HAL
open science

Transcribing Foucault's handwriting with Transkribus

Marie-Laure Massot, Arianna Sforzini, Vincent Ventresque

To cite this version:

Marie-Laure Massot, Arianna Sforzini, Vincent Ventresque. Transcribing Foucault's handwriting with Transkribus. Journal of Data Mining and Digital Humanities, 2019, Atelier Digit_Hum, 10.46298/jdmdh.5043 . hal-01913435v3

HAL Id: hal-01913435

https://hal.science/hal-01913435v3

Submitted on 18 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Data Mining and Digital Humanities

Transcribing Foucault's handwriting with Transkribus

Marie-Laure Massot[1], **Arianna Sforzini**[2], **Vincent Ventresque**[2]

[1]CAPHÉS [UMS 3610 / ANR FFL]
[2]Triangle [UMR 5206 / ANR FFL]

Contents

1 The ANR project Foucault's Reading Notes **2**
1.1 Overview 2
1.2 The Foucault Archives at the Bibliothèque nationale de France 2
1.3 General methodological remarks: which philosophical use for these archives? . 3
1.4 How to explore the corpus with digital tools? 5

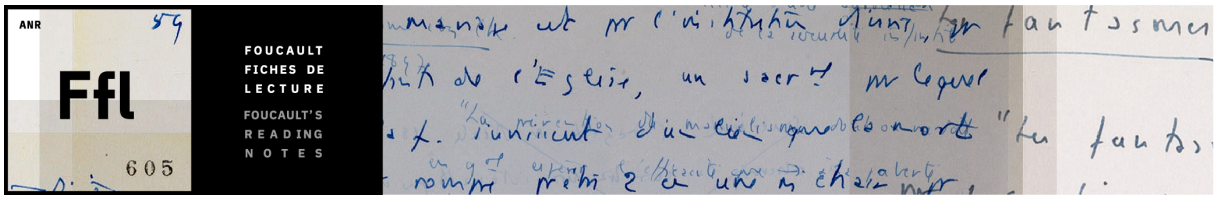
2 A tool for automatically transcribing manuscripts **8**
2.1 Creating learning data for Foucault's handwriting 9
2.2 Main difficulties encountered 11
2.3 Assessment of the experiment and prospects 14

Abstract

The Foucault Fiches de Lecture (FFL) project aims both to explore and to make available online a large set of Michel Foucault's reading notes (organized citations, references and comments) held at the BnF since 2013. Therefore, the team is digitizing, describing and enriching the reading notes that the philosopher gathered while preparing his books and lectures, thus providing a new corpus that will allow a new approach to his work. In order to release the manuscripts online, and to collectively produce the data, the team is also developing a collaborative platform, based on RDF technologies, and designed to link together archival content and bibliographic data. This project is financed by the ANR (2017-2020) and coordinated by Michel Senellart, professor of philosophy at the ENS Lyon. It benefits from the partnerships of the ENS/PSL and the BnF. In addition, a collaboration with the European READ/Transkribus project has been started so as to produce automatic transcription of the reading notes.

Keywords

Michel Foucault; archives; digitization; Handwritten Text Recognition; Transkribus; RDF.



1 The ANR project Foucault's Reading Notes

1.1 Overview

The Foucault Fiches de lecture (FFL) project aims to explore and make available online a large set of Michel Foucault's reading notes (quotes and references organized and commented) kept at the BnF (Bibliothèque nationale de France) since 2013.

The objective is to digitize, describe and enrich Foucault's reading notes relating to the preparation of his books, lessons, lectures, but also to allow a new approach to his work, based on the analysis of the philosopher's reading practices and the thought processes they help to trace. To put the reading notes online and collectively produce the data, the team develops a collaborative platform, based on RDF technologies[1], to bring together archival content and bibliographic data on the sources consulted by Foucault.

This project is financed by The French National Research Agency (ANR 2017-2020) and coordinated by Michel Senellart, professor of philosophy at the ENS Lyon. It benefits from the partnerships of the ENS/PSL and the BnF.

1.2 The Foucault Archives at the Bibliothèque nationale de France

What are we talking about when we refer to the Foucault archives? Since 1994, the department of manuscripts at the National Library of France (BnF) holds the original manuscripts of *The Archeology of Knowledge* (1969) and *The History of Sexuality* II and III (*The Use of Pleasure and The Care of The Self*, published in 1984). In 2013, the library added more than 37,000 pages of manuscripts and notes on readings, lectures, conferences, books that he never wrote, and the first drafts of books he did end up publishing, acquired from Daniel Defert. There is also an intellectual journal in which Foucault wrote his ideas, reflections, and quotes throughout his work in the 1960s and 1970s (mostly). These remarkable documents make up a heterogeneous archive, tracing more than 30 years of intellectual activity, from his education, until his death in 1984. The documents vary widely, from simple reading notes to detailed texts prepared for lectures or publication (such as *Les aveux de la chair*). Most recently, in 2015, a third series of Foucauldian archives was donated to BnF, including a certain number of documents from Foucault's youth, i.e. his masters thesis, his notes from his time as a student at the École normale supérieure, and a series of letters, some of which are family correspondence that the Foucault family doesn't want released for consultation until 2050. Thus, the Foucault archives

[1] An overview of the data model and mashup functionalities can be found on the BnF blog: http://bnf.hypotheses.org/files/2018/07/Slides_atelier_bnf_Corpus_2.pdf (accessed on October 22, 2018).

make up an extremely heterogeneous and varied set of documents, both in their content, their form, their presumed date of composition and in the way they were acquired by the BnF.

Due to the complex nature of the Foucault archives, the National Library (with the agreement and scientific support of Daniel Defert, the rights holders and the Association for the Michel Foucault Center) has launched the first systematic exploration of the documents acquired in 2013 — undeniably the largest part of the Foucault archives. The aim is to write a new inventory, more detailed than the one originally provided when the BnF acquired these documents, and make it available in the online catalog of the manuscripts[2].

1.3 General methodological remarks: which philosophical use for these archives?

The original folders distribution in the boxes was made not by Foucault himself but by Daniel Defert, who kept the order in which these texts were found in Foucault's office at the time of his death in 1984. Even though Daniel Defert himself is aware of possible misclassifications, and even though different parts of the same text were found in separate boxes, the decision was taken to keep the original order of the documents. Because Foucault had the habit of making different "montages" of his own texts, the "discontinuous" order in which several documents are found may be indicative of his approach and work. Moreover, every other methodological criteria would be a more violent intervention on the archives than preserving the existing order.

So, it is essential to weave intertextual references between the different boxes. These references highlight overarching themes in Foucault's work, the use of a same archive source on different occasions, and how he re-exploited the same document from different perspectives. The hundreds of boxes acquired by the BnF show Foucault reusing and repeating incessantly his own notes and texts. Very often, an earlier work is reused and rearticulated for a later project. Corrections with new styles of writing show different phases of his analysis. The succession of authors studied and archives explored tell us something about his own intellectual journey, with plenty of bifurcations, deviations, and repetitions along the way. Could we question these archives from the point of view of the "archive", in the Foucaultian meaning of the concept – "a practice that causes a multiplicity of statements to emerge as so many regular events, as so many things to be dealt with and manipulated"[3]? It is tricky to apply Foucault's methodology to his own archives, whose composite nature makes it difficult to consider them as a general and well defined discursive system. We had better be methodologically cautious. But it is true that these archives can be used as a mass of documents reflecting the practice of historical discontinuity that Foucault tried to describe in his works. Their heterogeneous density constitutes a concrete attempt to draw some general lines of the "diagnosis of the present" that "establishes that we are difference, that our reason is the difference of discourses, our history the difference of times, our selves the difference of masks". Foucault's archives are an example of the practice of that "difference" which, "far from being the forgotten and recovered origin, is this dispersion that we are and make"[4]. The process of intertextuality is therefore very important to make this "difference" appear through the archives, considering every manuscript, document and text cited as a node in a complex network of references.

[2] http://archivesetmanuscrits.bnf.fr (accessed on October 22, 2018).

[3] M. Foucault, *The Archaeology of Knowledge* [1969], transl. by A. M. Sheridan Smith, New York, Pantheon Books, 1972, p. 130.

[4] M. Foucault, *The Archaeology of Knowledge*, op. cit., p. 131.

We can thus formulate our first observations on these archives. At least three questions may be asked about the disposition of the documents in the boxes. As explained before, it is hard to tell if the original *order* of the documents is random or significant and meaningful for our understanding of Foucault's method and the development of his thought. The question might seem idle, but it is actually theoretically dense. Do these archives have a form and order that are significant in themselves? Second, one can question the *style* of Foucault's thought in formation, as it is captured in these archives: do the bifurcations, deviations, the tortuous paths and repetitions, taken as a whole, shape a specific form of philosophical work? Do the archives overlap with Foucault's explicit methodological statements in his published writing? Or do they represent a new way to think about speeches and texts, more concrete, more "surgical" — philosophy as a "scalpel" cutting through the body of words, according to a famous definition of Foucault in an interview with Yves Bonnefoy[5]? Lastly, we cannot neglect the impact of these archives on Foucault studies today: the point is to understand how these archives transform the readings and uses of Foucault, especially how they enable to appropriate his work in a certain way, to work on it while going beyond the mere commentary of Foucauldian texts which is sometimes reduced to a blind and sterile devotion. The point is to understand how these archives can reactivate the political and critical potentialities of his analyses.

The Foucault archives appear as a snapshot of his thought as he left it when he died. It is such necessary to browse it and keep it as a living trace of work in progress. If the archives are an incredible jumble of different sources and styles, the goal is not to put it in order, but rather to find the traces of "creative disorder" that the archives represent. We take as hypothesis that this creative disorder is not only an integral part, but represents one of the most significant moments in Foucault's style of thinking, who was so attentive to mixed materiality, multiple strategies, and heterogeneous temporalities that superimpose themselves without ever collapsing into a single reality.

His reading notes represent an important part of the archives. Daniel Defert's inventory estimates 20,300 handwritten notes, single or double-sided. 41 out of 117 boxes could be filled entirely by reading notes. They are never simple notebooks of references and citations. Foucault immediately treats his sources critically and philosophically: he organizes them by theme and key notions and then in the construction of a single concept mixes multiple sources, primary and secondary. Notes about the early church fathers, for example, simply called "Penitence Clothes[6]", which refers to an article in the second volume of a 1954 German Lexicon (*Reallexicon für Antike und Christentum*), simultaneously uses texts of Plutarch, references to Babylonian and Phrygian rituals, and quotes from Saint Hilary, Saint Jerome, and Saint Cyprian.

Even if we consider only Foucault's reading notes among his manuscripts and documents, many theoretic and hermeneutic problems remain therefore open. Michel Foucault's archives are a perfect example of philosophical "working archives", i.e. the files used by a philosopher to build his thought, collected in a specific and institutionalized space of memory (the library). But they are also a sort of "archive", in the Foucauldian sense of the notion: not only all the things said in a certain time and space, but the set of laws and relations composing them in ordered figures, growing linearity and regularity or blurring discontinuity through time.

This "archeological" dimension of Foucault's reading notes is essential to historical and philosophical analyses, but at the same time difficult to reconstruct due to the huge amount of doc-

[5] M. Foucault, *Le beau danger. Entretien avec Claude Bonnefoy*, Ph. Artières éd., Paris, Éditions de l'EHESS, 2011.

[6] Box #24, Archives Foucault, Bibliothèque nationale de France, Département des Manuscrits, NAF 28730.

uments provided. Digital tools become therefore a powerful ally to any research project on Foucault. It is impossible, without fully transcribing each reading note, to trace all the sources, ancient and modern, used by Foucault in relation to a specific research theme. Since Foucault's reading notes can not be fully published (it would make no sense), the most effective way to make the richness and precise developments of Foucault's notes available to researchers would be to digitally transcribe them, include intertextual references to the sources and developing new tools to browse between them and show their common stylistic, conceptual or bibliographical traits. The BnF is currently working on such a project as a partner of the ANR project: "Foucault's Reading Notes".

1.4 How to explore the corpus with digital tools?

A first step in Foucault's reading notes exploration was taken with the ANR Corpus project *La bibliothèque foucaldienne*[7], which led to publish digitized manuscripts online in 2011. The corpus consisted of 822 reading notes taken from working papers that Foucault gathered during the writing of *The Order of things* (1966). The digitization allowed, first, to share the manuscripts more easily, and then to index them in a digital inventory (XML-EAD encoding). For this purpose, a specific data model was added to EAD, so that researchers could describe precisely how Foucault collected, commented and reused bibliographic information[8]. Using these digital tools, researchers have studied the way Foucault combined primary and secondary sources to elaborate his concepts, his various quoting styles, or even the difference between the conception of books and lectures.

Thus, Foucault specialists discovered new insights into the philosopher's "archaeological" method, and a light could be shed on his original interpretation of structuralism. This both revealed how Foucault explored large periods of western culture, and how he reorganized his sources to unearth "epistemes"[9]. Besides, it became possible to analyze more extensively the relationships between the documents Foucault consulted, showing the importance of intertextuality in his original way to combine historical and philosophical approaches. More recently, some visualization tools have been added to the inventory, in order to give a transversal perception of the distribution of the sources in the notes, emancipated from the hierarchical structure of EAD[10].

Based on these first achievements, the ANR project Foucault's Reading Notes[11] (FFL) aims at extending the corpus to a new scale and also providing new digital tools, designed to enable collaborative work and data reuse. Since the BnF fonds has now more than 20 000 reading notes, the FFL project planned to digitize more than 10 000 new folios, amongst which 5 720 have already been released on the FFL prototype platform. As said above, it wouldn't be relevant to

[7] http://lbf-ehess.ens-lyon.fr (accessed on October 22, 2018).

[8] Samantha Saidi, Jean-François Bert, Philippe Artières. « Archives d'un lecteur philosophe. Le traitement numérique des notes de lecture de Michel Foucault », in *Codicology and Palaeography in the Digital Age, Schriften des Instituts für Dokumentologie und Editorik*, 2011.

[9] « Le fichier foucaldien, en tant que recueil de citations, procède d'abord d'une opération d'extraction, voire de fragmentation qui dénoue les parentés premières (celles liées à l'auteur ou à l'œuvre en particulier) pour réarticuler les archives du savoir (la masse des choses dites et pensées) à l'archive d'une époque[...]. » Philippe Sabot, notice des *Mots et les choses*, 2015 (in Foucault, *Œuvres*, Pléiade t. 1 p. 1595).

[10] 2015. http://lbf-ehess.ens-lyon.fr/pages/visualiser_les_donnees.html (accessed on October 22, 2018).

[11] Foucault Fiches de Lecture. See http://ffl.hypotheses.org (accessed on October 22, 2018).

try to render, nor restore, a hypothetical original order in the archival material. Moreover, being organized as a "file of files", where every folio may have been moved from a folder to another and reused in different contexts, as opposed to a bound document, the reading notes form what Bachimont called a "hyperdocument"[12]. Actually, the corpus is not a collection of continuous texts, but a series of fragments which tend to compose an anthology of commented references and quotations, that we could compare to a bibliographical database.

In order to take this specific documentary form into account and to help to understand the "creative disorder" of the papers, the platform has been designed to provide heterogeneous annotation categories, so that researchers can create multiple pathways between the documents. A specific data model, based on RDF, was implemented in the platform, so as to substitute an aggregation of metadata to the classical archival records.

Such records as XML-EAD actually represent all documents in a hierarchical classification plan, and assume that the same grid can be applied to each item; furthermore, it's very difficult to express uncertainty and interpretation within the fields: every piece of information is equivalent to another as regards reliability or trustfulness. On the contrary, the main principle of the FFL data model is that every piece of information is independent from a global and unique schema, so as to constitute an autonomous assertion with its own metadata. Hence we make no formal distinction between archival description and personal comment, and both are sibling classes of the generic `Annotation` class. All annotations share common properties, which enables to compile personal interpretation and descriptive elements easily *and* to distinguish them in the same time, for every annotation has its own unique identifier, an explicit type, and provenance information – as soon as a user creates it[13].

For instance, the EAD schema has a `<bibref/>` element, in which the users can mention a document and specify its title, publisher, etc. However, it's impossible to insert reliability or provenance information about this particular reference. Here is a representation of a bibliographic citation in the FFL model:

```
PREFIX f: <http://ffl.org>
f:readingNote1 f:hasAnnotation f:annot1 .
f:annot1 a f:BiblioCitation ;
  f:access "public" ;
  f:value f:Duby_Economie_rurale_II ;
  f:createdByUser f:user3 ;
  f:creationDate "2017-06-25_11:03:40" .
f:BiblioCitation rdfs:subClassOf f:Annotation ;
  f:requiresValueType f:Document .
f:Duby_Economie_rurale_II owl:sameAs <http://data.bnf.fr/ark:/12148/
   cb345892334#about>
```

As every annotation (description or personal comment) has its own URI, we can add as many

[12] « Un hyperdocument est moins qu'un document mais plus qu'un agrégat de documents. (...) Le propre des hyperdocuments est donc de rompre la linéarité du signifiant textuel pour suggérer des parcours non linéaires. Ces parcours peuvent être complémentaires et s'ajouter à un parcours linéaire canonique, comme la note de bas de page enrichit un texte principal, ou bien constituer la textualité elle-même, le lecteur devant affronter la multiplicité des parcours possibles pour construire sa propre lecture ». Bruno Bachimont. *L'ingénierie des connaissances et des contenus*, Hermes science publications, Lavoisier, 2007. Pp. 181-84. See also J.-F. Bert: *Une histoire de la fiche érudite*, Presses de l'enssib, 2016. Online: `http://ficheserudites.enssib.fr` (accessed on October 22, 2018).

[13] Cf. note 1.

properties as necessary: for instance the user can decide if the annotation will be public, restricted or private. Similarly, it is very easy to add a degree of certainty:

```
f:annot1 f:certaintyDegree "0.6" .
```

This mechanism is based on the same principle as RDF reification[14], since an assertion becomes the matter of another assertion. However we found it more convenient for our objectives to create a custom `Annotation` class, because it's easy to add more properties later (or modify existing properties).

Actually, we found it more straightforward to create two children classes (`ApproximateBiblioCitation` and `ExactBiblioCitation`) than to ask the users to provide a confidence rate for each citation. Here are the corresponding triples:

```
f:ApproximateBiblioCitation rdfs:subClassOf f:BiblioCitation ;
  rdfs:label "Use this type of citation when you're not sure if the
     chosen edition is the good one"
f:ExactBiblioCitation rdfs:subClassOf f:BiblioCitation ;
  rdfs:label "Use this type of citation when there's enough evidence
     that Foucault consulted the chosen edition."
```

Lastly, the provenance information (`createdByUser` and `creationDate`) allows to gather all the data that have been created by a user in a *personal user profile page*, and give the complete history of their contributions. Thus one can insert various types of bookmarks when exploring the corpus, and browse the documents according to their own research objectives.

We can therefore say that there is not one corpus, but multiple viewpoints on the manuscripts and the connections between the reading notes contents. In conjunction with this information architecture, the project aims at developing new ways of searching and browsing the manuscripts and tools that facilitate the representation and analysis of relationships between Foucault's sources. Because a tree structure could not render all the connections between the reading notes, and also as a tribute to Foucault's conception of intertextuality[15], the FFL project offers a network cartography system that shows the distribution of sources and concepts between the manuscripts, and gives access to them via hyperlinks. The first visualization attempts, for La bibliothèque foucaldienne, were realized with Gephi, and only gave a statical overview of the corpus: the new tool integrated in the web platform is based on a javascript library[16] (cf. figure below).

In addition to that, since a large part of the investigation consists in identifying Foucault's sources, the platform enables to create named entities and enrich them with biographical and bibliographical information, that the users can search and retrieve from BnF RDF data[17]. Based on the analysis of the first corpus, which contained about 590 references disseminated in the

[14] Cf. Bizer, C.; Carroll, J.; Hayers, P. and Stickler, P. (2004): *Named graphs, provenance and trust*. Online: http://wifo5-03.informatik.uni-mannheim.de/bizer/SWTSGuide/carroll-ISWC2004.pdf (accessed on October 22, 2018).

[15] « C'est que les marges d'un livre ne sont jamais nettes ni rigoureusement tranchées: par-delà le titre, les premières lignes et le point final, par-delà sa configuration interne et la forme qui l'autonomise, il est pris dans un système de renvois à d'autres livres, d'autres textes, d'autres phrases: nœud dans un réseau ». Michel Foucault, *l'Archéologie du savoir*, 1969 (p. 34).

[16] Vis.js: http://visjs.org (accessed on October 22, 2018).

[17] This way of referring to named entities allows to reuse data easily, therefore to shorten the work of collecting the references, but also to normalize data.

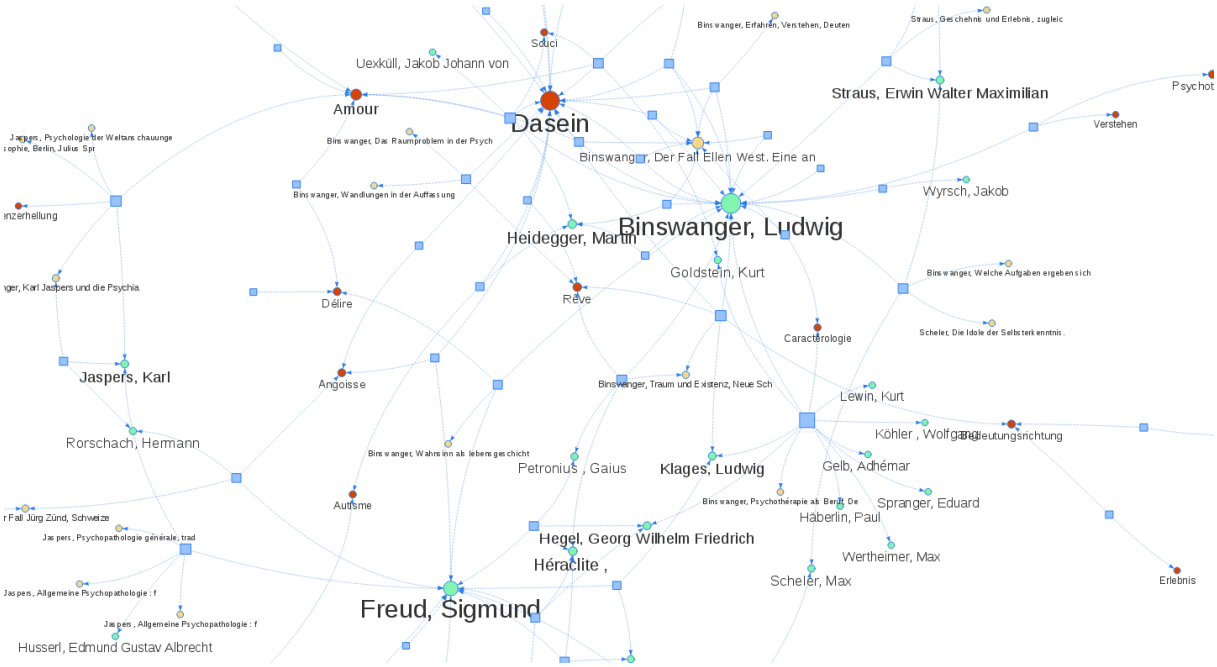


Figure 1: **A network cartography of reading notes in box #38** (annotated and described in the platform by E. Basso).

822 reading notes, one can assume that the new corpus may gather thousands of references and quotations. Hence it will be a huge amount of work to identify all the sources, and the project considers two complementary methods. The first one relies on collaborative work, and is already being used. The second one consists in trying to transcribe the manuscripts automatically, in order to index the corpus and process the transcripts with named entities recognition tools (NER).

Of course it isn't possible to expect perfect transcriptions and entities recognition, but the team has tested a remarkable handwritten text recognition (HTR) expert system, Transkribus, which could change our way to explore the corpus. As this system, still experimental, gave very promising results, a detailed reporting of the tests is given in next section.

2 A tool for automatically transcribing manuscripts

N.B.: the tests described below were made in 2017 (Transkribus versions 1.2.3 to 1.3.1).

An international collaboration with the European READ/Transkribus[18] project has been started for the automatic transcription of Foucault's manuscripts from his reading notes. Transkribus is an automatic handwriting recognition software, accompanied by a platform for transcribing digitized images of manuscripts and a classic OCR. The Main uses of Transkribus are:

- Train the handwriting recognition engine (HTR: Handwritten Text Recognition[19]) then use it to automatically transcribe the images provided;

[18] https://read.transkribus.eu/transkribus (accessed on October 22, 2018).
[19] Automatic handwritten text recognition.

- Transcribe documents for scientific editing (including: WYSIWYG input interface with TEI encoding and creation of custom tags[20]);
- Search handwritten documents for terms, including related terms ("fuzzy" search, "keyword spotting"[21]).

Transkribus is an expert system, based on machine learning technology, i.e. a system capable of "learning" to decipher the writing of a given scripter[22]. This learning is based on a training dataset and results in the production of a neural network (also named HTR model) which is specific to the scribe.

It is therefore necessary to provide digitized images accompanied by their transcription (digital text tagged with markup), while ensuring line-by-line correspondence between the image and the text. Once the HTR model has been created by the Transkribus team, it becomes possible to automatically transcribe a batch of images using the HTR engine. In addition, the model thus produced can be improved later, by adding new images and transcriptions to the initial data of the dataset training. An iterative process can therefore be carried out: at the end of the first training, the HTR engine is used to produce a first batch of automatic transcriptions, which are manually corrected, and which make it possible to carry out a second training, and so on. However, it should be noted that the learning process requires at least 200 images to function effectively: the use of Transkribus is therefore relevant when a large quantity of images is available to transcribe[23].

The services offered by the platform are free. Simply register on the website, contact the READ project team, and follow the instructions in the introductory document "How to use Transkribus in 10 steps"[24] to install the software and work with your own documents after uploading them to the Transkribus server. Learning to use the software takes some time, but Transkribus tutorials are clear and the team is actively involved in this learning[25].

2.1 Creating learning data for Foucault's handwriting

Initially, a test was carried out by Vincent Ventresque with 200 images (about 30 000 words) taken from the manuscript of *Théories et institutions pénales*: the team already had a transcription realized by Elisabetta Basso, which made it possible to accelerate considerably the production of the first dataset training.

It remained to align this transcription with the images, and to make it conform as closely as possible to the original: Transkribus does not use a dictionary and does not try to recognize

[20] Various formats are available for export, in addition to TEI. Transkribus uses PAGE XML: http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15.

[21] See documentation: https://transkribus.eu/wiki/images/1/1b/HowToTranscribe_Keyword_Search.pdf (accessed on October 22, 2018).

[22] It can also work for several scribes, if their number is not very high for the same corpus.

[23] Automatic transcription saves time even if it has to be corrected, and provides an initial digital text in which it is immediately possible to search for terms. However, the time spent producing learning data represents a significant investment. In the case of Foucault's reading notes and in the current state of the tool, we estimate that this investment becomes justified beyond 500 images to transcribe, among which 200 images can be used for a first test.

[24] https://transkribus.eu/wiki/images/7/77/How_to_use_TRANSKRIBUS_-_10_steps.pdf (accessed on October 22, 2018).

[25] In addition, some institutions are starting to offer training.

words, but analyzes the lines of text *character by character*[26]. To ensure optimal learning, it is therefore recommended to scrupulously respect the letter of the manuscript: it was necessary to restore abbreviations when they had been developed in the initial transcription, but also spelling mistakes and missing accents, and add markup to identify unclear or illegible passages[27].

This alignment between the images and the text requires a segmentation phase, i.e. the identification of text areas (blocks and lines) in the image. At the time of our test, we had to manually draw the large text regions on each image[28], start the automatic line recognition process, manually correct wrongly recognized lines, and finally, add the digital text on each line[29].

At the end of this first phase which consisted in producing training data, and lasted approximately two weeks, the READ project team produced (in a few hours) a first HTR model of Foucault's handwriting: the average recognition rate was 85% per character. This result was considered very encouraging by the READ team, and we decided to continue with manuscripts from the FFL corpus, for which we had no pre-existing transcription.

Indeed, G. Mühlberger, head of the READ team, assured us that with about 400 additional images, the average recognition rate could go above 90% per character, and that we could automatically produce an automatic transcription over our entire corpus[30].

The second phase therefore consisted in transcribing and aligning more than 400 images (about 50 000 words), chosen in the corpus whose transcription had been planned in the project.

This transcription work was carried out by Marie-Laure Massot and Arianna Sforzini and concerned selected files in the following boxes:

- box #1 (preparation for the 1971-1975 lectures and *Surveiller et punir*),
- box #51 (preparation for *La volonté de savoir*)[31].

The aim was to produce an improved neural network, and thus to obtain more accurate automatic transcriptions: this would both save time for future transcriptions and produce a digital text that could already be used for "full text" search, even before automatic transcriptions were proofread and corrected[32].

[26] It was possible to add a dictionary in Transkribus, but when the test was carried out, Transkribus team explained it wouldn't make a significant difference. However, this might not be the case with the latest versions of Transkribus.

[27] With TEI encoding and formatting use, see next section.

[28] Since then, line recognition has been greatly improved, and it is no longer necessary to manually cut text areas, which saves considerable time - still correcting poorly recognized lines.

[29] See screenshots of figures #2 and 3. Transkribus requires to make a direct link between the images in the document and the corresponding transcribed text. To do this, it was previously necessary (it is no longer the case) to manually define text regions on several images and then automatically detect the lines using the button "Find lines in text regions" in the "Tools" tab. Next, align the baselines of the image with the lines of the transcription window (for each baseline of the image, there is a corresponding line in the text editor). The text must therefore be transcribed line by line, exactly as it appears in the image.

[30] More detail is given in Gunter Mühlberger and Tamara Terbul's article (complete reference in 2.3).

[31] A total of 431 additional images: box #1: 1-300; 300-324; 335-352; 369-380; box #51: 33-62, 81-126.

[32] The FFL project initially planned to produce transcripts for a limited portion of the reading notes corpus. Thanks to Transkribus, it becomes possible to transcribe more files (saving time for reading and typing) and to immediately have a corpus transcribed imperfectly, certainly, but automatically.

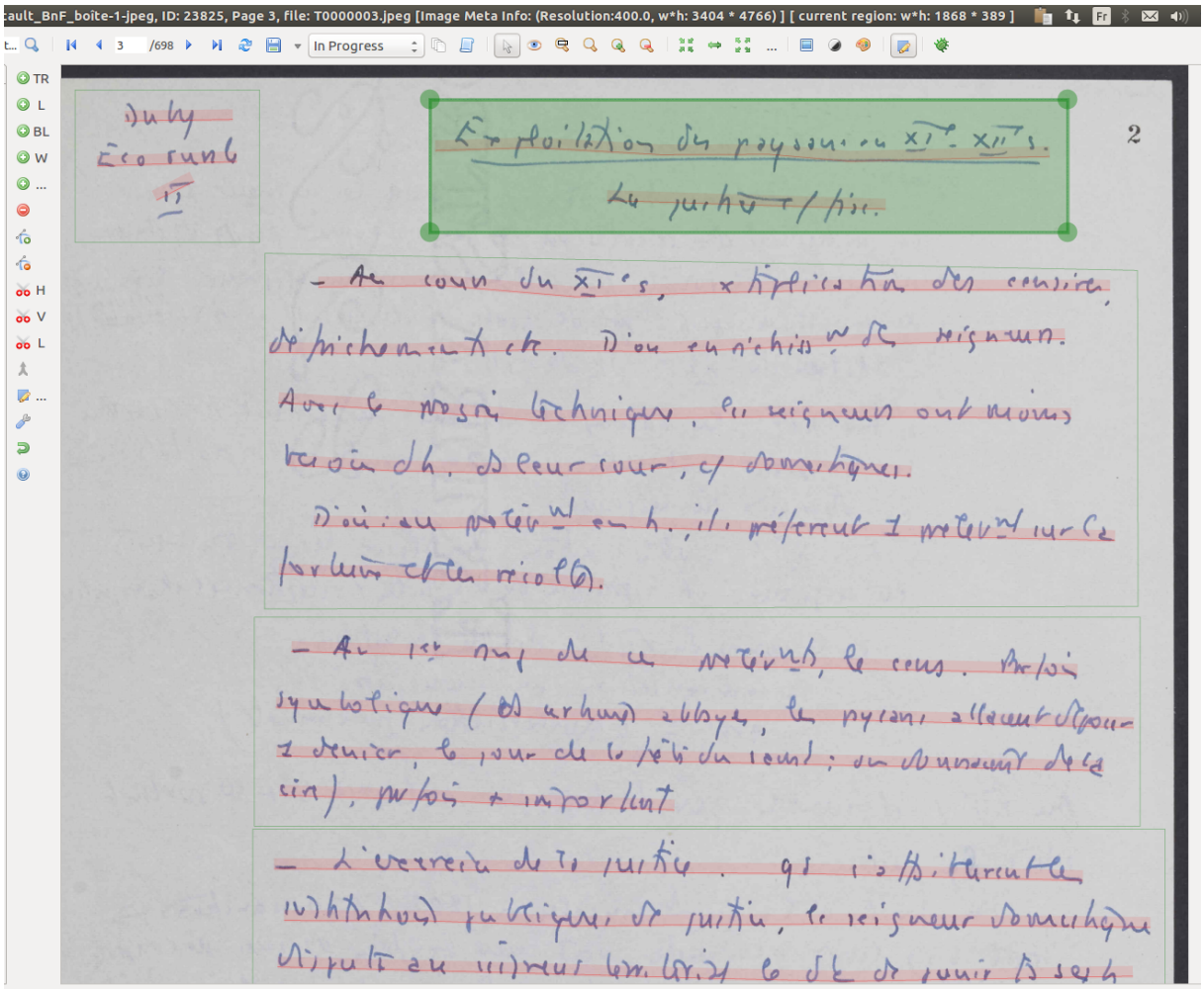


Figure 2: **Document segmentation:** text regions (textregions: green rectangular frames), lines and baselines (baselines: red).

2.2 Main difficulties encountered

– **Foucault's handwriting**

These are reading notes, with many abbreviations. Foucault's writing and abbreviations are often difficult to decipher. Several abbreviations are equivocal, which can mean different words in different contexts. Foucault also wrote several letters in the same way, once again making deciphering his writing extremely difficult without taking into account the general context (this is particularly true for names of authors and titles of works in the reading notes). For the Transkribus test, we have identified illegible words, with TEI encoding (element `<gap/>`), and a strikethrough font (`<strikethrough/>`) for doubtful, misshapen or erased words.

It would then be interesting to submit certain images to the Foucault community of specialists, having identified words that are illegible or difficult to decipher. Similarly, a dictionary of difficult words and abbreviations could be considered, with screenshots of words and abbreviations, which would facilitate future transcription, and more generally, access to Foucault's manuscripts for researchers. In addition, the project seeks to inventory the works and authors cited, as well as the historical figures, by aligning these "entities" with the lists of authorities

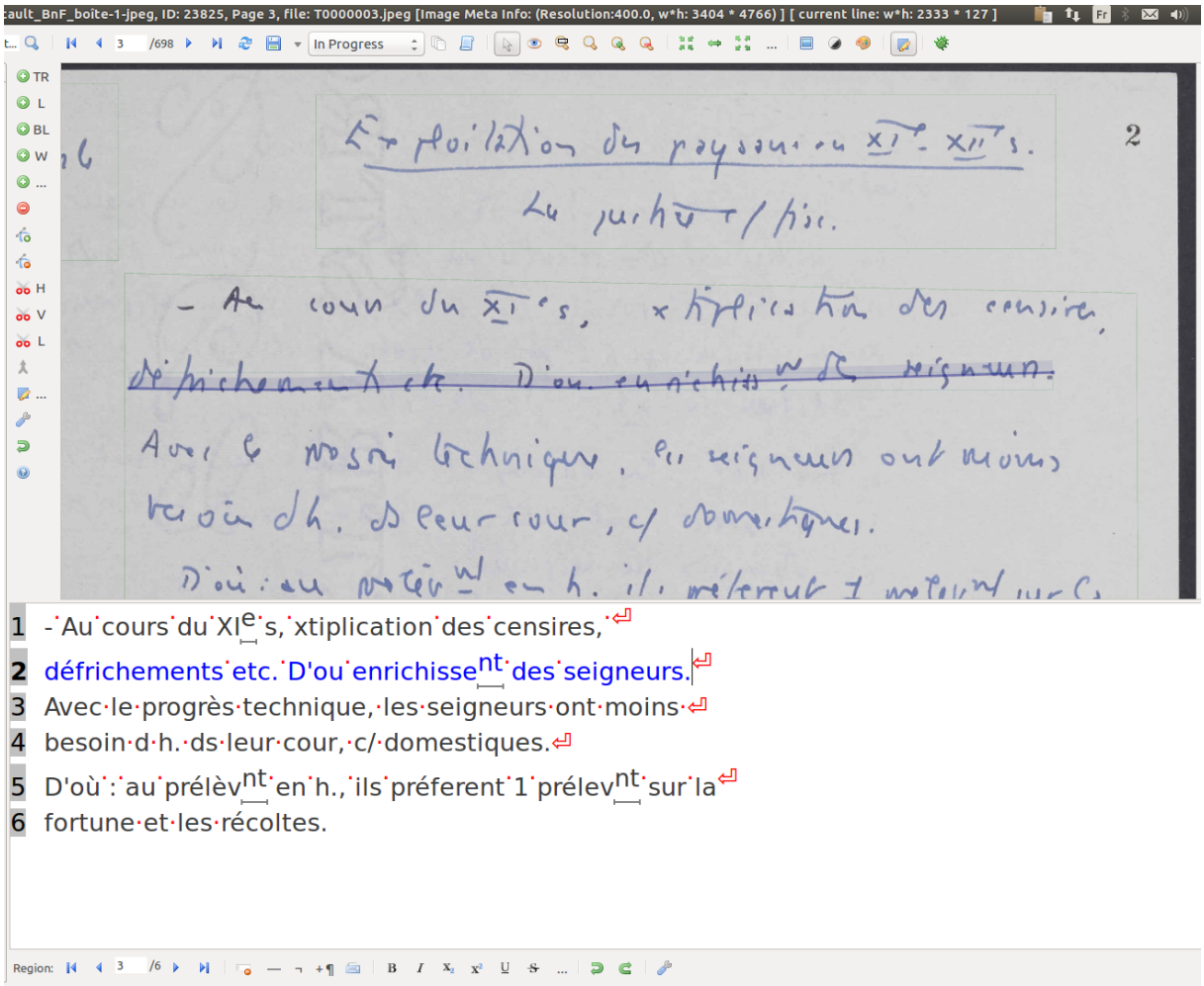


Figure 3: **Line by line transcription of the manuscript** (1st § of the folio; the line highlighted in blue in the image corresponds to the second line of the text editor, also in blue).

and bibliographic records of the BnF and other institutions[33]. This study will allow researchers to know the sources used by Foucault, but also to decipher the reading notes more easily.

In all cases, a review of the transcriptions by one or more Foucault's specialists should be considered in the short or medium term, on the Transkribus platform or a different one, a more easily accessible one.

– **Getting to grips with the software and average time per folio**

It takes a few hours to learn how to use Transkribus efficiently in a project: it is fast enough for the basic transcription functions (segmentation and typing in the editor window), but some reflexes are gradually acquired and save time on typing.

Thus, after becoming familiar with the writing of Foucault and having transcribed about thirty images, the average time of transcription (segmentation, capture) then of re-reading of a folio (2 images) was from 30 to 40 min according to the length of the note (1 page and a half or 2 full pages) and the presence of proper names that are difficult to read, and for which it was necessary to carry out a documentary research. Note however that these difficulties are not specific to the use of Transkribus, and are common to any transcription work.

[33] As explained in previous section.

– **Software Ergonomics**

It should be noted that Transkribus is still under development, and that the READ team is focusing on improving text and character recognition features. Therefore, the interface remains to be improved to facilitate the reading of images and the typing of digital text.

For example, on a laptop computer with a 13-inch screen, it is not always easy to read the manuscript and enter the transcript at the same time. The input space is very limited and the line being typed is highlighted in blue on the image, which can affect legibility. It might be more practical to be able to choose the layout of the windows ("floating" windows), to compare the image and the digital text rather than one above the other.

Another important limitation: it is not easy to navigate from one record to another to check a word, nor to compare several images, or even to have a global view of the record to search for a particular word. Loading the next or previous image makes the image disappear during consultation, and the loading time of the image and data (text areas, transcription) can be quite long, especially if you do not have a fast Internet connection. The comparison of several images is particularly problematic: there is a thumbnail gallery function, but here too the loading is very long and can sometimes block the software. Currently, it is not possible to choose the number of images to load - all the thumbnails of a collection are loaded at once - and we have uploaded our images in batches corresponding to the archive boxes (between 700 and 1140 images per box[34]).

– **Multi-handed transcripts**

In the event of multiple people contributing, some problems arise in standardizing transcripts, as interpretations can be different.

For example, Foucault often writes almost continuously, without visible space between words. Some transcriptors will tend to separate the words, others to remain more faithful to the image and to attach them. Or he shortens some words (e.g. "x" for "not"), but not systematically. Some transcribers will tend to develop the abbreviation, others will not. Lastly, accents and apostrophes are rarely marked distinctly by Foucault, they are often tied to previous or following letters, we guess them more than we can distinguish them. There again Foucault is not systematic.

In all these cases, the transcriber's interpretation plays an important role in the capture of this type of document. However, it seems impossible to envisage the capture of such an extensive and complex corpus if not collaboratively, and one of the objectives of the FFL project is precisely to provide to the community of specialists of Foucault a collaborative transcription and annotation space. It will therefore be essential to provide a collaborative tool that is easier to access than Transkribus[35], and to establish a transcription guide defining the principles and rules common to all contributors.

[34] For future training phases, it might be wise to create "sub-collections", limiting the number of images per collection. On the other hand, it would involve additional manipulation to circulate in the corpus. Another solution could be to work with a local folder of images, if synchronization with the server is possible.

[35] It would be unrealistic, in particular, to ask each potential contributor to install the software, to learn how to use it, and then to juggle between the FFL project platform and the Transkribus interface. We are therefore considering the use of the Eman platform (see next paragraph and note 38).

2.3 Assessment of the experiment and prospects

After this "training" on approximately 600 images transcribed manually (transcriptions made by Marie-Laure Massot and Arianna Sforzini, and reuse of a transcription of Elisabetta Basso for the manuscript of *Théories et institutions pénales*), the results of the Transkribus software tests are very encouraging: we obtained an average success rate of 92% on the characters[36]. In addition, future corrected transcripts can be fed back into Transkribus to complete the training and increase the accuracy rate.

Obviously the automatic transcription of the notes must be taken up manually by editors. We also notice a certain difference in the degree of automatic character recognition according to the different boxes used: in particular, the software is sensitive to the paper transparency of the scanned folios and sometimes also reads the characters on the back. Further software developments could perhaps correct this difficulty[37]. Nevertheless, the results obtained remain very positive, since the automatic transcription allows a faster manual transcription, and even helps with the deciphering of certain words that are difficult to read.

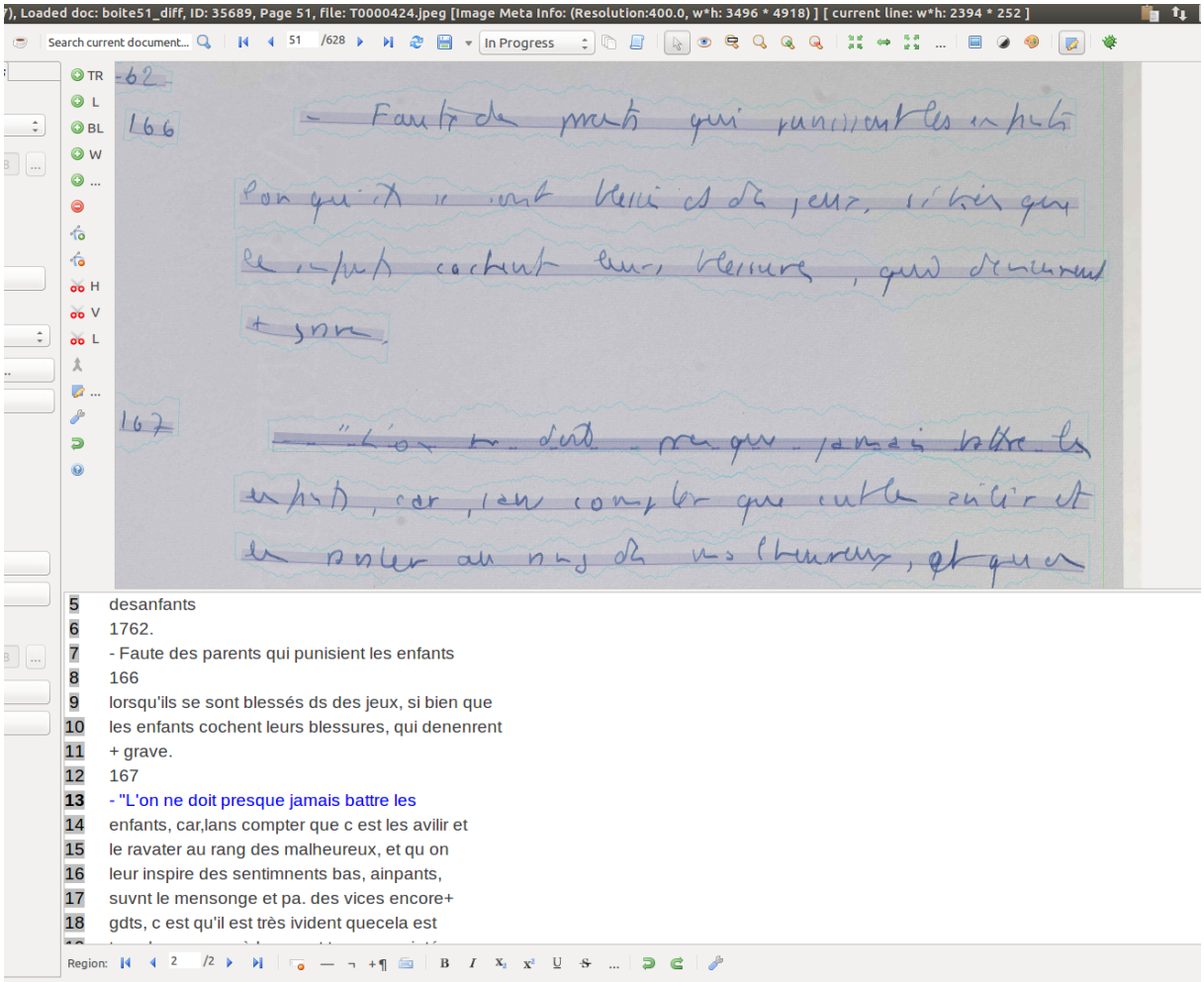


Figure 4: **Automatic transcription**: box 51, note on Venel, *Essai sur la santé et l'éducation médicinale des filles* (1776).

[36] See figures #4 and 5.

[37] For example, by taking into account the contrast differences between the front and back characters that appear in transparency.

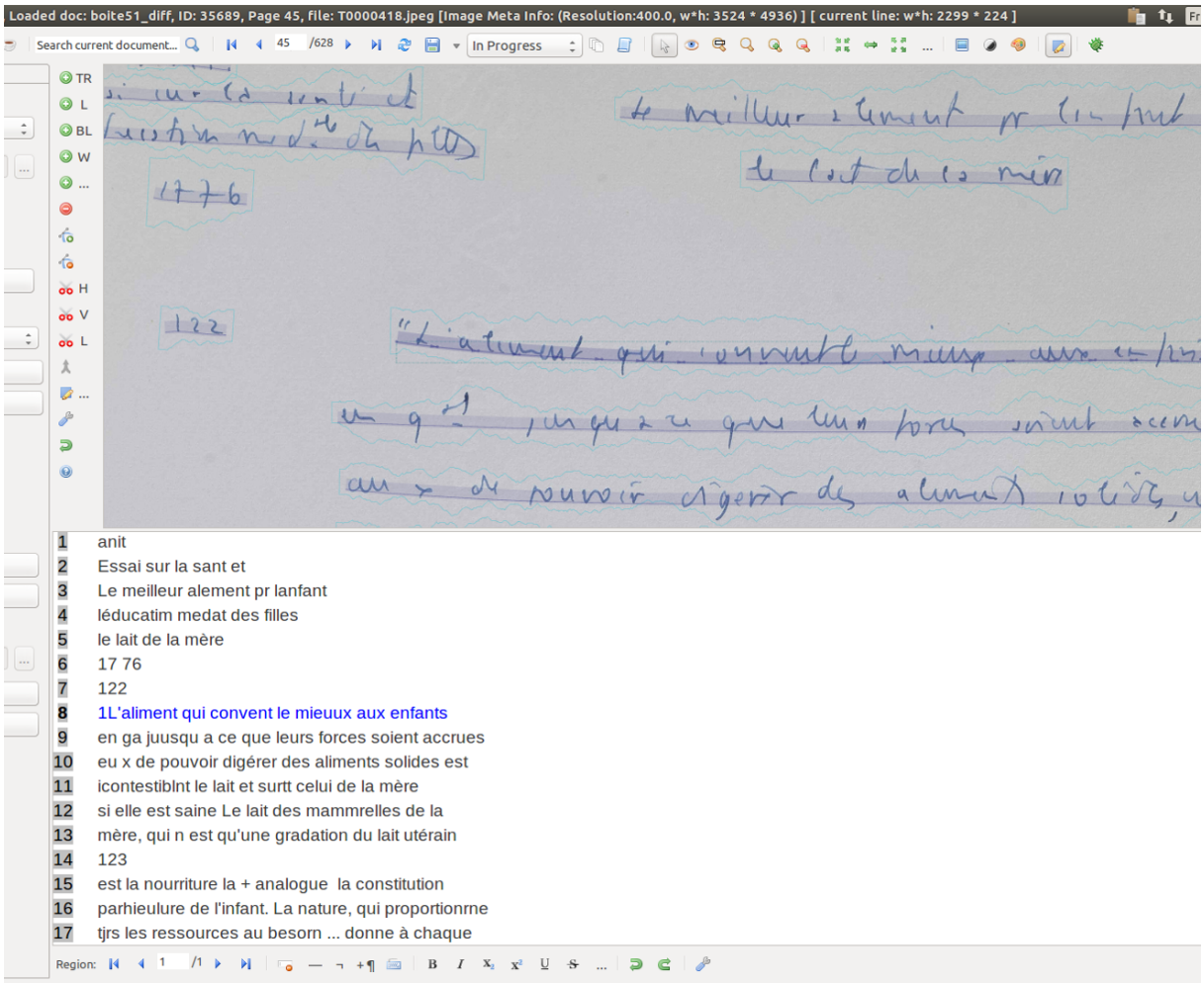


Figure 5: **Automatic transcription**: box #51, note on Ballexserd, *Dissertation sur l'éducation physique des enfants* (1762).

We are therefore considering producing automatic transcripts for all images already digitized or to be digitized. It will obviously not be possible to correct all of these transcriptions within the context of the ANR project itself, that is why we are also considering the use of Omeka/eman[38], a platform to set up a collaborative correction system for these automatic transcriptions. After a while, we will be able to create accounts for scholars interested in this work. It is actually more interesting to create a space for sharing around Foucault's manuscripts, opening up the possibility of gradually completing the transcriptions and annotations of the corpus, than to propose a series of "read-only" transcriptions in a limited number.

Finally, despite their imperfection, automatic transcriptions are already usable to perform "full text" searches. One of the most relevant features of Transkribus for researchers is undoubtedly the "full text" search in images or texts transcribed by Foucault. It is possible, even in the presence of an incorrect transcription, to search for keywords ("fuzzy" search). At first during the automatic transcription process (HTR), the software does not correctly recognize the handwritten word, but it is able to recognize the similarity between this word and the term the user wants to retrieve. Obviously, this could be decisive, on a corpus of more than 10 000 reading sheets,

[38] Developed by Richard Walter and ITEM: http://eman-archives.org (accessed on October 22, 2018). Eman is based on the Omeka Content Management System (CMS): https://omeka.org (accessed on October 22, 2018).

for cross-cutting search by concept, author, work, and imperfectly recognized terms as well.

However, this search tool is only available to Transkribus users who have downloaded and installed the software. In the medium term, access to Transkribus will also become chargeable. We therefore plan to implement fuzzy search[39] features within the Eman/Omeka platform, and the RDF based annotation prototype. Besides, it should also be remembered that Transkribus does not natively use a dictionary for the automatic transcription phase[40], but analyses the manuscripts letter by letter: the results can therefore be improved by using automatic correction algorithms or by searching for similarities to "clean" the automatically generated data.

Conclusion

Collaborating with the European Read/Transkribus project has really given a new dimension to the ANR Foucault's reading notes project: By expanding the transcriptions and the explorations possibilities, the Trankribus software will enable the Foucault's specialists to embrace a larger corpus of his archives, and so to have a better understanding of the philosopher work, in a larger scale but in a more accurate level as well. The prototype, the Transkribus software and the Eman/Omeka platform needs to be linked and to communicate in order to create new data: testing new ways of enriching the corpus, transcribing, or correcting and sharing, the spaces are complementary. We are also convinced that this collaboration benefits the Read/Transkribus project by improving the neural network model and by sharing our experience of the tool. This new dialog between computer science and Humanities is the very place where new opportunities and new paths could be find to explore and enrich data, and finally to facilitate research and disseminate knowledge.

Additional Bibliography

– BONHOMME, M.-L. *Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture: Rapport exploratoire*. [Contrat] Inria, p.1-10 (2018). Online: https://hal.inria.fr/hal-01949198v2 (accessed on October 22, 2018).

– CRISTOFOLI, P. "Principes et usages des dessins de réseaux en SHS", in *Revue Histoire et Informatique*, 18-19, 2015. Online: http://blog.ahc-ch.ch/wp-content/uploads/2015/09/1-Cristofoli.pdf (accessed on October 22, 2018).

– CHRISTLEIN, V., NICOLAOU, A., SCHLAUWITZ, T., SPÄTH, S., HERBERS, K. & MAIER, A. *Handwritten Text Recognition Error Rate Reduction in Historical Documents using Naive Transcribers*. In: Burghardt, M. & Müller-Birn, C. (Hrsg.), INF-DH-2018. Bonn: Gesellschaft für Informatik e.V (2018). Online: https://dl.gi.de/handle/20.500.12116/16993 (accessed on October 22, 2018).

– GRÜNING, T., LEIFERT, G., STRAUSS, T. and LABAHN, R. (2018) *A Two-Stage Method for Text Line Detection in Historical Documents*, Arxiv (2018). Online: https://arxiv.org/abs/1802.03345 (accessed on October 22, 2018).

[39] See for example Lucene's *FuzzyQuery* feature: https://www.tutorialspoint.com/lucene/lucene_fuzzyquery.htm (accessed on October 22, 2018).

[40] It is possible to add a dictionary, but the procedure seems complex and does not seem to significantly improve the results of the HTR module.

– KAHLE, P., COLUTTO, S., KACKL, G., MÜLBERGER, G. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", *14th IAPR International Conference on Document Analysis and Recognition* (2017). Online: https://ieeexplore.ieee.org/document/8270253 (accessed on October 22, 2018).

– MÜHLBERGER, G, TERBUL, T. "Handschriftenerkennung für historische Schriften. Die Transkribus Plattform", in *Bibliothek. Information. Technologie.*:, 21 (2018). Online: https://b-i-t-online.de/heft/2018-03/fachbeitrag-muehlberger.pdf (accessed on October 22, 2018).

– SOCIETY of AMERICAN ARCHIVISTS. *Encoded Archival Description*. Online: Official site: http://www.loc.gov/ead/index.html (accessed on October 22, 2018).

– STRAUSS, T., LEIFERT, G., GRÜNING, T., LABAHN, R. (2016) "Regular expressions for decoding of neural network outputs", *Neural Networks* 79, p. 1-11 (2016). Online: http://linkinghub.elsevier.com/retrieve/pii/S0893608016000447 (accessed on October 22, 2018).

– WEINGART, S. "Demystifying Networks", *Journal of Digital Humanities*, 1, 2011. Online: http://journalofdigitalhumanities.org/1-1/demystifying-networks-by-scott-weingart (accessed on October 22, 2018).