Transcribing and editing digitized sources on work in the textile industry

Jean-Damien Généro, Alix Chagué, Victoria Le Fourner, Marie Puren

TRANSCRIBING AND EDITING DIGITIZED SOURCES ON WORK IN THE TEXTILE INDUSTRIES

Jean-Damien Généro (CNRS), Alix Chagué (Inria), Victoria Le Fourner (MESHS), Marie Puren (Epitech)

September 9th, 2021

TIME-US. Remuneration and Time Use in the Textile Industries in France in the Longue Durée

OUTLINE

1. Between tradition and modernity

2. From data to corpora

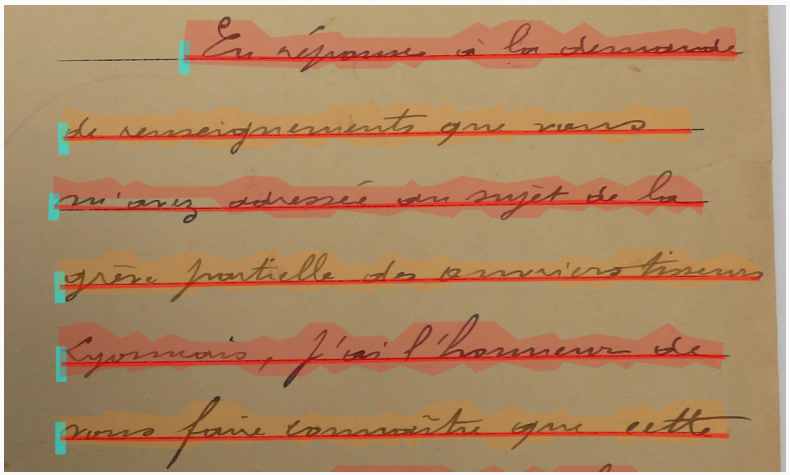
3. Information extraction

4. Paving the way for future projects

1. BETWEEN TRADITION AND MODERNITY

Transcribing

Produce, with a computer, textual data extracted from a corpus of digitized primary sources



Editing

Structure the extracted historical information and provide online access to it

```
    <div n="002" type="section">
     <head type="section" xml:id="part_02">
      OBSERVATIONS PRÉLIMINAIRES  DÉFINISSANT LA CONDITION DES DIVERS MEMBRES DE LA
     </head>
     <div n="001" type="sub_section">
      <head type="sub_section" xml:id="part_02_01">
       I. Définition du lieu, de l'organisation industrielle et de la famille
      </head>
      <div n="001" type="sub_sub_section">
       <head type="sub_sub_section" xml:id="part_02_01_01">
        § 1er. — ÉTAT DU SOL, DE L'INDUSTRIE ET DE LA POPULATION.
       </head>
       <p xml:id="para_6_67_12">
        <pb facs="#facs_72" n="67" source="lesouvriersdesde03sociuoft_0075.xml"/>
        La commune de P*** qu'habite la famille est située à 35 kilo-mètres E.-N.-E
E. d'Aix ; elle a pour chef-lieu de canton le village de T* dont elle n'est éloigné
territoire de la commune s'étend sur le flanc d'un coteau exposé au nord, et descen
une plaine que traverse la route d'Aix à Dragguignan et dont l'horizon est borné pa
Toute cette contrée repose sur les calcaires marneux désignés par les géo
        <pb facs="#facs_73" n="68" source="lesouvriersdesde03sociuoft_0076.xml"/>
```

Computer long used by historians (Lemercier and Zalc 2019)

A few milestones :

– 1959 : 'mechanographic' (computer) methods for reading notarial archives (Daumard and Furet 1959)
– 1960s : emergence of historians making intensive use of computers (Corvisier 1964, Le Roy Ladurie 1966)
– 1968 : Emmanuel Le Roy Ladurie, « La fin des érudits » in Le Nouvel Observateur : "The historian of tomorrow will be a programmer or he will not be [1]. "
– 1971 : contribution of computers to the reconstruction of demographic and economic data series (Furet 1971)

Studying words extracted from historical sources : an activity that has interested historians **since the 1970s** in the context of discourse analysis. Pioneering work by Régine Robin (Robin 1973)

1. Translation from : Rabb, Theodore K. and Rotberg Robert I. (2017) The New History : The 1980s and Beyond. Princeton University Press

1.3 AN ETERNAL RETURN (1)

Renewed interest in quantitative history after some disaffection (Ruggles 2021)

Explosion of the Web at the end of the 1990s

– New way of communicating results with colleagues and consumers of history : birth of the concept of ”digital history” in the Anglo-Saxon world (Mintz 2008)

– Access to a reservoir of historical data : digitization of Cultural Heritage (Google Books, Gallica, etc.) + ”born-digital archives” which require new ways of reading sources (Clavert 2014 and 2016)

Revived interest in methods for exploring these large corpora of data

– Early 2000s : ”distant reading” advocated by Franco Moretti in the context of literary history (Moretti 2013 and 2005)

– ”Culturomics” project supported by Google (Michel et al. 2011) : birth of Ngram Viewer

– Debate sparked (Annales 2015/2) by Jo Guldi and Armitage’s The History Manifesto (2014)

1.4 AN ETERNAL RETURN (2)

Renewed interest in quantitative history after some disaffection (Ruggles 2021)

Advances in computing

– Microcomputing

– Easy-to-use softwares (TXM, Gephi…)

– Promising new methods. For example : LDA algorithm (Blei, Ng, Jordan 2003) which gives rise to topic modelling, AI (neural network) and its applications in computer vision (**OCR = Optical Character Recognition** and **HTR = Handwritten Text Recognition**)

Construction and organization of the digital humanities community in the Anglo-Saxon world (Schreibmann, Unsworth, Siemens 2004) and then in France (Mounier 2010)

– Increasing visibility of digital methods in the human and social sciences

– Growth of resources for training (Web + growth of academic training such as masters in digital humanities)

2. FROM DATA TO CORPORA

– Monographies des Ouvriers des deux mondes (1857-1913, 1930) :

+ **124 sociological surveys** published by **F. Le Play** (†1882) and the **International Society of socio-economic practical studies**;
+ **same organization** all along + c**ross reference system**.

– Minutes des prud'hommes parisiens (1847-1848, 1858, 1878) :

+ **Legal records, written by the court clerks**;
+ Study per decade of audiences : Structure following the course of the hearing.

2.2 LOGICAL STRUCTURE EXTRACTION 1

Goal : Starting from the digitized images, reproduce the logical structure of the monographies to create interactive tables of contents

1. Collecting the digitizations;
2. Semi-automatic pre-processing of series and images;
3. Collecting clues allowing to detect the logical structure;
4. Automatic analysis of the layout;
5. Automatic transcription;
6. Generating standardized files (TEI XML) fully structured;

2.3 LOGICAL STRUCTURE EXTRACTION 2

We want to find **a balance between the benefits** of automation (saving time) **and its drawback** (more noise, risk to lose information).

- The resulting files must be controlled and manually corrected;
- Each automated step increases the risk of errors or loss;
- Keeping links between images and text ease controls;
- We don't have to automate tasks that are too difficult (e.g. : processing tables);

Then, we can add in the files metadata collected by researchers

2.4 PRUD'HOMMES' STRUCTURE 1

Objective : Make the **internal structure** of the minutes **visualizable and searchable** in order to identify each day, each hearing, each trial and the parties to the trial.

The automatic identification of these parties enables researchers to **easily compare the same type of data**.

As **the procedure of the hearing is always the same**, the structure of the minutes is perfectly homogeneous :

- identification of parties;
- point de fait;
- point de droit;
- judgement.

2.5 PRUD'HOMMES' STRUCTURE 2

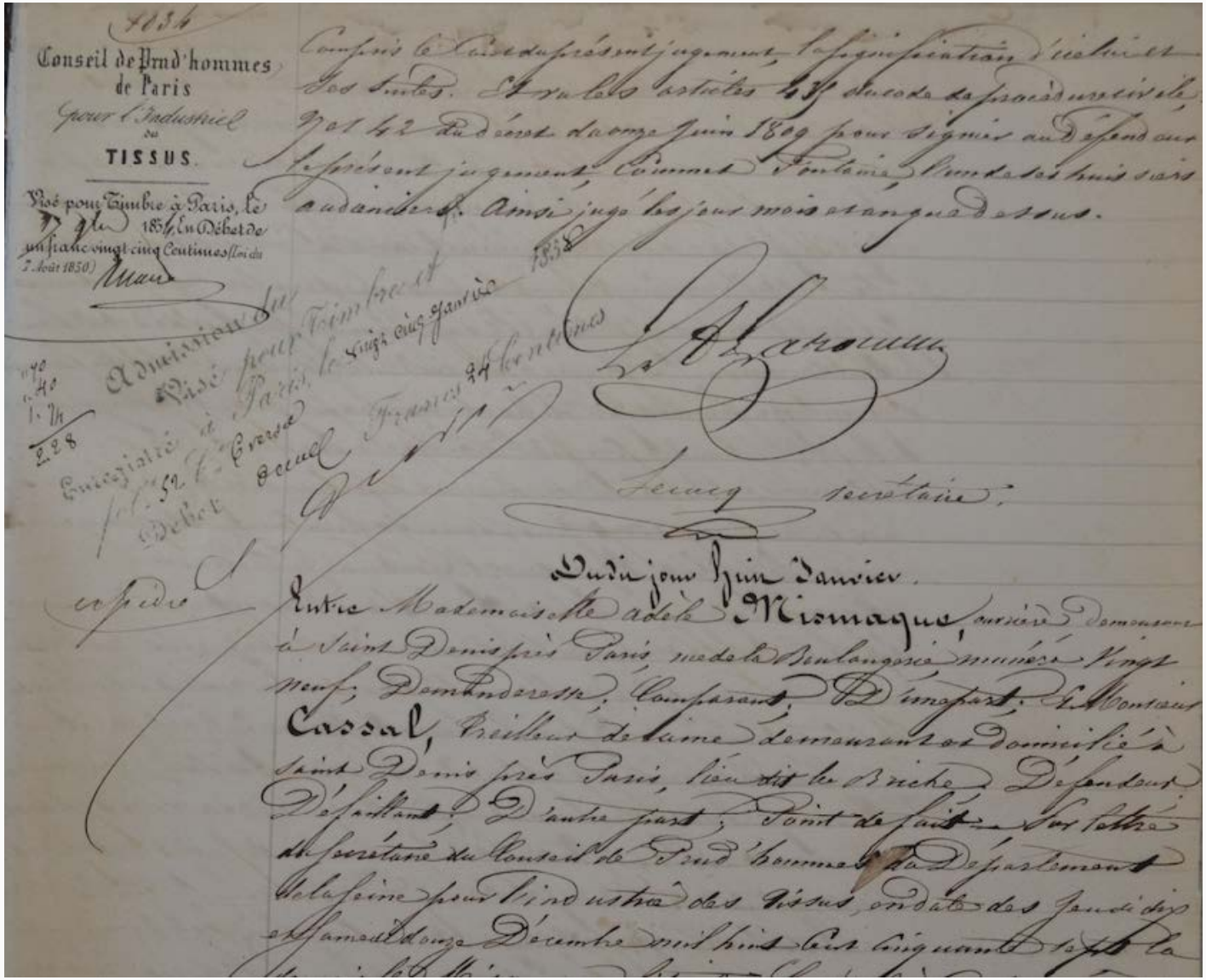


FIGURE : View of a minutes archive page

2.6 PRUD'HOMMES' STRUCTURE 3

```
<div n="1" type="courtHearing">
    <opener>Audience Du Vendredi huit Janvier mil huit en cinquante huit
        Siégeaient : Messieurs Larouette, vice président du Conseil de Prud'hommes
        du Département de la Seine, Président d'audience en remplacement de Monsieur
        Biétry, président du dit conseil, empêché pour cause légitime Adolphe
        Michel, Marienval, Dubosc et Ancelin prud'hommes assistés de Monsieur
        Lecusq, secrétaire du dit conseil. </opener>
    <div n="1" type="case">
        <div type="identificationParties">
            <p>Entre Monsieur Frédéric Auguste Albarède, demeurant à paris, rue
                notre dame de Nazareth, numéro trente cinq, agissant au nom et comme
                chargé de l'administration de la personne et des biens de sa fille
                mineure Hélène Albarède, ouvrière modiste ; Demandeur ; Comparant ;
                D'une part ; et Mademoiselle Desgranges, maitresse modiste,
                demeurant et domiciliée à Paris, rue Montorgueil, numéro quarante
                sept ; Défenderesse, défaillant. D'autre part ; </p>
        </div>
        <div type="pointDeFait">
            <p>Point de fait – Par lettres du secrétaire du Conseil de Prud'homme
                pour l'industrie du tissus, établi à Paris pour le Département de la
```

FIGURE : Extract from the automatically annotated file

A website was designed thanks to a searchable XSLT transformation :
`http://timeusage.paris.inria.fr/prudhommes-paris-19e/home.html`

3. Information Extraction

§6 « Propriétés » (Properties);

– « Immeubles » (estate);
– « Argent » (money);
– « Animaux » (animals);
– « Matériel spécial des travaux et industries » (specific equipment for works and industries).

§10 « Habitation, mobilier et vêtements » (Housing, furniture, clothing).

– « Meubles » (furniture);
– « Linge » (laundry);
– « Ustensiles » (utensils);
– « Vêtements » (clothing).

VÊTEMENTS DE LA FEMME (447f 50) : ceux portés par la classe ouvrière aisée.

1° *Vêtements du dimanche.* — 20 chemises en toile, 70f ; — 1 robe en laine brochée, 25f ; — 1 robe en mérinos, 20f ; — 1 robe en reps gris, 15f ; — 1 robe en flanelle à pois, 18f ; — 1 robe en indienne, 16f ; — 1 tablier en soie noire, 6f ; — 1 jupon en mérinos ouaté, 7f ; — 1 tablier en mérinos noir, 4f ; — 1 jupon de molleton marron, 10f ; — 1 jupon de flanelle verte, 6f ; — 1 jupon en indienne, 3f ; — 1 jupon en reps gris, 4f ; — 1 jupon en calicot blanc, 6f ; — 2 cols brodés, 3f ; — 4 cols unis, 3f ; — 1 bonnet à rubans blancs, 12f ; — 1 bonnet à rubans bleus, 6f ; — 4 bonnets unis, 8f ; — 1 foulard en soie, 6f ; — 1 châle en mérinos noir, 8f ; — 1 châle en mousseline fantaisie, 10f ; — 1 châle en laine à fleurs, 10f ; — 1 caraco en mérinos ouaté, 15f ; — 1 caraco en orléans, 6f ; — 3 paires de bas noirs en laine, 6f ; — 3 paires de bas blancs en coton, 6f ; — 1 paire de souliers en drap, 6f ; — 1 paire de souliers en cuir, 8f ; — Sabots, chaussons, etc., 6f. — Total, 329f.

FIGURE : « L'Ouvrier éventailliste de Sainte-Geneviève (Oise - France) », Les Ouvriers des deux mondes, n°40, 1875 [1885], p. 124.

– Initial situation : §6 and §10 only are identified in the XML.

– Goal : identify sections' divisions and made up a database.

– Tools : Python & SQL.

1. Add `<div>` with a specific `@type` for each inventory;
   $\longrightarrow$ MANUAL TASK (HUMAN)

2. Get the text of the `<div>` and inject it into a database;
   $\longrightarrow$ AUTOMATIC TASK (PYTHON & PYTHON SQL ALCHEMY)

3. Make an SQL query to display the desired text;
   $\longrightarrow$ AUTOMATIC TASK (PYTHON SQL ALCHEMY)

4. Display : website.
   $\longrightarrow$ AUTOMATIC TASK (PYTHON FLASK / HTML)

4. Paving the way for future projects

Accept and use trial and error

Difficulties inherent in this type of project that should not be hidden

- Data vs capta (Drucker 2011) : data is never ”given” but ”taken”
- A lot of work, time and money to get textual data that can be used by a computer
- A big challenge to appreciate and deal with these difficulties

Digital humanities context allows for more artisanal approaches

- Field still under construction that encourages experimentation
- DH = ”lieu de bricolage” (Clavert and Schafer 2019)

Design acquisition and processing chains for historical sources

Share them

Facilitate their appropriation by historians themselves, to allow their evaluation and calibration (Lamassé and Rygiel 2013) by :

- Giving access to the algorithms and code
- Allowing fine-tuning of these instruments
- Enabling users to transform and modify these tools according to their needs

Work as a team

Central role of the historian in this type of project (Lamassé and Rygiel 2013, Grandi and Ruiz 2012).

- Constitution of corpora
- Information extraction (what questions can the data answer ?)
- Structuring of historical information

Impossible to work alone : necessary to rethink the ”graph of one’s professional relations” (Lamassé and Rygiel 2013)

- Usual collaboration with librarians and archivists
- Necessary **collaboration with computer scientists** : initiate a dialogue aimed at understanding the specificities of the historical material, and those of the computer tool (Grandi and Ruiz 2012)
- New collaboration with physicists (signal acquisition) and mathematicians (Lamassé and Rygiel 2013)

Experimental aim of TIME US : new questions arise from collaboration

Train yourself and others

To be able to fulfil the above objectives, it is absolutely necessary to train historians (Heimburger and Ruiz 2011) in order to :

- Make informed choices
- Facilitate the appropriation of these new tools
- Understand the limitations of these methods

An appropriate training to be widely disseminated in order to :

- Develop a digital literacy common to all historians
- Abolish the frontier between ”digital history” and history as such. (Grandi and Ruiz 2012)

Developing a common digital culture between historians, engineers and computer scientists

PLEASE RT!

Thank you for your attention!

Jean-Damien Genero @JdGnr

Alix Chagué @Alix_Tz

Victoria Le Fourner @victoriavicoos

Marie Puren @MariePuren

BIBLIOGRAPHY

– Blei David M., Ng Andrew Y. and Jordan Michael I. (2003). ”Latent dirichlet allocation”, In : Journal of machine Learning research vol. 3, p. 993-1022

– Clavert Frédéric (2014). « Vers de nouveaux modes de lecture des sources », In : Le temps des humanités digitales. FYP EDITIONS.

– Clavert Frédéric (2016). « Une histoire par les données ? Le futur très proche de l'histoire des relations internationales », In : Bulletin de l'Institut Pierre Renouvin, 2016, pp.119-130.

– Clavert Frédéric et Schafer Valérie (2019). « Les humanités numériques, un enjeu historique », In : Quaderni n°98 2019/1, pp. 33-49.

– Corvisier André (1964). L'Armée française de la fin du XVIIe siècle au ministère Choiseul. Publications de la Faculté des lettres et sciences humaines

BIBLIOGRAPHY

– Daumard Adeline et Furet François (1959). « Méthodes de l'Histoire sociale : les Archives notariales et la Mécanographie », In : Annales. Economies, sociétés, civilisations 1959/4. pp. 676-693

– Drucker Johanna (2011). "Humanities Approaches to Graphical Display", In DHQ : Digital Humanities Quaterly, vol. 5, n°1, 2011.

– Furet François (1971). « Histoire quantitative et construction du fait historique », In : Annales. Economies, sociétés, civilisations 1971/1, pp. 63-75.

– Grandi Elisa e Ruiz Émilien , « Ce que le numérique fait à l'historien.ne. Entretien avec Claire Lemercier », In : Diacronie n°10, 2 | 2012. URL : http ://journals.openedition.org/diacronie/2780.

– Guldi Jo and Armitage David (2014). The History Manifesto, Cambridge University Press.

BIBLIOGRAPHY

– Heimburger Franziska et Ruiz Émilien (2011). « Faire de l'histoire à l'ère numérique : retours d'expériences », In : Revue d'histoire moderne & contemporaine 2011/5, n°58-4bis. URL : https ://www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2011-5-page-70.htm.

– « La longue durée en débat » (2015). In : Annales. Economies, sociétés, civilisations 2015/2, 240 pages.

– Lamassé Stéphane et Rygiel Philippe (2013). « Nouvelles frontières de l'historien », In : Revue Sciences/Lettres, 2 | 2014. URL : http ://journals.openedition.org/rsl/411.

– Lemercier Claire and Zalc Claire (2019). Quantitative Methods in the Humanities. University of Virginia Press.

– Le Roy Ladurie Emmanuel (1966). Les paysans de Languedoc. S.E.V.P.E.N.

BIBLIOGRAPHY

– Michel Jean-Baptiste et al. (2011). ”Quantitative Analysis of Culture Using Millions of Digitized Books”, In : Science, vol. 331, issue 6014.

– Mintz Steven (2008). ”Interchange : The Promise of Digital History”, In The Journal of American History, vol. 95, no 2, 2008, pp. 25-29.

– Moretti Franco (2005). Graphs, Maps, Trees : Abstract Models for Literary History. Verso.

– Moretti Franco (2013). Distant Reading. Verso

– Mounier Pierre (2010). « Manifeste des Digital Humanities ». In : Journal des anthropologues 122-123, pp. 447-452.

– Robin Régine (1973). Hstoire et Linguistique. Armand Colin.

– Ruggles Steven (2021). ”The Revival of Quantification : Reflections on Old New Histories”, In : Social Science History 45(1), pp. 1–25.

– Schreibman Susan, Siemens Ray and Unsworth John (ed.) (2004). A Companion to Digital Humanities. John Wiley & Sons.