Transcribing Handwritten Documents Using Artificial Intelligence

Presentation to the Canadian Research Knowledge Network (CRKN) Virtual Conference

Maxime Gohier, UQAR   October 13, 2020
Rénald Lessard, BAnQ
Jean-François Palomino, BAnQ

Nouvelle France NUMÉRIQUE

BAnQ

Presentation Plan

- Digitalization and dissemination within BAnQ
- The Digital New-France Project
- Technology for re-transcription and digital editions of handwritten documents
- BAnQ corpus as a trial project

BAnQ Mission

The purpose of the BAnQ is to collect, conserve and disseminate the documentary heritage of Quebec.

This heritage consists of millions of documents which are a testimony to the history and culture of the Quebecois.

The First Revolution: Digital Collection and Dissemination

- An effective way that helps BAnQ promote research of its collections and holdings
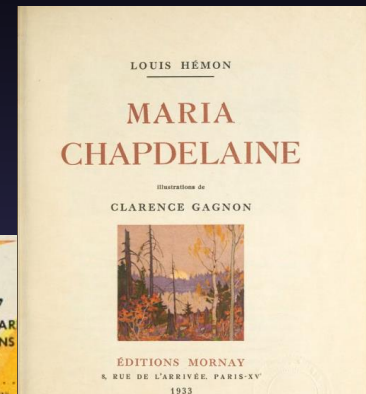
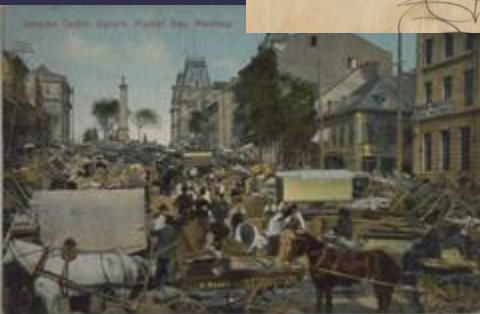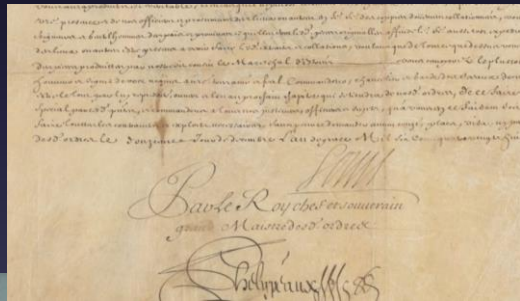# Digitization and Dissemination within BAnQ

- Digital services have continued to grow over the last 20 years

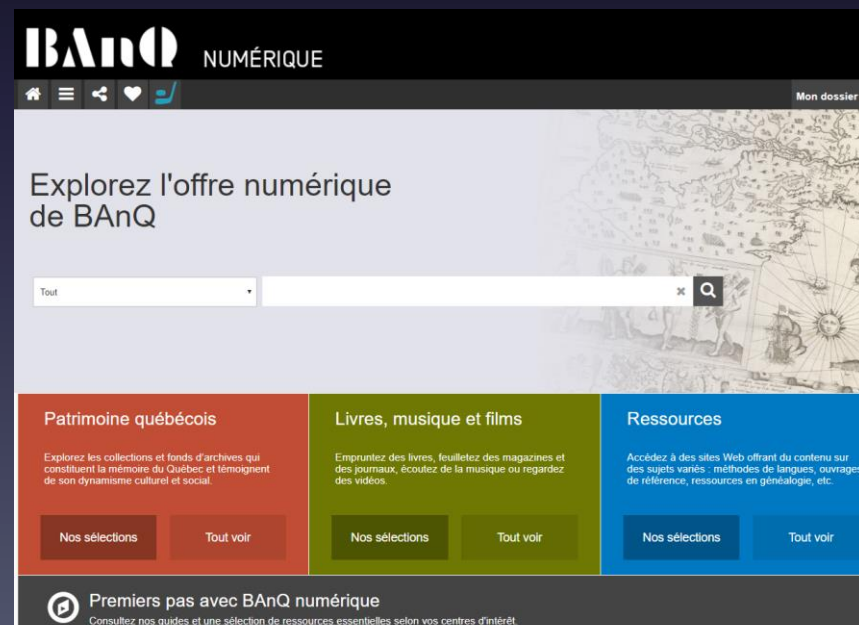*Insert here a graphic on the annual volume of digitized documents or page*

Digitization and Dissemination within BAnQ

- A variety of digitized, published and handwritten documents

Digitization and Dissemination within BAnQ

- Digitization of several million files, disseminated through the **BAnQ numérique** and the **Advitam** catalogue

The Second Digital Revolution: Optimal Character Recognition (OCR)

Launch search using keyword "Papineau" in the **BAnQ numérique**: 264,750 results !

Huge research potential.

BAnQ is looking for potential partnerships to:

- Provide better visibility for institutions and their documentary corpus;
- Promote and facilitate research of collections and holdings;
- Promote better knowledge of the history and culture of Quebec, based on documents in BAnQ.

UQÀM | LHPM
Laboratoire d'histoire et
de patrimoine de Montréal
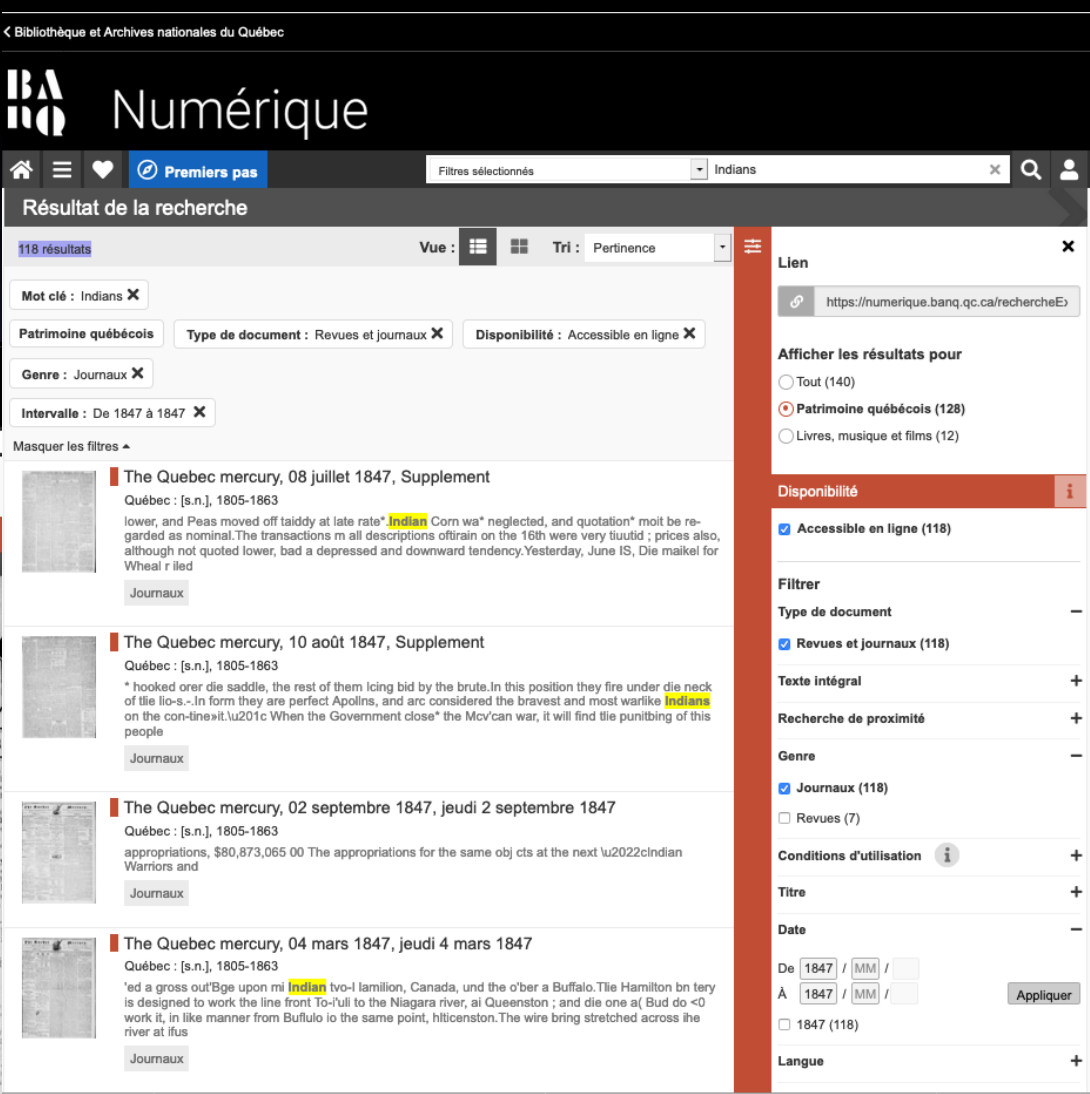
THÉÂTRES ASSOCIÉS

{BnF

WIKIMEDIA
CANADA

(…)

BAnQ is looking for potential partnerships

A partner to work with BAnQ in the next revolution of documentary research...

The Digital Revolution and *Big Data*

A worldwide movement

Gutenberg Project
- Since 1971 more than 53,000 books have been digitally transcribed

ICMH and *Early Canadiana Online*;
Canadiana.org
- Since 1999: 19 million Canadian print documents are available online

BAnQ numérique
- Since 2003: 5,873 journals; 5.3 million journal pages and Quebecois journals

GoogleBooks
- Since 2004: 25 millions books have been scanned in every language
- Ngram Viewer

The Digital Revolution

# The Digital Revolution
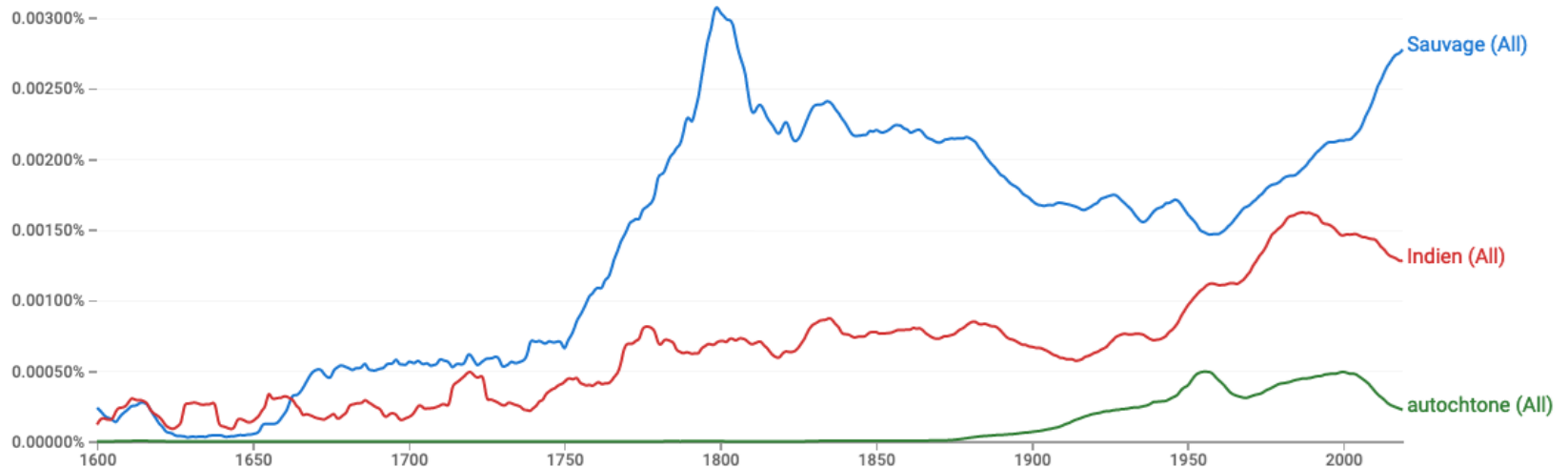
Challenges of OCR

- Relative quality of OCR: Creation of an "illusionary order"

Il nous dit ensuite que nous avons mal traduit le mot " *the Indians*" qui signifie, selon lui, *les Indiens*, et non les sauvages. Si l'écrivain du *Transcript* était plus fami- lier avec la langue française, il saurait que nous désignons plus souvent sous le nom " d'Indiens" les habitants des Indes, que les habitants des forêts du nouveau monde où la civilisation n'a pas encore pénétré ; il saurait que le mot " sauvage" a deux ac- ceptions en français. Il ne signifie pas toujours " féroces et barbares," puis qu'on dit en français, " les sauvages civilisés, les sauvages instruits, les sauvages du Sault St. Louis, les sauvages de St. Régis, les sauvages de St. François, &c." Mais on se sert presqu'invariablement de ce mot pour désigner les habitants primitifs de nos fo- rêts, les naturels du pays, toutes les peu- plades sauvages, en un mot. Il nous sert

11 uous dit ensnile ue nous avnus mal
Cruduit le mut ** the Indians? qui signitie,
selon Ini, les Indiens, ex non les sauvages.
Ni l'écrivain du Transcript était plus fummi-
per aveo a langue française, «1 sausait que
ous d ignous plus avuveut xoux le nom
« d'ludiens" les hubitante des ludes, que
les habitants des tucets du nouveau moude
où la civilisation n'a jus encore pénétré
aurai que le mot a get deux ues
ceptions en françai, l ue signilie pas
toujours " féroeus et barbarue,*" pois qu'
Jiten francais, * les sauvages ci vilises, jen
auvages Fruits les ages du Sanit
o Lonis, les sauvages de SL [Reis Jon
guvages de St, François, Le."

Challenges of OCR

- More democratic access and document marginalization: the arduous task of working with manuscripts
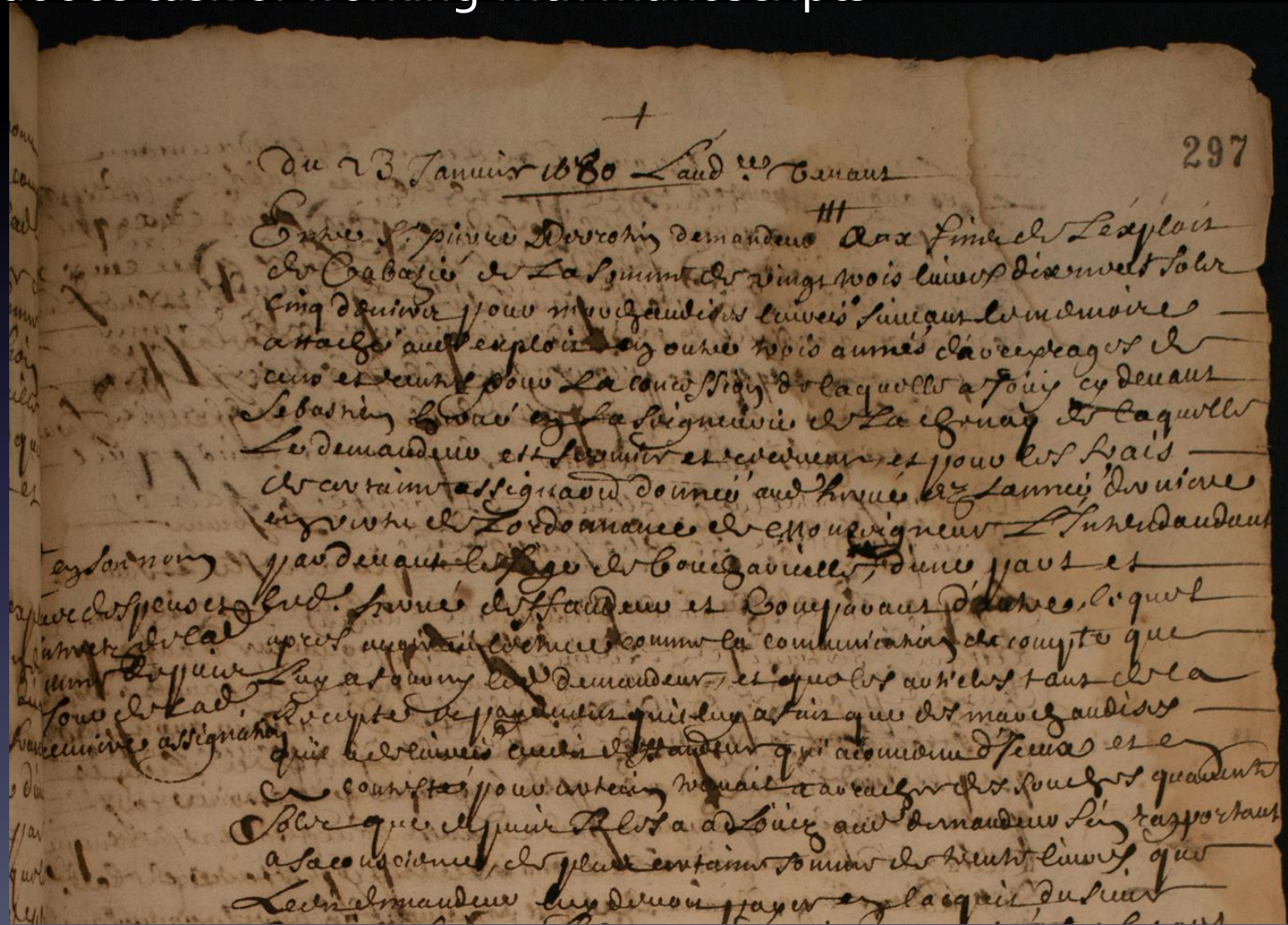
Research Challenges Working with Handwritten Documents in the Digital Era

- Access to archives: democratization vs marginalization

- The need for paleographic expertise

- Data produced by researchers and archives is far from being uniform

- Variability of document types

- Sustainability of research data

Digital New-France:
A Large-Scale Digital Publication Project

Goal:

- Develop technology tools which will allow researchers to pool their research data, thus facilitating consultation and digital processing, as well as the continuous enrichment of research.

Historical-Literary Approach: Document as Text and Object

The materiality of the document (structure, cursive, deletions, illustrations, etc.) provide information about the writing process and the historical nature of the document.

A Collaborative *and* Research Partnership Project

Goals:

- Ensure compatibility and inter-operability of digital data generated by researchers;

- Ensure sustainability of data: allow the possibility of ongoing correction and gradual enrichment;

- Ensure that access to open-source data is based on the respect for the mandates of various archival centres.

Collaborative *and* Partnership Approach

1- Develop a digital repository to share documentary research by scholars;

2- Develop working protocols to ensure the uniformity and compatibility of data produced;

3- Support archival institutions in their mission to facilitate accessibility and improvement of their collections using digital tools.

The Digital New-France Project: Digital Repository and Research Tool

- Scanned images of archive documents

- Automated transcription

- Collaborative working protocols

- Structural and textual tags (encryption)

- Expertise in paleography (permanent workshop on documentary analysis by Dominique Deslandres)

- Research and consultation interface (next step)

Partners

- Library and Archives Canada
- Bibliothèque et Archives nationales du Québec
- Rare Books and Special Collections Library, University of Montreal
- Document and Archives Management Division, University of Montreal
- Museum of Civilization of Quebec
- READ Cooperative (Transkribus)

Others

- Archives nationales d'Outre-Mer (Aix-en-Provence)
- Abenaki Museum in Odanak

An Extensive Team… Destined to Grow

Management Team:

- Catherine Broué (UQAR)
- Maxime Gohier (UQAR)
- Léon Robichaud (Université de Sherbrooke)
- Jean-François Palomino (BAnQ)

Other Researchers:

- Mehdi Adda (UQAR)
- Alain Beaulieu (UQAM)
- Arnaud Bessières (Université Laurienne)
- Philippe Charland (Collège Kiuna)
- Colin Coates (Glendon College)
- Dominique Deslandres (Université de Montréal)
- Helen Dewar (Université de Montréal)
- Alexandre Dubé (Washington University)
- Gilles Havard (CNRS-EHESS)
- Daniella Kostroun (Indiana University)
- Jean-François Lozier (Université d'Ottawa)
- Marie-Christine Pioffet (York University)
- And many more to come…

Multiple Corpora

- Collections and minutes of French-Indigenous diplomatic encounters (Maxime Gohier)

- Reports on the exploration of Louisiana (Catherine Broué)

- Indigenous language dictionaries (Philippe Charland and Marie-Christine Pioffet)

- Geographical and toponymy maps (Jean-François Palomino)

- Land grids in Montreal (Dominique Deslandres)

Technology for the Digital Publication of Handwritten Documents

- The READ Cooperative (Recognition and Enrichment of Archival Documents) and Transkribus software
- XML-TEI mark-up language

Transkribus Software

- Text spatialization

- Handwritten Text Recognition (HTR)

- Mark-up language

- Research

Transkribus Software and Functions
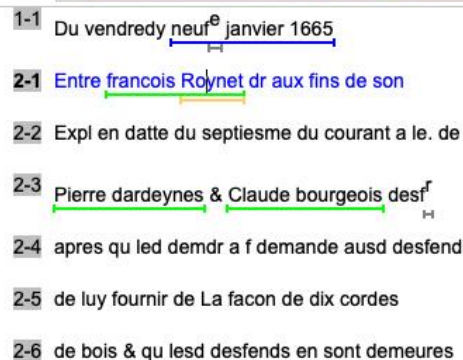
Spatialization of Transcriptions Within the Image

Handwritten Text Recognition (HTR)

Details

Name:	Language:
Français ANOM	French

Description:	Parameters:
Documents from series F3, C11A and C11E. 17th and 18th Centuries	Nr. of Epochs 200

Show advanced parameters...

Nr. of Words:	Nr. of Lines:
128175	16392

Save	Show Train Set	Show Validation Set	Show Characters

Learning Curve

Accuracy in CER (100% – 0%) vs Epochs (0 – 200)

— CER Train   — CER Validation
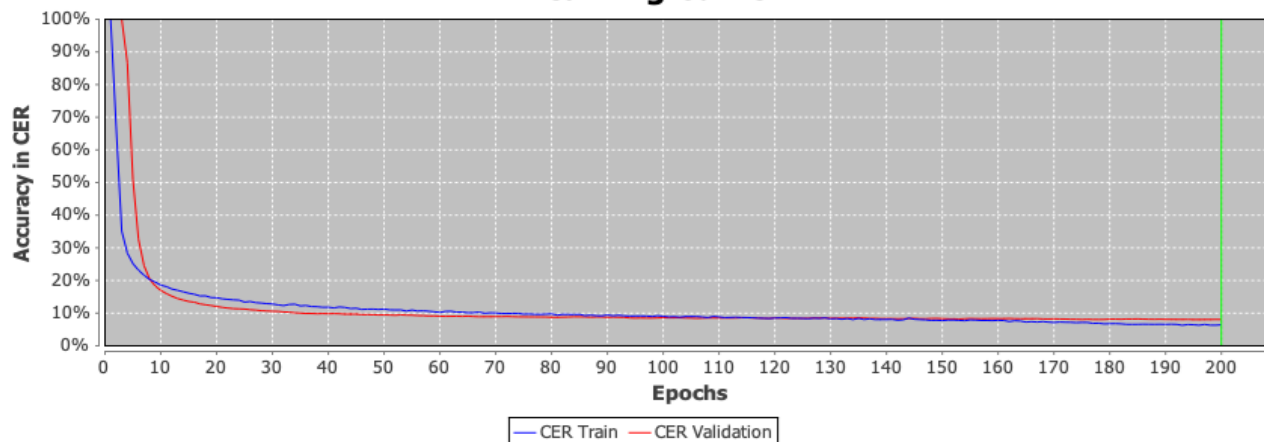
CER on Train Set:	6.43%	CER on Validation Set:	8.05%

Details

Name:

| Relation de 1668 |

Language:

| French |

Description:

| 18th Century |

Parameters:

| Nr. of Epochs 50 |

[Show advanced parameters...]

Nr. of Words:

| 13856 |

Nr. of Lines:

| 1873 |

[💾 Save] [Show Train Set] [Show Validation Set] [Show Characters]

Learning Curve

CER Train — CER Validation
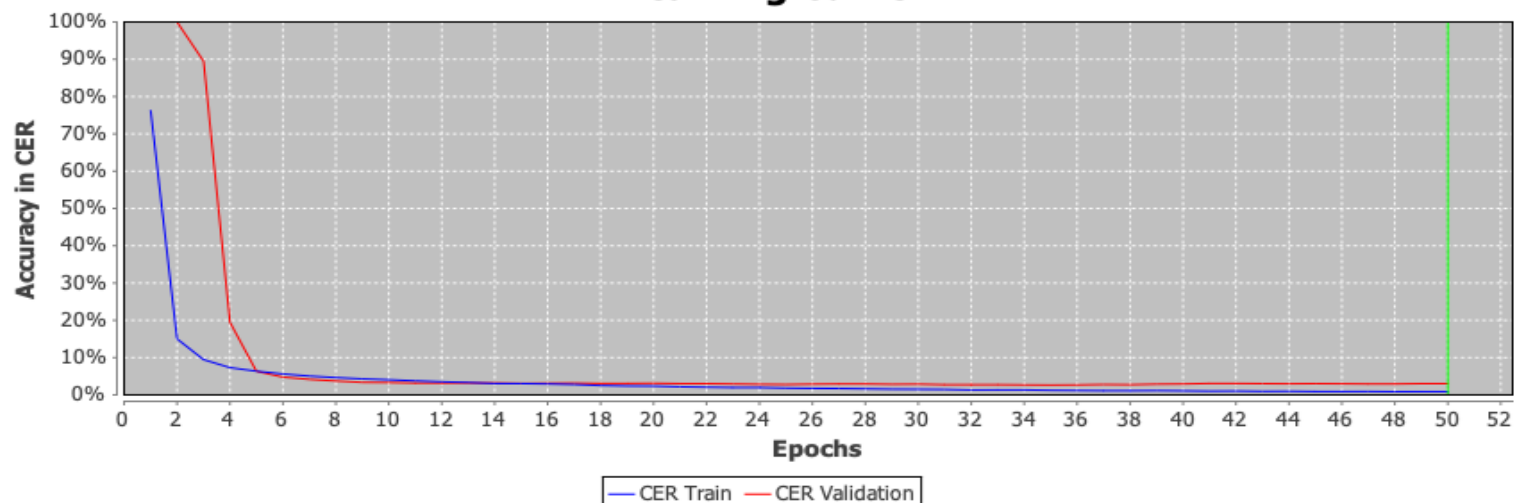
CER on Train Set: | 0.87% | CER on Validation Set: | 3.02% |

Name:	Language:
Théodore De Pincier	French

Description:

Journal d'arpentage de Théodore de Pincier, 1798-1799 (35 pages)

Parameters:

Nr. of Epochs 200

Show advanced parameters...

Nr. of Words:	Nr. of Lines:
6277	558

Save	Show Train Set	Show Validation Set	Show Characters

Learning Curve

— CER Train — CER Validation

Epochs

Accuracy in CER
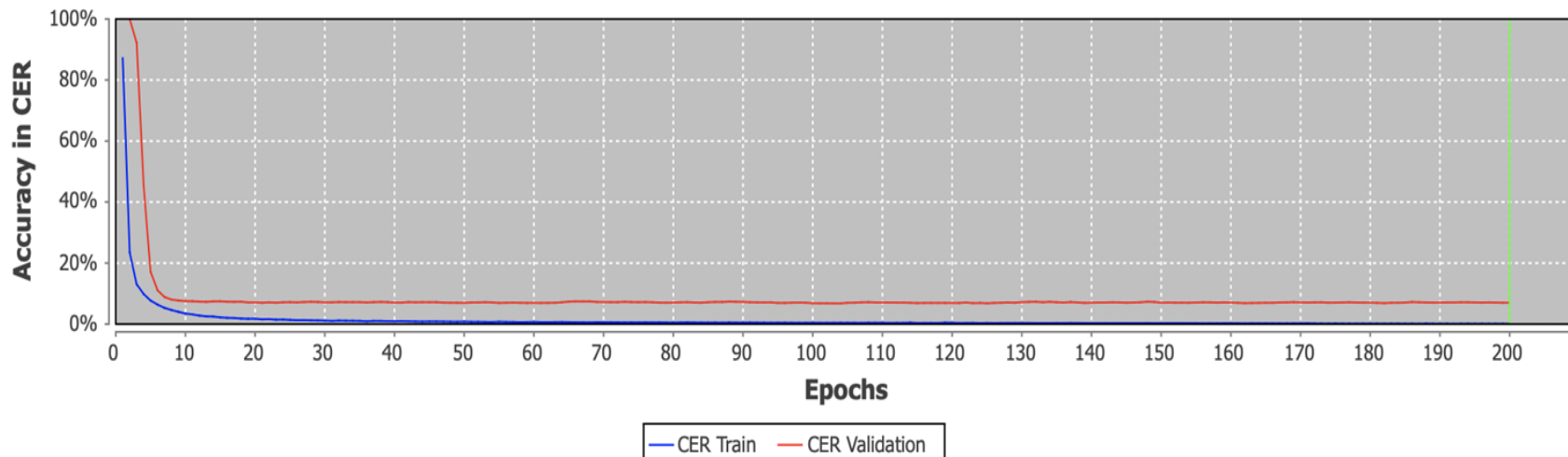
CER on Train Set:	0.11%	CER on Validation Set:	6.98%

du xiiie desd mois & an a le. de Charles Martin
desfr apres qu led dr a ft demande aux desfr
de la somme de vingt une livres ts a chacun

a payer de trente livres ts pour le louage de la
maison appt au nomme St. George suivant le
transport ft par led dr aud desfr de louage
d'Icelle maison Ensemble de Reparer la Closture
des pieux a la Reserve de vingt ung pieux dune part

et den a donné autre part ld desfr est dem
daccord ordonne que led desfendr payera
Donc lad somme de vingt une livres ts
& ce dans quinzaine dhuy a faulte fera fr lad closture
Dans Le premier temps commode a la defense

Lesd et aux desps

Dessailly

DeMonchy
Greffier

dud Jour et an que

Entre charles LeMoyne dr aux fins
de son expl du xiiie du pn mois & an
a le de pierre Raguedeau desfendeur
apres qu led Dr a ft damande aud desfendeur
de la somme de Cent dix livres ts
tant pour marchandises fournyes par led
dr aud desfend, quargent deub a lesglise
de ce lieu duql le dr est dispensable
& que led desfr a ft demande de Conte

What is the Text Encoding Initiative (TEI)?

" XML-TEI provides a framework to represent everything we deem important about a text and not uniquely its appearance, so that software can process different characteristics that have been identified, thus generating new visualizations and new perspectives."

(Lou Burnard, *Qu'est-ce que la* Text Encoding Initiative*?*, Marseille, OpenEdition Press, (n.d.) [en ligne:](généré le 04 novembre 2015), p. 10.

XML-TEI Mark-Up Language

- Identifying and indexing of identified entities (surnames, place names, ethnic names) and dates (numbered or text format);
- Identification and resolution of abbreviations;
- Identification of ambiguous passages and alternate proposals;
- Identification of structural components of text (paragraphs, page numbers, titles, signatures, etc.);
- Identification of non-textual entities (tables, graphs, illustrations, etc.);
- Identification of successive interventions regarding a document (additions and deletion of passages, changes to cursive style);
- Clarifications concerning material support and its state of conservation;

Structural tags are used to provide a summary of the form.

- Titles and subtitles
- Paragraphs
- Page numbers
- Signatures
- Legend
- Drawings
- Greeting or salutation
- etc.

Identification and Description of Non-Textual Entities
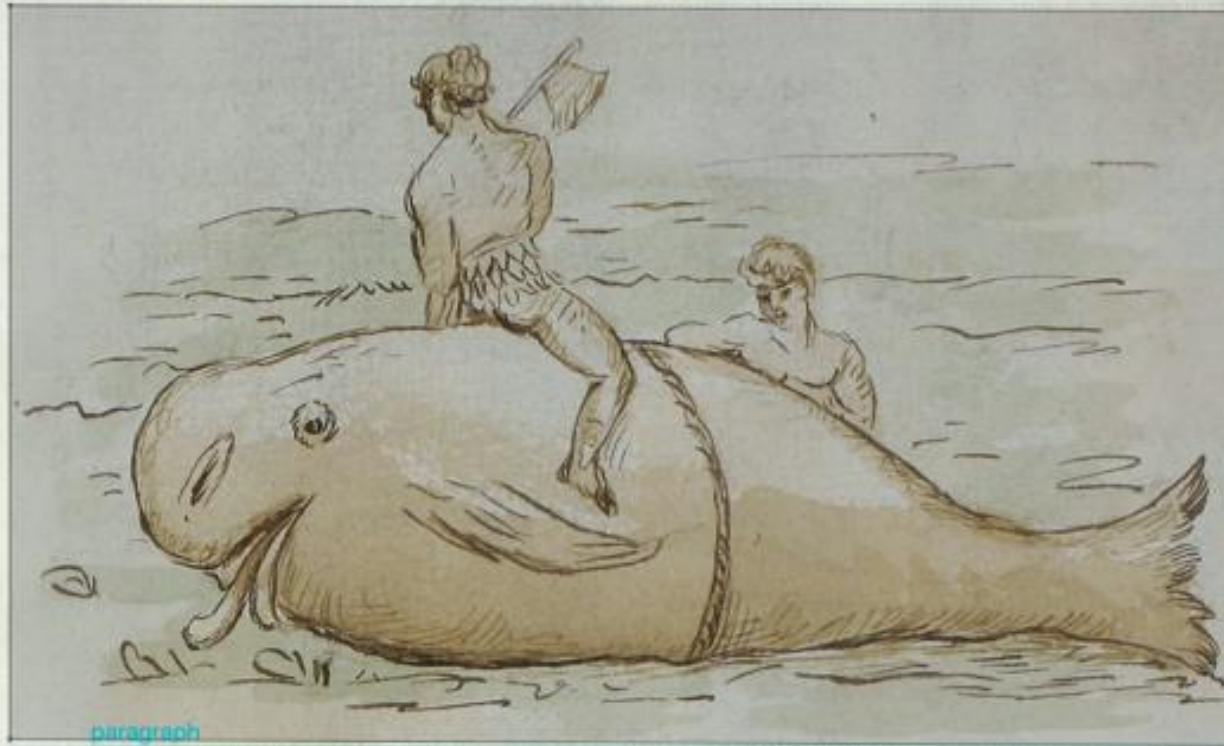
se concurent contre le froid. Qui leur monstreroit à vinure, & leur enseigneroit le labourage des terres, & autres choses, ils apprendroient fort bien: car

Textual mark-ups are used to take into account the contents and to facilitate tracking.

Easy-to-use, including:
- Copy-paste-click
- Tags and mark-ups can be personalized at all times

Textual Mark-Ups are Used to Underscore the Writing Process and the Historical Authenticity of Documents

The database consists of various corpora ensuring cross reference searches (ex. toponymy or terms derived from Aboriginal languages in the glossaries).

Old maps contain a great deal of text in which the search function is not normally available.

Collective Construction

- A collective decoding operation
- Open data format
  - Possibility of correcting and improving transcription
  - Use of mark-ups is potentially infinite
  - The open nature of TEI-Consortium (possibility of adding tags depending on the research interests)
- Transkribus and READ-Coop
  - Open software
  - Cooperative model: funding and cost of use
  - From training to technology
- The more researchers continue to participate, the greater will be the utility of the database.

The BAnQ Corpus
Falls within scope of this Project

- Decrees issued by the Intendant of New-France (approximately 11,000 pages have already been scanned)
- Notary records (digital processing of these records is underway by Professor Dominique Deslandres)
- Geographical maps from New-France

Benefits for BAnQ

- BAnQ is active in a project that uses the most up-to-date tools of artificial intelligence
- BAnQ then receives textual retranscriptions produced by handwritten language recognition and can use these texts within its search engines (BAnQ numérique and Advitam)

Questions Arising During Project

- What are the rules to manage and share retranscriptions between researchers and partners?

- What is an acceptable level of error in light of re-dissemination in the BAnQ search engines (BAnQ numérique and Advitam)?

- What corpora should have priority?

- What form will be used for a potential central repository for these retranscriptions using the sharing and dissemination tool specific to each partner institution?

The Floor is Yours...