Exploring silent social changes: the digital exploitation of a big mine of historical data (Catalonia, Eighteenth century)

Rosa Congost, Jordi Regincós, Rosa Ros & Enric Saguer

Universitat de Girona

Fundación BBVA

A collaborative project

Pattern Recognition and Human Language Technology Research Center

PRHLT

Universitat de Girona 25 anys

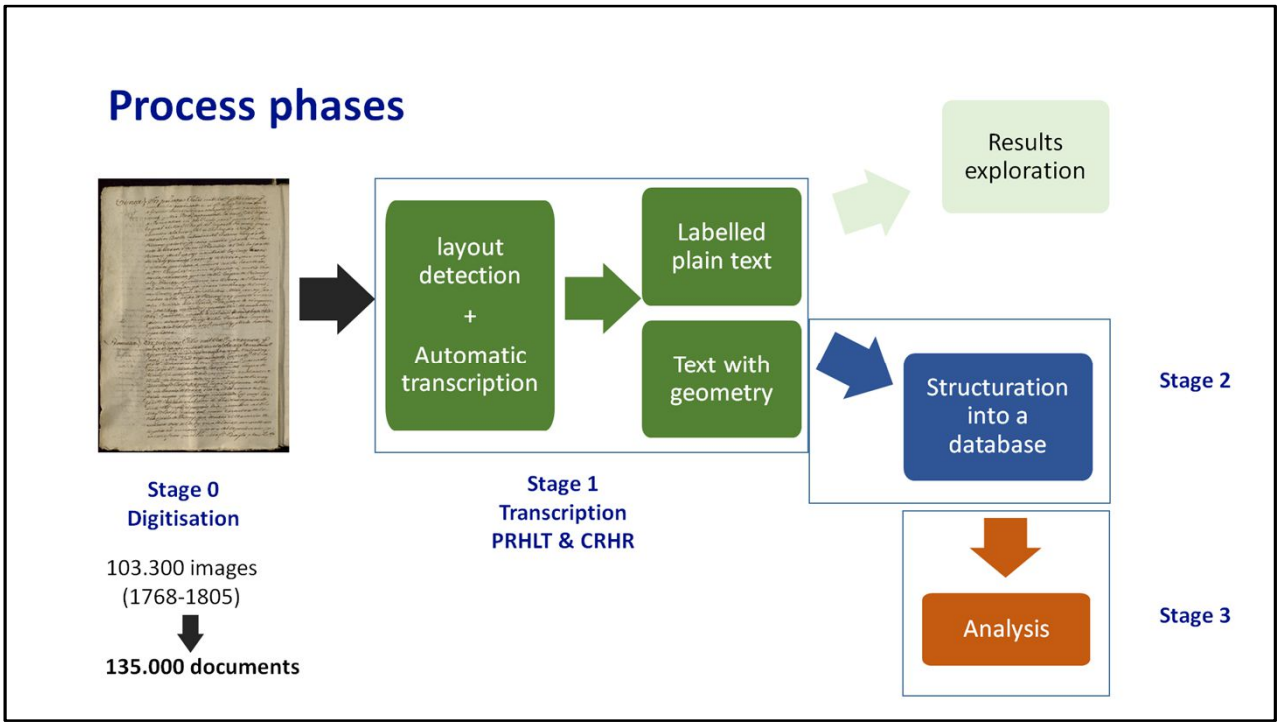
Centre de Recerca d'Història Rural (CRHR)

OBJETIVE

To apply automatic handwritten text recognition techniques to the digital images of a large documentary source

el *Registro de Hipotecas* (Registry of Deeds, 1768–1861) of the Girona region

in order to analyse processes of social change from mid eighteenth century to mid nineteenth century.



The project is being developed in four stages:
Firstly, the digitisation of the images, for the moment centred in the period 1768-1805.
Secondly, the transcription (layout detection + automatic transcription), from which we obtain a labelled plain text with geometry. As we will explain later, this plain text allows us to obtain some results. However, in order to increase the analytical potential, the plain text is structured into a database, which will be the main instrument for the study of social change.
This procedure makes it possible the analysis of our 135.000 documents, which was impossible using the traditional methods. We have estimated that the construction of a database manually would have needed 117.000 hours of work, that is, for example, the work of 6 students, 15 hours a week during 25 years!

What we are going to explain is:

1. Content and possibilities of the Registry of Deeds

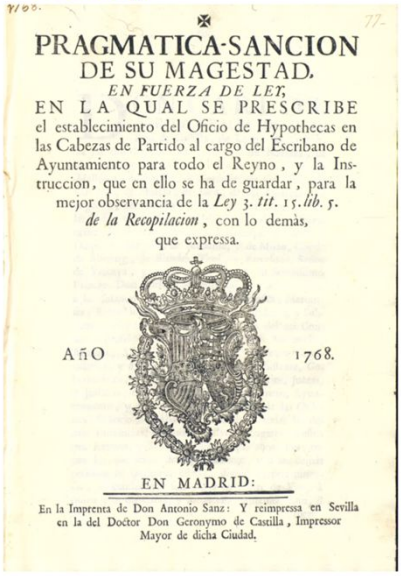
2. Current state of the project

3. What do we do with the transcript and labelled text?

4. Conversion of the plain text into a database

The *Registro de Hipotecas (Registry of Deeds)*

- Registry institution created in 1768
- It contains **notarial contracts** with charges and mortgages
- The application criteria were not homogeneous in the Spanish territory
- In Catalonia, it includes a large number of notarial documents (all of those containing general mortgage guaranties)



PRAGMATICA-SANCION
DE SU MAGESTAD.
EN FUERZA DE LEY,
EN LA QUAL SE PRESCRIBE
el establecimiento del Oficio de Hypothecas en las Cabezas de Partido al cargo del Escribano de Ayuntamiento para todo el Reyno, y la Instruccion, que en ello se ha de guardar, para la mejor observancia de la *Ley* 3. *tit.* 15. *lib.* 5. *de la Recopilacion*, con lo demàs, que expressa.

AÑO 1768.

EN MADRID:

En la Imprenta de Don Antonio Sanz: Y reimpressa en Sevilla en la del Doctor Don Geronymo de Castilla, Impressor Mayor de dicha Ciudad.

Content of the Registry of Deeds in Girona

Type of documents in the Registry of Deeds, (Girona region, 1768-1770)

	n	%
Loan and other credits contracts	3.256	29%
Sales and purchases	2.465	22%
Marriage contracts and wills	1.428	13%
Emphyteutic contracts	1.143	10%
Leases and sharecropping contracts	896	8%
Receipt of payment	418	4%
Donations	275	2%
Judicial and extrajudicial agreements	215	2%
Inventoies and auctions	135	1%
Guarantees	132	1%
Price and rent modifications	69	1%
Account description	52	0,5%
Capbrevacions	39	0,3%
Powers of attorney	8	0,1%
other	705	6%
	11.236	**100%**

nota: Estimación realizada a partir de los 4 primeros volúmenes

The first of its advantages is the wide range of notarial documents that it contains. It includes notarial acts such as sales and purchase agreements, leasing and emphyteutic contracts, credit operations (of diverse types, especially census and liabilities), marriage contracts, etc. To this diversity of operations, we may also add its considerable representativeness and its limited social bias, since Catalonia was a region having a strong tradition of visits to the notaries, even by the more humble groups. All this contributed to enrich the volume of information contained in the source.

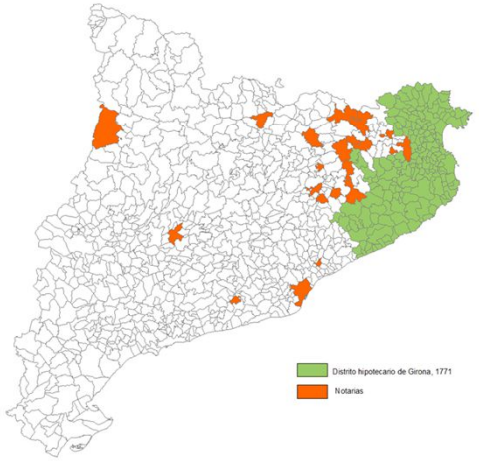
Content of the Registry of Deeds in the Girona region

- Wide extension
  - 1768-1773 = 3.884 km^2
- Obligation to present the notarial deed in the registry office where the affected estate is located

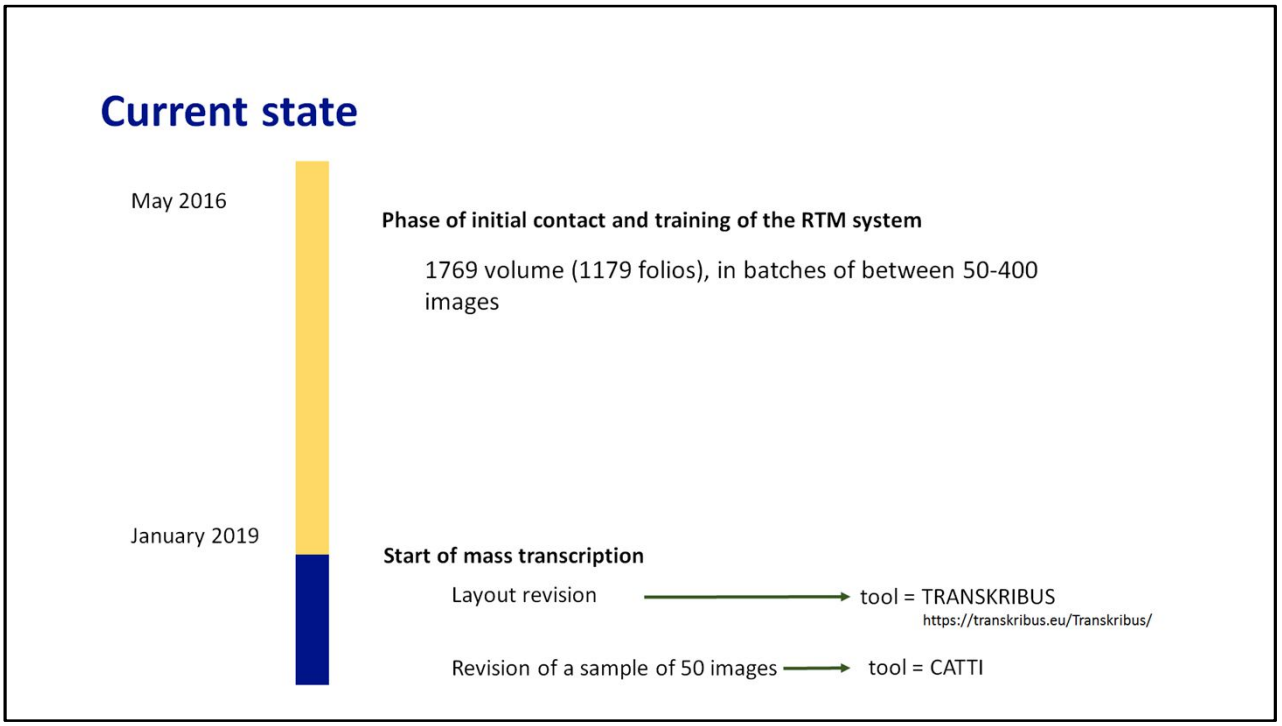
  ⇩

  The registry includes all the deeds referred to estates located in the territory of the Registry office

Notary offices from where the deeds registered in the *Registro de Hipotecas* came, 1771



The geographic and temporal coverage of the source is another of its strong advantages. It collects, in a centralised manner, all information relating to a relatively extensive area, including those notarial deeds formalised in notary offices outside the region. Working with regional areas allows us to reach beyond the local level. In terms of chronology, the source covers a period of intense institutional change, including the transition of a society of the *Ancien Régime* to a liberal one.



After a period of initial development of the project and training the transcription system taking as a basis the 1769 volume, the mass transcription started in January 2019.

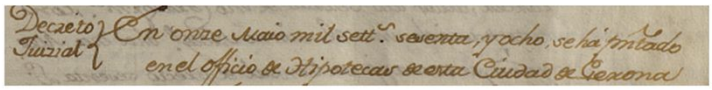
Current state

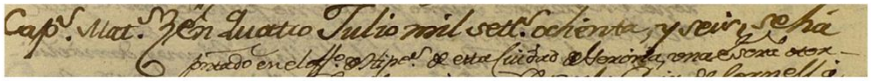
		Fase 0	Fase 1		Fase 2	
libro	año	digitalizado	revisión layout	transcripción	inclusión en BD	num imágenes
1	1768					2091
2	1769					2373
3	1769					1785
4	1770					2307
5	1770					1934
6	1771					2459
7	1771					2652
8	1772					2170
9	1772					1774
10	1772					1676

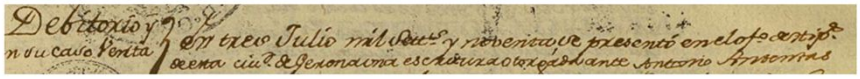
Now, we have the first 5 volumes transcripted and the first 4 with their data structured into a database.

A challenge: changes in the typeface

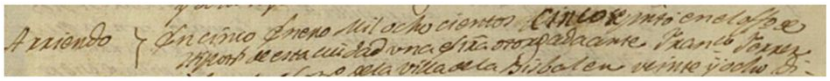
Training the transcription system → typeface 1769

1768 

1786 

1790 

1792

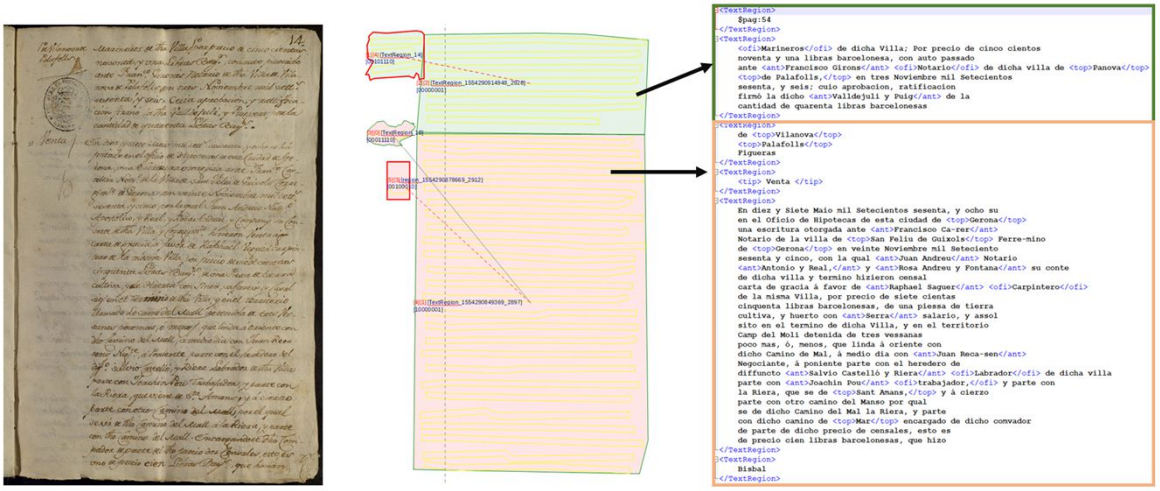
1805 

How will the transcription system work when the typeface changes?

The transcription system has been trained with the typeface of 1769. For this reason, one of the most important uncertainties is how the transcription will work when the typeface changes, especially from the last decade of the eighteenth century, when changes are more pronounced. It may be necessary to carry out a new training process.

What do we obtain?

1. A file of plain text for each image, distinguishing text regions



What we obtain with the transcription process is:

1. A file of plain text for each image with labels that are used to: a) distinguish text regions; b) qualify some words as anthroponyms, place names or occupations.

Transcript enhanced with xml labels + expanded abbreviations + word partitions

Text region

Place name

anthroponym

Occupation or social label

```
<TextRegion>
    <tip> Venta </tip>
</TextRegion>
<TextRegion>
    En diez y Siete Maio mil Setecientos sesenta, y ocho su
    en el Oficio de Hipotecas de esta ciudad de <top>Gerona</top>
    una escritura otorgada ante <ant>Francisco Ca-rer</ant>
    Notario de la villa de <top>San Feliu de Guixols</top> Ferre-mino
    de <top>Gerona</top> en veinte Noviembre mil Seteciento
    sesenta y cinco, con la qual <ant>Juan Andreu</ant> Notario
    <ant>Antonio y Real,</ant> y <ant>Rosa Andreu y Fontana</ant> su conte
    de dicha villa y termino hizieron censal
    carta de gracia á favor de <ant>Raphael Saguer</ant> <ofi>Carpintero</ofi>
    de la misma Villa, por precio de siete cientas
    cinquenta libras barcelonesas, de una piessa de tierra
    cultiva, y huerto con <ant>Serra</ant> salario, y assol
    sito en el termino de dicha Villa, y en el territorio
    Camp del Moli detenida de tres vessanas
    poco mas, ó, menos, que linda á oriente con
    dicho Camino de Mal, á medio dia con <ant>Juan Reca-sen</ant>
    Negociante, á poniente parte con el heredero de
    diffuncto <ant>Salvio Castelló y Riera</ant> <ofi>Labrador</ofi> de dicha villa
    parte con <ant>Joachin Pou</ant> <ofi>trabajador,</ofi> y parte con
    la Riera, que se de <top>Sant Amans,</top> y á cierzo
    parte con otro camino del Manso por qual
    se de dicho Camino del Mal la Riera, y parte
    con dicho camino de <top>Mar</top> encargado de dicho comvador
    de parte de dicho precio de censales, esto es
    de precio cien libras barcelonesas, que hizo
</TextRegion>
```

tei/170025120000001,0036.xml

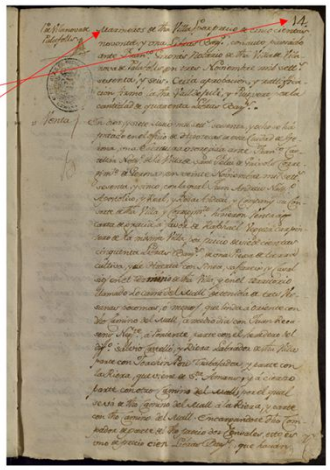
For example...

What do we obtain?

2. A file that relates each text region with the coordinates of the image → necessary for the database to identify the different parts of a document and to unify them

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="
http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd">
  <Metadata>
    <Creator>P2PaLA-PRHLT</Creator>
    <Created>2019-03-28T18:37:06</Created>
    <LastChange>2019-04-03T13:36:29.508+02:00</LastChange>
  </Metadata>
  <Page imageFilename="170025120000001,0036.tif" imageWidth="2790" imageHeight="4027">
    <ReadingOrder>
      <OrderedGroup id="ro_1554291391601" caption="Regions reading order">
        <RegionRefIndexed index="0" regionRef="TextRegion_1"/>
        <RegionRefIndexed index="1" regionRef="TextRegion_1554290914948_2926"/>
        <RegionRefIndexed index="2" regionRef="TextRegion_14"/>
        <RegionRefIndexed index="3" regionRef="TextRegion_18"/>
        <RegionRefIndexed index="4" regionRef="TextRegion_1554290849369_2897"/>
        <RegionRefIndexed index="5" regionRef="region_1554290878669_2912"/>
      </OrderedGroup>
    </ReadingOrder>
    <TextRegion id="TextRegion_1" custom="readingOrder {index:0;} structure {type:$pag;}">
      <Coords points="2666,102 2659,82 2619,58 2579,66 2448,70 2437,78 2437,90 2430,94 2426,129 2437,149 2430,153 2430,173 2452,188 2561,184 2611,196 2659,196
      2666,169"/>
      <TextLine id="TextLine_0_1" custom="readingOrder {index:0;} structure {type:$pag;}">
        <Coords points="2631,145 2562,117 2488,141 2503,188 2567,170 2563,171 2631,200 2644,148"/>
        <Baseline points="2500,177 2561,157 2623,183 2635,146"/>
        <TextEquiv>
          <Unicode>$pag:54</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
      <Coords points="755,157 1710,145 2107,153 2662,204 2630,1008 688,988"/>
      <TextLine id="TextLine_7_23" custom="readingOrder {index:0;} structure {type:$pac;}">
        <Coords points="2641,200 2371,194 1756,174 709,188 784,238 1753,224 2369,244 2640,250"/>
        <Baseline points="796,227 1754,212 2370,232 2641,238"/>
        <TextEquiv>
          <Unicode>$ofi:Marineros de dha$.dicha Villa; Por precio de cinco cientos</Unicode>
        </TextEquiv>
      </TextLine>
      <TextLine id="TextLine_5_23" custom="readingOrder {index:1;} structure {type:$pac;}">
        <Coords points="2584,289 2539,301 2396,302 792,308 793,358 2397,352 2545,351 2597,338"/>
        <Baseline points="815,349 2397,340 2544,339 2594,326"/>
        <TextEquiv>
          <Unicode>noventa y una libras bar.$^s$.barcelonesa, con auto passado</Unicode>
        </TextEquiv>
```

page/170025120000001,0036.xml



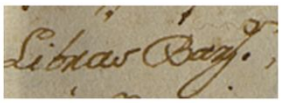
A transcript with a *certain* margin of error

1. Irrelevant errors

- Page numbers

$pag:54

- Gender or number concordance



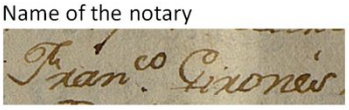
libras barcelonesa,

The result is a transcript with a certain margin of error. Some errors, such as page numbers or gender concordance, are irrelevant.

A transcript with a certain margin of error

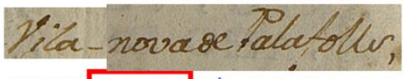
2. Easily recognizable and amendable errors

- Name of the notary

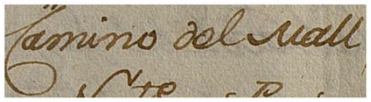


Francisco Girons

- Place names



<top>Panova</top> <top>de Palafolls



Camino del Mal
camino del Manso
camino de <top>Mar

Other errors are easily recognizable and amendable. For example, many place names or the name of the notary.

A transcript with a certain margin of error

3. Errors difficult to amend

- anthroponyms

Puig</ant>

- Incorrect labelling and transcription

<ant>Antonio

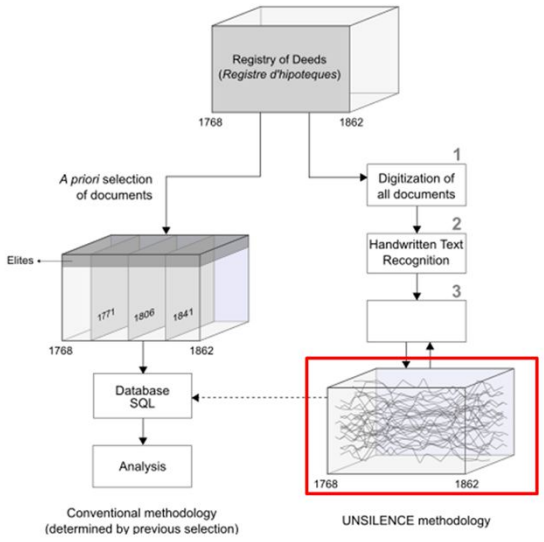
- Not transcribed words

se de dicho Camino

Others are more difficult or impossible to correct, like anthroponyms, incorrect labelling and not transcribed (or badly transcribed) words. The use of this technique implies the assumption of a certain margin of error, not as different as, for example, the tools widely used to find words in historical press.

What do we want to do with the transcript and labelled text?

General objective: Study of social change from c. 1750 to c. 1850



Until now, there were two ways of working with the data contained in a source like the Registry of Deeds. Both of these methods, represented on the left hand side of the graph, required the prior selection of materials and have been experimented in the books of the Registries of Deeds of Girona. The first consisted of selecting those contracts in which the elite were the key actors. For example: deeds of sale of estates having a value exceeding 1000 Catalan pounds or marriage contracts with dowries exceeding this same quantity. But this form of action, while offering some interesting results, makes it difficult to reconsider our opinion of the importance of the elite in history. The second consisted of examining all of the deeds corresponding to specific years. The results of this were quite spectacular, given that they highlighted an unexpected key role of the humble groups, highlighting the bias that may result when exploring social dynamism when only data from the elite or dominant groups is available. Given that these results came from relatively short periods of one year, the problem was how to correctly interpret the social changes that were detected between them. Did they occur quite suddenly or were they more gradual? Were they the result of some historic event? For example: regarding the changes observed in 1841, compared to those from 1806, were they the result of the Liberal Revolution? In order to appropriately respond to these questions, it is necessary to analyse additional years. However, the selection of the years may once again contaminate the

results. The only solution would be to conduct an exhaustive examination, year after year, of all of the data. However, the costs of this research would be very high.
The methodology used in this project, represented on the right hand side of the graph, is quite different from what has been completed to date. Not only it makes it possible to work with more data. It is also a huge step forward from an analytical point of view, given that it allows 1) to formulate questions regarding social change, which do not condition their responses beforehand; and 2) in the likely case that the obtained responses give rise to new questions, to be capable of responding to them without the need to go back to the starting point.

Specific Objective

- Processes of social change previous to the crisis of the Ancien Regime

A new relatively enriched group emerging from the humblest ranges of the rural society

From TREBALLADORS to MENESTRALS

Consolidation of a group of landowners, from peasant origins, which turns into dominant class

From PAGESOS to HISENDATS

In this framework, we have detected two processes of social change previous to the crisis of the Ancien Régime:

a) The emergence of a new relatively enriched group (those who were called *menestrals* since the end of the eighteenth century) from the humblest ranges of the rural society: the so-called *treballadors* (literally, workers).
b) The consolidation of a group of landowners, from peasant origins, which turns into dominant class.

We are going to focus our attention on the former.

Analytic objectives

1. *Treballadors*' role in the land market, credit market, marriage practices…

   - Example 1: formal **credit** market
     - Selecting documents referring to different types of loan (*censal, debitorio, obligación, venta a carta de gracia*,...)
     - Identifying creditors and debtors (both on an individual and social scale)
     - Measuring credit flows

   - Example 2: **marriage** practices
     - Evolution of dowries' value
     - Social homogamy
     - Spatial scope of the marriage market

Analytic objectives

2. Analysis of individual trajectories (prosopographic approach)

   - example: emergence of a new social group (*menestrals*) from the ranks of *treballadors*

     - Identifying the social origins of *menestrals*
     - Studying their living standards and the chronology of the accumulation processes.
     - Analysing their marriage strategies (with whom they got married, dowries' values …)

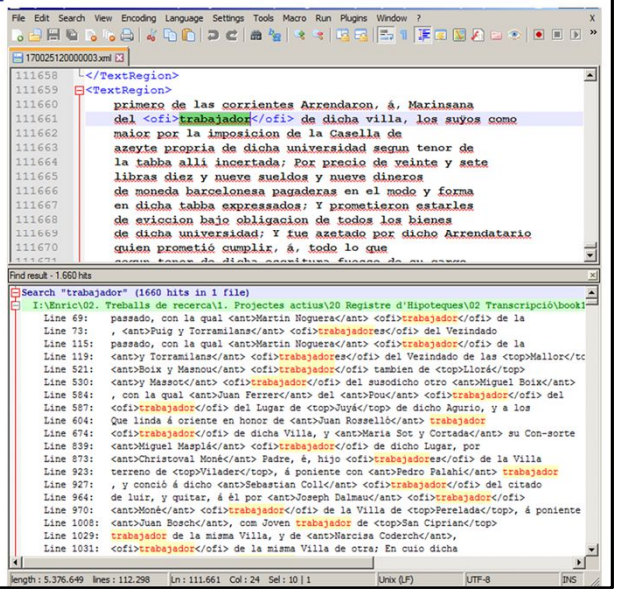
How can we use the transcripts?

Search of the word "trabajador"

Option 1: Exploring the files in plain text

Direct search through **key words**

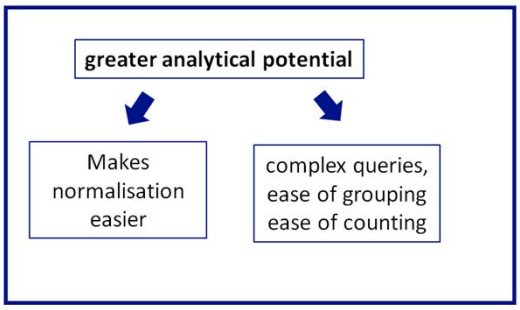
- Simplicity
- Accessible to any user (intern or extern)
- Limited analytic potential

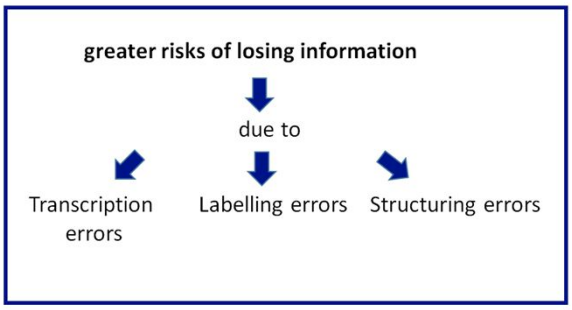


How can we use the transcripts?

Option 2: generation of a structured database → basic structure:

- Notarial reference (notary and office)
- Type of deed
- Date
- People involved (including occupation and place of residence)
- Places
- Value or price of the operation





An example of query into the database : Social labels with more than 100 references

4 books
(June 1768- June 1770)

11.262 documents

occupation	n
labrador	10107
trabajador	6645
jornalero	1757
Negociante	965
presbítero	814
Albañil	688
Marinero	592
Ciudadano	520
tejedor	517
Herrero	478
Sastre	465
Carpintero	461
zapatero	391
Pescador	353
Cirujano	304
Paraire	302
Boticario	285
Droguero	252
Soguero	240
tendero	213
comerciante	172
noble	167
Arriero	161
carretero	156
Molinero	144

→ Priority objectives

In order to test the possibilities of the database obtained using the transcripts for the study of humble rural groups, we are going to show some very preliminary results offered by just 4 books referred to 2 years, from June 1768 to June 1770.
In the first place, it is necessary to stress the fact that those humble groups (those labelled as *trabajador* and *jornalero*) had a strong presence in the Registry of Deeds. They are the second most frequent group, only after farmers.

Treballadors in different types of notarial deeds

Tipology (in Catalan)	Tipology	total	with treballadors	%
venda	Sale / Purchase	2205	1354	61,41
censal	Annuity	1320	548	41,52
capítols matrimonials	Marriage contract	1191	566	47,52
establiment	Emphyteutic contract	1143	725	63,43
debitori	Loan	897	356	39,69
arrendament	Lease	892	248	27,80
acollament	Annuity transmission	778	414	53,21
carta de pagament	Receipt of payment	367	75	20,44
donació	Donation	275	107	38,91
lluïció	Annuity redemption	269	48	17,84
revenda	Resale	258	142	55,04
testament	Will	231	49	21,21
concòrdia	Extrajudicial agreement	214	53	24,77
inventari	Inventory	125	43	34,40
indemnitat	Indemnity	122	61	50,00

* Only types of deeds with more than 100 documents

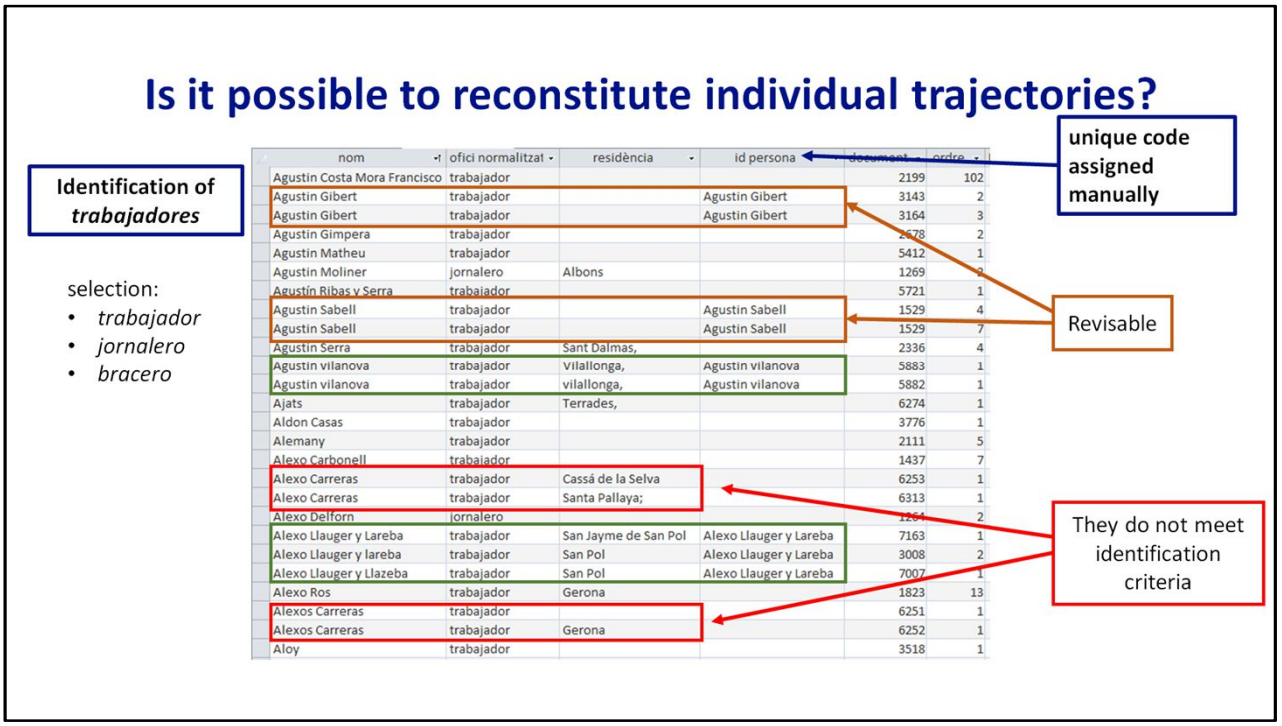
They were especially present in the land market (there were *treballadors* in more than 60% of sale/purchase deeds and emphyteutic contracts), but also in the credit market (arround 40% of annuities and loans with *treballadors*), marriage contracts (47,52%) or inventories (34,4%).

Is it possible to identify *treballadors*' role? The sale/purchase example

Treballadors' role in sale/ purchase deeds	n
possible seller	633
possible purchaser	605
possible owner of adjacent estates	1.210
Not identified	168
TOTAL	**2.616**

A key question is whether it is possible to identify the role of each participant in the deeds. We have tested it in the case of *treballadors'* role in sale/purchase deeds. 2.616 *treballadors* appear in 1.354 sale/purchase deeds. Using the procedure of identification through position respect the key word, we can deduce their role as is shown in the table.

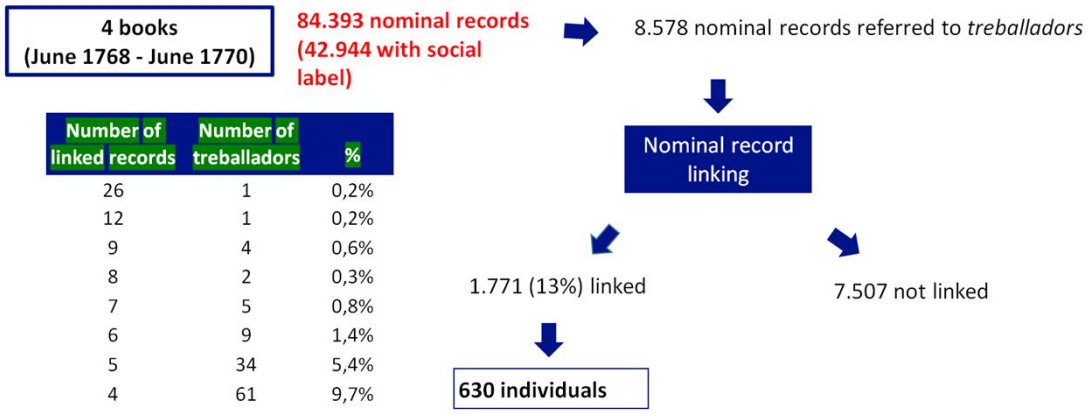
It is worth noting that we find more or less the same number of *treballadors* as sellers and as purchasers, being the first case slightly higher than the second one.



nom	ofici normalitzat	residència	id persona	document	ordre
Agustin Costa Mora Francisco	trabajador			2199	102
Agustin Gibert	trabajador		Agustin Gibert	3143	2
Agustin Gibert	trabajador		Agustin Gibert	3164	3
Agustin Gimpera	trabajador			2678	2
Agustin Matheu	trabajador			5412	1
Agustin Moliner	jornalero	Albons		1269	2
Agustín Ribas y Serra	trabajador			5721	1
Agustin Sabell	trabajador		Agustin Sabell	1529	4
Agustin Sabell	trabajador		Agustin Sabell	1529	7
Agustin Serra	trabajador	Sant Dalmas,		2336	4
Agustin vilanova	trabajador	vilallonga,	Agustin vilanova	5883	1
Agustin vilanova	trabajador	vilallonga,	Agustin vilanova	5882	1
Ajats	trabajador	Terrades,		6274	1
Aldon Casas	trabajador			3776	1
Alemany	trabajador			2111	5
Alexo Carbonell	trabajador			1437	7
Alexo Carreras	trabajador	Cassá de la Selva		6253	1
Alexo Carreras	trabajador	Santa Pallaya;		6313	1
Alexo Delforn	jornalero			1264	2
Alexo Llauger y Lareba	trabajador	San Jayme de San Pol	Alexo Llauger y Lareba	7163	1
Alexo Llauger y lareba	trabajador	San Pol	Alexo Llauger y Lareba	3008	2
Alexo Llauger y Llazeba	trabajador	San Pol	Alexo Llauger y Lareba	7007	1
Alexo Ros	trabajador	Gerona		1823	13
Alexos Carreras	trabajador			6251	1
Alexos Carreras	trabajador	Gerona		6252	1
Aloy	trabajador			3518	1

Another key question is whether it is possible to reconstitute individual trajectories. We use a nominal record linking procedure based on name and surname, social label (considering *trabajador, jornalero* and *bracero* as equivalents) and place of residence. If there is a total coincidence in all these items, we consider that the individuals are the same. When some information, such as place of residence, is missing, we assign a revisable linking. Finally, records do not meet identification criteria when some items are clearly different, as can be seen in the example.

Results of the nominal record linking (*treballadors*)



Number of linked records	Number of treballadors	%
26	1	0,2%
12	1	0,2%
9	4	0,6%
8	2	0,3%
7	5	0,8%
6	9	1,4%
5	34	5,4%
4	61	9,7%
3	152	24,1%
2	361	57,3%
	630	**100 %**

The results of this procedure is that, for the short period of two years, we can link 1.771 registers (13% of those referred to *treballadors*) involving 630 individuals. We consider that this is not a small figure, taking into account that we are analysing just two years.

Conclusions

- We do not have robust results since this is a work in progress. We need a longer period to capture the trends of social change. For example: at this time, we have only found one *menestral*, but we know that this social label is very frequent at the end of the eighteenth century.
- The preliminary results allow us to observe important presence of *treballadors* in the notarial deeds, which not only confirms that the source is suitable to study this social group, but also suggests the dynamic role of *treballadors* in the land market and the credit market.
- The methodology, assuming its imperfection, opens new ways to explore large documentary sources at an affordable cost.