This presentation introduces Time-Us research project involving economic history and gender studies, on the topic of textile trades in France from the late 17th century to the early 20th century. It started in January 2017 and is to last until late 2019.

The project is still at an early stage of development, so I don't intend to present results, but the project's main goals, and rather the methods put into place to digitally process our historical sources and to generate high-quality data.

The role of women in industrial development is now largely recognized in both sociological and economic studies on developing countries, and the historiography of the first industrial revolution in Europe. Yet data on their remuneration, schedules and domestic work and that of men working in the same sectors remain deficient for many regions, especially for France. The project aims at reconstructing remunerations and time budgets of women and men working in the French textile industry. It focuses not only on paid work, but also unpaid work such as domestic tasks, in the four main French industrial regions : Lille, Paris, Lyon and Marseille.

The project plans:

- on treating both qualitative documents and quantitative data to better document the role of women in industrial development on a long-time period;
- it also plans on creating high-quality data in the form of time series data, statistics, vocabularies and annotated corpora.

To achieve these goals, Time-Us gathers together economic and labour historians, but also sociologists and statisticians to tackle these data ~~and produce knowledge~~. Natural Language Processing experts and data curators help ~~to~~ produce and modelize these data.

The project is based on the cooperation of five laboratories located all over France. Each of these laboratories tends to focus on their own geographical area. Each laboratory is to share, confronte and compare the sources and information they can find.

On top of these 5 labs, the Almanach team, based at Inria in Paris, brings a technical support for the implementation of natural language processing analysis and data curation, and the choice of softwares and modelizations.

One of the main questions raised by Time-Us is a methodological one: what steps to follow to go from paper sources to NLP treatment?

Tackling this question brings up the need for guidelines to implement good practices from the team members on many different levels, including how to get good quality digitizations, metadata, transcriptions… Moreover, we need to create a gold standard corpus that will be used by machine learning experts. This gold standard is the final version of the annotated data, tagged according to the most recent guidelines. Besides we talk about "gold" data because we are not only providing an up-to-date annotated corpus, but we are also following an international standard to modelize these data, namely the Text Encoding Initiative.

Slide 5

Teams involved in the project have adopted different strategies to collect primary sources. It is important to underscore that all laboratories collaborate but still work independently from each other. They adopt their own approach to the research, adapt to their sources and still make sure crossing sources is possible.

First, Time-Us researchers tried to find similar series for the four regional case studies, for instance professional court records.

The strategies adopted for 19th century judiciary sources in Lille, Paris and Lyon offer a good illustration of the variety of our approaches.

- Marseille judiciary archives about textile trades during the modern period were partially lost and turned out to be too sporadic to be of any use. It was decided to leave them out of the study.
- In Lille, existing handwritten labour court archives are collected only when court decisions include at least one woman. When no woman was involved, the court decision was left out. Such a strategy means that women are often enough the protagonists in such judgements, and therefore that they do play a significant role in the textile industry in the North of France.
- On the contrary, the important mass of court decisions in Paris doesn't make it possible to efficiently sort court decisions. So in Paris the strategy is slightly different: every court decision on textile for one sample year (1858) is collected, whether it is dealing with women or not.
- In Lyon, there no longer are any handwritten archives from the labour court. Therefore, the team focuses on collecting workers' printed press were labour court decisions are summed up, whether or not they are dealing with women. Researchers collect also other kind of archives, such as police reports on strikes.

So far, the team gathered a great variety of document spanning from the very late 17th century to the early 20th century. Part of these documents are printed. They include posters and petitions produced by workers, workers'/working class newspapers, and sociological surveys on working classes and their life conditions such as Le Play's monographies *Les ouvriers des deux mondes*.

However, most of the sources gathered for the project are handwritten documents, such as court decisions - and more specifically labour court decisions -, police reports, company's archives, personal archives, surveys and petitions. If OCR technologies let us work with printed documents rather easily, the wide variety of paper and writing qualities and the very nature of handwritten text have posed a challenge to us.

Since the project aims at gathering an unprecedented amount of archival documents, manually transcribing each one of them cannot be satisfying. In such case, using automated transcription is necessary.

Time-Us uses the platform Transkribus, which enables detecting lines of text and aligning them with their transcriptions. These textual data are used as ground truth to train models.

Which means the accuracy of these data is guaranteed, because transcribers precisely respect the original text during the transcription phase.

With about 50 to 80 pages manually transcribed, the Transkribus team is able to train an Handwritten Text Recognition model on these materials. It should be noted that this HTR model can be trained and applied to an homogeneous corpus of documents, namely written by the same hand. Then it is possible to automatically transcribe hundreds of remaining pages. So since the writing style are various, several sets of training data need to be produced and corrected.

In order to spot and extract data to create time series data or vocabularies, we need to annotate these transcriptions.

Transkribus provides a tool to manually tag the corpus, so we have created our own set of elements to match our needs. It goes along with an annotation guide available for all of our collaborators on our “wiki” website. Annotating the text includes tagging elements such as matrimonial statuses, job titles, but also sums of money, types of products, work rhythms, performed tasks and any sort of passage in a sentence evoking a remuneration.

Tags are viewable in Transkribus and can be exported into XML files. But we need to process these files with XSLT stylesheets to obtain good quality, TEI-conform documents.

Transcribed and annotated texts will be displayed on the wiki website eventually.

Our intention is to make this layer of annotation visible on the website as well, by using and adapting the TEITags extension, designed for the Transcribe Bentham project. For the moment, we did a few tests, but we are currently waiting to stabilize our annotation model before adding new tags into this extension.

We also intend for the layer of annotation to be automatically generated thanks to manually annotated texts used, here too, as training data. This is where NLP specialists are essential.

Not only would NLP treatments enable us to automatically annotate the text, it also makes it possible to correctly link informations to recreate their context. This is necessary if we want to be able to generate time series: wages and payments are expressed in such a great variety of forms that isolating them from the context given by the sentence or even the document would make them meaningless.

We will use FRMG parser in order to produce dependencies between words and reconstruct data as complete as possible. We can identify who does what, when, how and against what payment. We will be able to associate a matrimonial status to a type of remuneration in a sentence, or a type of product to the payment it induces.

In our case, specific issues may arise because of the quality of the transcriptions and the peculiarities of the language used, which contains archaic constructions, whereas our parser was designed for contemporary French.

Time-Us aims at treating qualitative documents to produce quantitative data with the help of computational analysis, especially methods provided by Natural Language Processing.

- The project is still at an early stage of development. We are still running some tests, and we don't have significant results yet. Fortunately the project will run for another 18 months.
- Even if significant progress have already been made, we are still preparing the corpus. So far,
    - the available historical sources have been listed;
    - series of comparable handwritten documents have been photographed, and then transcribed, and a part of the rich corpus formed by Le Play's printed monographies has been ocerized;
    - some parts of the corpus have been annotated and will be of use for conducting the first NLP analysis.
- And last but not least, establishing such a methodology has clearly demonstrated that a set of best practices are needed for this kind of project, from its very beginning. For instance, it is necessary to take good-quality pictures of the archives, by respecting a set of simple guidelines. Producing a gold standard corpus for machine learning from various historical sources takes time, but our workflow will ensure the quality of the final results. This workflow will be soon stabilized, and could then be reused by similar research projects.

Bibliography:

Budin, Jean-François, « Les ouvrières de la soie à Lyon au XVIIe siècle », in *Le travail avant la révolution industrielle*, 127e congrès national des sociétés historiques et scientifique, éd. Maurice Hamon, 2006, éd. électronique, pp. 117-126. http://cths.fr/ed/edition.php?id=4401

Burnette, Joyce (2008), *Gender, Work and Wages in Industrial Revolution Britain*, Cambridge, Cambridge University Press.

Davidoff, Leanore. and Hall Catherine (1987), *Family Fortunes*: *Men and Women of the English Middle Class 1780-1850*, Chicago, University of Chicago Press.

Duflo, E. (2012), "Women Empowerment and Economic Development", *Journal of Economic Literature* 50(4), 1051–1079.

Goldin, C. (1995), "The U-Shaped Female Labor Force Function in Economic Development and Economic History", in Schultz, TP, *Investment in Women's Human Capital and Economic Development*, Chicago, University of Chicago Press, pp. 61-90.

Hafter, Daryl, *Women at Work in Preindustrial France*, University Park, PA, Pennsylvania State University Press, 2007.

Hafter, «Daryl, Stratégies pour un emploi : travail féminin et corporations à Rouen et à Lyon, 1650-1791 », *Revue d'histoire moderne et contemporaine*, janvier-mars 2007, 54-1, pp. 98-115.

Horrell, Sara and Humphries, Jane (1995), "Women's Labour Force Participation and the Transition to the Male-Breadwinner Family, 1790-1865", *Economic History Review* 48: 1, pp. 89-117.

Humphries, J., and C. Sarasua (2012), "Off the Record: Reconstructing Women's Labor Force Participation in the European Past", *Feminist Economics* 18:4, pp. 39-67.

Humphries, J. (1996), “Women and Paid Work”, in Purvis, J. (ed.), *Women's History: Britain, 1850–1945*, edited by, UCL Press, 85–106.
Humphries, J., and J. Weisdorf (2015), “The Wages of Women in England, 1260-1850”, *Journal of Economic History* (forthcoming)
Juratic Sabine and Pellegrin, Nicole, « Femmes, villes et travail en France dans la deuxième moitié du XVIIIe siècle : quelques questions », *Histoire, économie et société*, 1994, n° 3, pp. 477-500.
Manzel, K. and J. Baten (‘2009), “Gender Equality and Inequality in Numeracy – the Case of Latin America and the Caribbean, 1880-1949”, *Revista de Historia Económica – Journal of Latin American and Iberian Economic History* 27:1, pp. 37-74.
Monique Meron, Margaret *Maruani, Un siècle de travail des femmes en France,* Paris, La Decouverte, 2012.
Nederveen Meerkerk, Elise van (2010), “Market Wage or Discrimination? The Remuneration of Male and Female Wool Spinners in the Seventeenth-century Dutch Republic”, *Economic History Review* 63, pp. 165-186.
Ogilvie, Sheilagh (2003). *A Bitter Living: Women, Markets, and Social Capital in Early Modern Germany*. Oxford: Oxford University Press.

Schmidt, Ariadne and Nederveen Meerkerk, Elise (2012), “Reconsidering the ‘first male breadwinner economy’: long-term trends in female labour force participation in the Netherlands, c. 1600-1900’, *Feminist Economics* 18, pp. 69-96.

Schweitzer, Sylvie, *Les femmes ont toujours travaillé. Une histoire du travail des femmes aux XIXe et XXe siècles*, Paris, Odile Jacob, 2002

Shaw-Taylor, Leigh (2007), “Diverse experiences: the geography of adult female employment in England and the 1851 census,” in N. Goose (ed.), *Women’s work in Industrial England. Regional and local perspectives*, Hatfield, Hertfordshire University Press.

Valenze, D.M. (1995), *The First Industrial Woman*, New York, Oxford University Press.

van Zanden, J.L. (2011), “The Malthusian intermezzo: Women's wages and human capital formation between the late Middle Ages and the demographic transition of the 19th century”, *History of the Family* 16, pp. 331–342.

Wall Richard, « The contribution of married women to the family economy under different family systems: some examples from the mid-nineteenth century from the work of Frédéric Le Play ». In Antoinette Fauve-Chamoux, Sölvi Sogner (eds), *Socio-economic consequences of sex-rations in historical perspective, 1500-1900*, Università Bocconi. Vol. B5 Procedins Eleventh International Economic History Congress, 1994, p. 139-148