compter du 1<sup>er</sup> avril 1945 la (134) carte d'identité professionnelle sera obligatoire.
- IV. Le Ministre de l'Information fixe les prix de vente et répartit les contingents de papier. La Fédération a demandé audience au Général DE GAULLE. En même temps paraissent dans "Franc-Tireur", "Combat" et "L'Humanité" des éditoires se plaignant de ce que la Fédération (135) n'ait pas été consultée sur le texte des ordonnances. Au cas où la presse du pays de votre résidence ferait état de ces incidents ou si des questions vous étaient posées à ce sujet, mais dans ces cas seulement, je vous serais obligé de faire état des indications ci-après:

- I. La presse (136) bénéficie d'une liberté totale. Les critiques elles-mêmes des journaux concernant les ordonnances prises en sont la preuve.
- II. Celles-ci ont été signées par le Général DE GAULLE après avoir été adoptées en séance extraordinaire. Le Gouvernement dans son ensemble en est donc avec son Chef principalement responsable.

File PG- 3671 (continued) Examination Unit,



Champs extraits automatiquement :

- ▶ Dates, numéros de télégramme
- ▶ Expéditeurs et destinataires
- ▶ Lieux (NER géographique — spaCy)

Champ	Taux de succès
Numéros	~95 %
Dates	~80 %

	A	B	C	D	E	F	G	H	I	J	K
	image_id	telegram_count	has_multiple_telegrams	is_multipage	from_location	to_location	via_routing	message_number	message_number_type	date_sent_parsed	date_received_parsed
1	t17425_0001	0	False	False	0	0	0	0	0	0	0
2	t17425_0002	0	False	False	0	0	0	0	0	0	0
3	t17425_0003	1	False	False	Paris	Washington	0	31-32	range	1944/10/03	1944/10/11
4	t17425_0004	1	False	False	Algiers	Ottawa	0	28720-28722	range	1944/10/10	1944/10/11
5	t17425_0005	1	False	True	Washington	London	0	0	0	0	1944/10/10
6	t17425_0006	1	False	True	0	0	0	0	0	0	0
7	t17425_0007	1	False	True	0	0	0	0	0	0	0
8	t17425_0008	1	False	False	Washington	Algiers	0	0	0	0	1944/10/11
9	t17425_0009	1	False	False	Washington	Algiers	0	0	0	0	1944/10/11
10	t17425_0010	1	False	True	0	0	0	0	0	0	0
11	t17425_0011	1	False	True	0	0	0	0	0	0	0
12	t17425_0012	1	False	True	Washington	London	0	0	0	0	1944/10/10
13	t17425_0013	1	False	False	Santiago de Chile	Algiers	0	275	single	1944/10/03	1944/10/11

# Phase 5 — Diffusion

## Zenodo

Jeux de données bruts & transformés (CSV, JSON)

DOI persistant · Licence CC0 · Accès ouvert

Principes FAIR · Reproductibilité · Science ouverte

GitLab + Zenodo (archivage du code  
et des jeux de données)

## Omeka Classic

Plateforme de consultation publique

Universal Viewer · IIIF · CC BY-SA 4.0

Recherche plein texte · Métadonnées Dublin Core

[omeka.uottawa.ca/examination-unit](https://omeka.uottawa.ca/examination-unit)

# The Canadian Vichy Intercepts

[HOME](#)[THE COLLECTION](#)[SEARCH GUIDE](#)[BIBLIOGRAPHY](#)[COLLECTIONS](#)[ABOUT](#)

Explore nearly 14,000 declassified pages of diplomatic telegrams intercepted and deciphered by Canada's Examination Unit during the Second World War. This unique collection reveals Vichy France's wartime diplomacy through SECRET-classified communications between Washington, Vichy, Ottawa, and other diplomatic posts from 1941 to 1945.

These documents—once classified SECRET and now digitized—offer unprecedented insights into international diplomacy, cryptographic intelligence, and Canada's early signals intelligence capabilities during a critical period of world history.

## Collection at a Glance

**13,847 Documents | 1941–1945 | 5 Microfilm Reels | French & English**

- First civilian cryptographic operations in Canadian history
- Diplomatic correspondence between major Allied and Axis powers

Le corpus vivant

## Principe directeur : l'ouverture

- ▶ code disponible (GitLab)
- ▶ documentation
- ▶ jeux de données (Zenodo)
- ▶ utilisation de logiciels libres et développés par une communauté (Python, Omeka)
  - ▶ **MAIS** : Mistral AI (très) fermé - n'a pas répondu à nos sollicitations

# D'un processus unidirectionnel à un écosystème participatif

## **Modèle traditionnel**

Institution → Public

Corpus figé, produit par un seul acteur

## **Corpus vivant**

Institution <-> Communauté

Corpus dynamique, co-construit et évolutif

# Scripto : co-construction communautaire

## Plugin Omeka Scripto

- ▶ Image originale <-> texte transcrit  
côte à côte
- ▶ Validation visuelle immédiate
- ▶ Chaque utilisateur devient  
**contributeur potentiel**



# Gouvernance communautaire

Système de validation en trois étapes :

**Transcription → Révision → Approbation**

- ▶ Pages de discussion **MediaWiki** pour les cas ambigus :
  - ▶ abréviations diplomatiques
  - ▶ lieux historiques
  - ▶ conventions documentaires
- ▶ Versionnement sur Zenodo : l'évolution du corpus est traçable

## Qui pourrait contribuer ? (tout le monde)

Profil	Contribution
Historiens	Interprétation, contextualisation
Linguistes	Français des années 1940, abréviations
Archivistes	Conventions documentaires, lieux
Spécialistes des RI	Réseaux diplomatiques
Citoyens	Transcription, corrections

## Exemples d'utilisation

# Trouver et explorer le corpus

## **Omeka Classic**

- ▶ Recherche plein texte sur les transcriptions OCR
- ▶ Filtrage par métadonnées (date, expéditeur, destinataire)
- ▶ Visualisation IIIF avec Universal Viewer

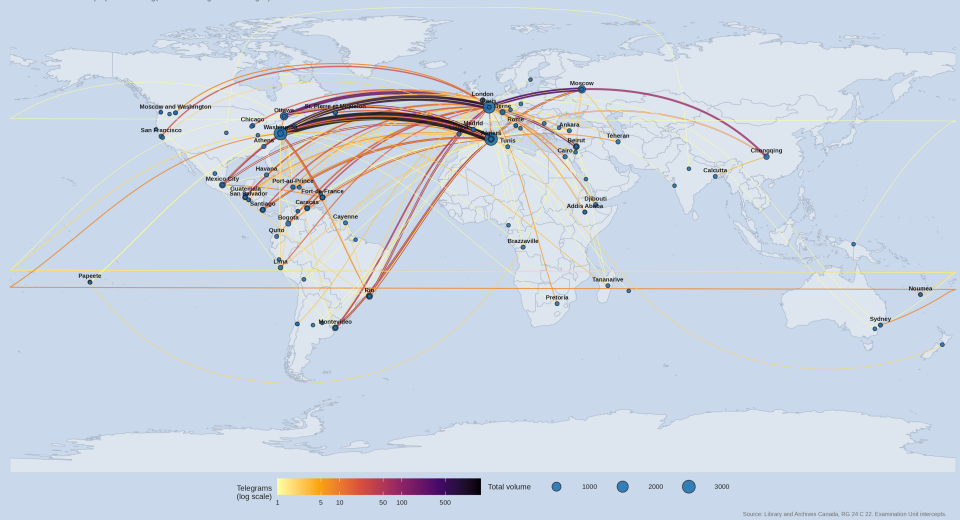
`omeka.uottawa.ca/examination-unit`

- ▶ tag
- ▶ requêtes booléennes

# Cartes et analyse de réseaux diplomatiques

## Free France Diplomatic Telegrams — Interception Network (1941–1945)

Line width and colour proportional to log(number of telegrams exchanged)



## Extensions prévues (non planifiées)

- ▶ **Analyse de réseaux** : cartographie des acteurs et des flux
- ▶ **Fouille de textes** : thématiques diplomatiques, évolution du discours
- ▶ **NER complet** : personnes, lieux, organisations
- ▶ **Cryptanalyse historique** : reconstruction des méthodes de l'Examination Unit
- ▶ **Édition savante** : annotation sémantique, contextualisation historiographique

# Planification et budget

- ▶ cout
- ▶ durée

## Perspectives et conclusion



## Projets à venir

- ▶ Court terme
- ▶ Activation de Scripto ?
- ▶ NER complet sur le corpus - amélioration de la nomenclature
- ▶ Publications en préparation (Humanistica, Cryptologia ou équivalent)
  - ▶ tentative d'évaluation des impacts environnementaux et sociaux

# Le corpus vivant comme modèle

- ▶ **Triple dialogue** : extraction automatique <-> expertise humaine · images <-> textes · contributeurs entre eux
- ▶ Versionnement documenté sur Zenodo : chaque intervention est traçable
- ▶ De l'archive individuelle à l'**intelligence collective**

L'enrichissement continu *est* la production historiographique



Merci

---

Omeka



GitLab



Zenodo

---

