Retos y oportunidades de la transcripción automática de textos manuscritos.

La explotación de los libros del Oficio de Hipotecas

Enric Saguer

Centre de Recerca d'Història Rural
Universitat de Girona

Planteamiento

- Trabajo en colaboración con el PRHLT de la UPV (> 2016)

Pattern Recognition and
Human Language Technology
Research Center

- Experiencia de usuario de las técnicas de transcripción asistida
  - expectativas desde la perspectiva de un historiador
  - resultados obtenidos
  - retos, dificultades y obstáculos

Foco --> extracción y procesado de los topónimos

Interés = obtención de representaciones cartográficas y realización análisis espaciales

El reto de la consulta masiva de series documentales

Existencia de grandes series documentales difíciles de explotar sistemáticamente por razones de extensión:

protocolos notariales
padrones y censos de población
registros parroquiales
catastros
oficio de hipotecas
...

Digitalización series = 1r paso en la accesibilidad

http://arxiusenlinia.cultura.gencat.cat

↑ volumen documental → reto a 3 niveles:

- LECTURA de la documentación
- CONSULTA (localización de datos simples, puntuales, discretos)
- TRATAMIENTO de datos (agregación, obtención de pautas generales,...)

La transcripción automática como respuesta

- posibilidad de **mecanizar una de las tareas** más absorbentes de la investigación histórica

> "Hemos de confesar, previamente, que la investigación en los Archivos de Protocolos no está esmaltada con los inesperados descubrimientos que constituyen el estímulo de los historiadores en los Archivos generales. (...). Muchas veces la labor se convierte en la práctica de una rigurosa contabilidad, en la que prevalece la táctica de retener el más mínimo detalle. (...) Largas hora de trabajo, pues, sin que las salpique el menor aliciente, en una **tarea gris, plúmbea y, al parecer, anodina**..." (Jaume VICENS VIVES (1954) *El gran sindicato remensa (1488-1508),* Madrid, CSIC)

- posibilidad de **gestionar cantidades ingentes de documentación** no asumibles con procedimientos manuales
  → salto cuantitativo con **efectos cualitativos**

> "... a medida que transcurren los días y se acumula el material investigado y las fichas se alinean en sus casilleros respectivos, el historiador se percata de que va pisando terreno firme. (...). Aquí **son los hechos que hablan al investigador; no el investigador quien dispone de los hechos**. Tal es la grandeza y la miseria del método estadístico, que conduce a la verdad condenando a la rutina al que lo utiliza" (Jaume VICENS VIVES (1954) *El gran sindicato remensa (1488-1508),* Madrid, CSIC)

- reto: podemos aprovechar las nuevas oportunidades abiertas por la tecnología de reconocimiento de textos manuscritos para desarrollar **nuevos enfoques?**
  - experiencia fondos hemerográficos (OCR) → han modificado la manera de trabajar? O, simplemente, facilitan los procedimientos habituales?

El primer dilema: opciones de reconocimiento de texto manuscrito

Word spotting

- detección de palabras
- método de indexación probabilística (grado de confianza ajustable)
- permite búsquedas directas sobre imágenes
- no proporciona la transcripción de la imagen
- procedimiento menos costoso y más rápido
  - adecuado para grandes series documentales que sólo han de procesarse mediante búsquedas discretas/simples (similitud Google)
  - permite consultas booleanas
  - requiere entrenamiento del sistema

ejemplo: Carabela (BNE)
teatro manuscrito del Siglo de Oro
40.000 imágenes/páginas

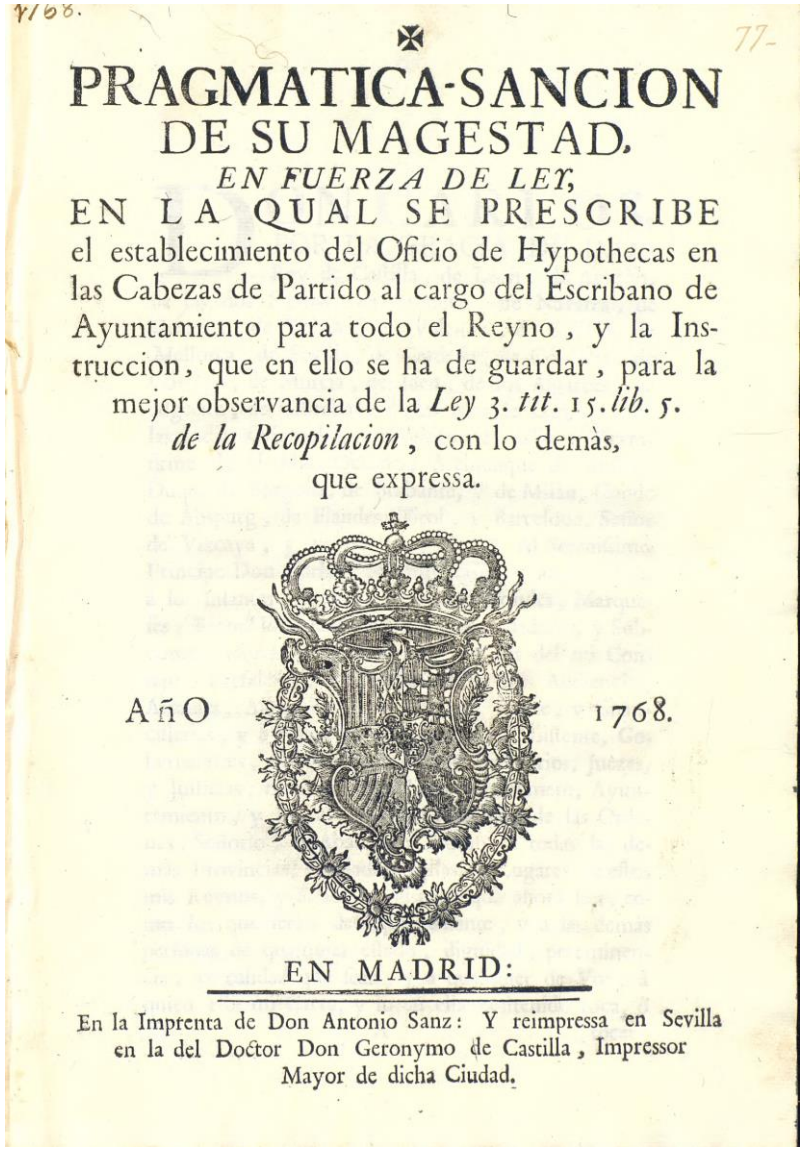
Transcripción completa

- objetivo: hallar la secuencia de palabras más probable para cada imagen de texto manuscrito
- tiene en consideración la cadena de palabras (n-grams) en el proceso de transcripción
- requiere detección y ordenación previa de las regiones de texto (layout analysis)
- procedimiento costoso
  - requiere mayor entrenamiento del sistema
  - especialmente costoso en la fase de detección del layout

El Oficio de Hipotecas

- Institución de publicidad registral
- Creación mediante Real Pragmática 5 febrero 1768
- Cambios en la denominación:

  1768 – Oficio de Hipotecas
  1829 – Contaduría de Hipotecas
  1845 – Registro de Hipotecas



1768.

77-

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.

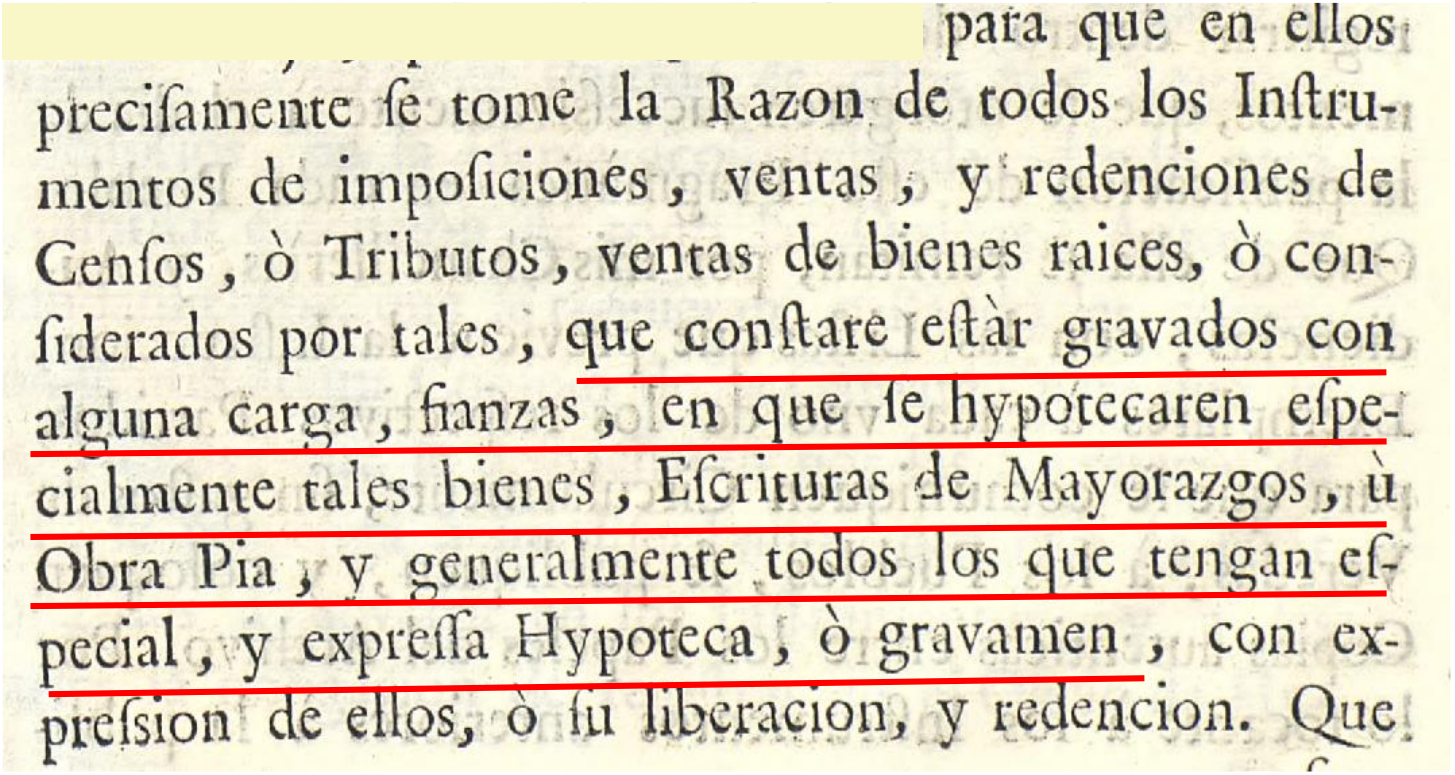
EN MADRID:

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

Contenido del Oficio de Hipotecas de Girona

- contratos con cargas e hipotecas sobre bienes
- criterio de aplicación no homogéneo
- Cataluña = inclusión de un amplio número de actas notariales (garantías hipotecarias generales)

↑↑↑ colaboración notarios



para que en ellos
precisamente se tome la Razon de todos los Instru-
mentos de imposiciones, ventas, y redenciones de
Censos, ò Tributos, ventas de bienes raices, ò con-
siderados por tales, que constare estàr gravados con
alguna carga, fianzas, en que se hypotecaren espe-
cialmente tales bienes, Escrituras de Mayorazgos, ù
Obra Pia, y generalmente todos los que tengan es-
pecial, y expressa Hypoteca, ò gravamen, con ex-
presion de ellos, ò su liberacion, y redencion. Que

Contenido del Oficio de Hipotecas de Girona, 1768-1784

Transferencia (mercantil) de inmuebles y derechos	n	%	% (grupo)
ventas	12.611	16,4%	27,1%
establecimientos emf.	4.660	6,1%	
reventas	1.798	2,3%	
donaciones	1.739	2,3%	
Instrumentos de crédito y garantía			
censos consignativos (censales)	9.464	12,3%	32,7%
debitorios	5.175	6,7%	
ventas a carta de gracia	1.997	2,6%	
encargamientos	4.522	5,9%	
luiciones	2.737	3,6%	
indemnidad	1.187	1,5%	
Régimen hereditario			
capitulos matrimoniales	8.244	10,7%	14,6%
heredamientos	2.212	2,9%	
inventarios de bienes	728	0,9%	
Otros			
arrendamientos	6.128	8,0%	25,6%
cartas de pago	4.064	5,3%	
concordias	1.303	1,7%	
cabrevaciones	775	1,0%	
promesas	723	0,9%	
fundaciones	665	0,9%	
otros	5.975	7,8%	
	76.707	**100%**	**100%**

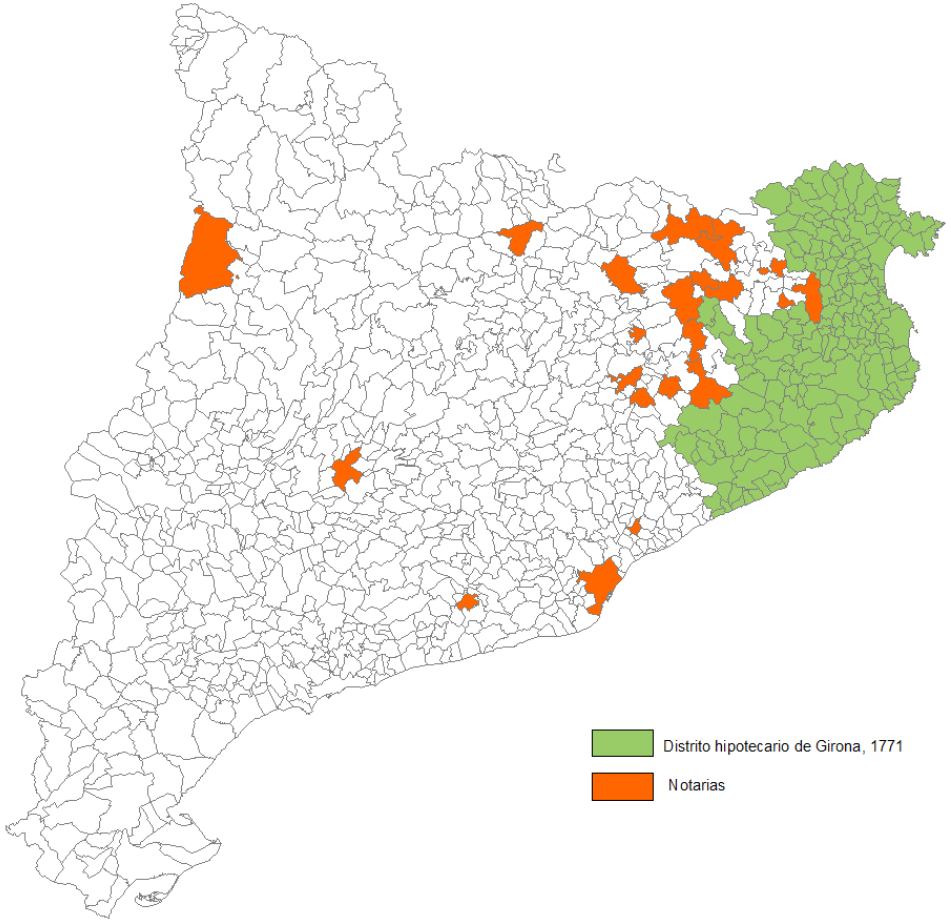
Contenido del Oficio de Hipotecas de Girona

- amplia extensión territorial
  - 1768-1773 = 3.884 km^2

- Obligación de presentar la escritura en la oficina donde se hallan los bienes afectados

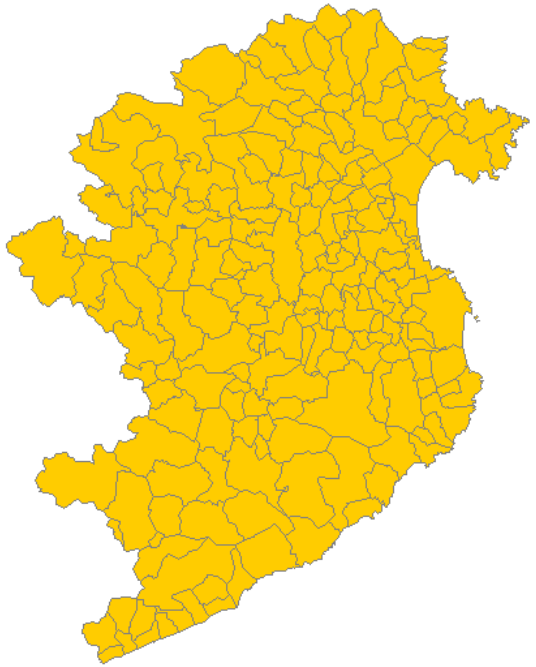
Inclusión de todas las escrituras referidas a fincas del territorio de cada oficina

Notarías de procedencia de las escrituras registradas en el Oficio de Hipotecas de Girona, 1771



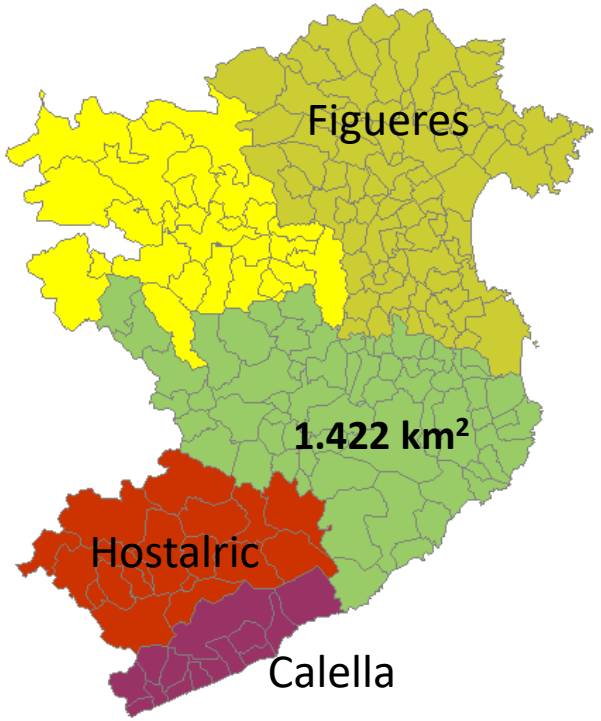
Modificaciones territoriales

- Pragmática 1768 → 1 oficina / cabeza de partido
  - Inexistencia de Partidos en Cataluña → Corregimientos (7)
  - Traslación Pragmática al Corregimiento de Girona: 2 oficinas (Girona + Besalú → Alcaldía Mayor)
  - 1774 → Segregación de los Oficios de Figueres i de Hostalric (1780 Calella)



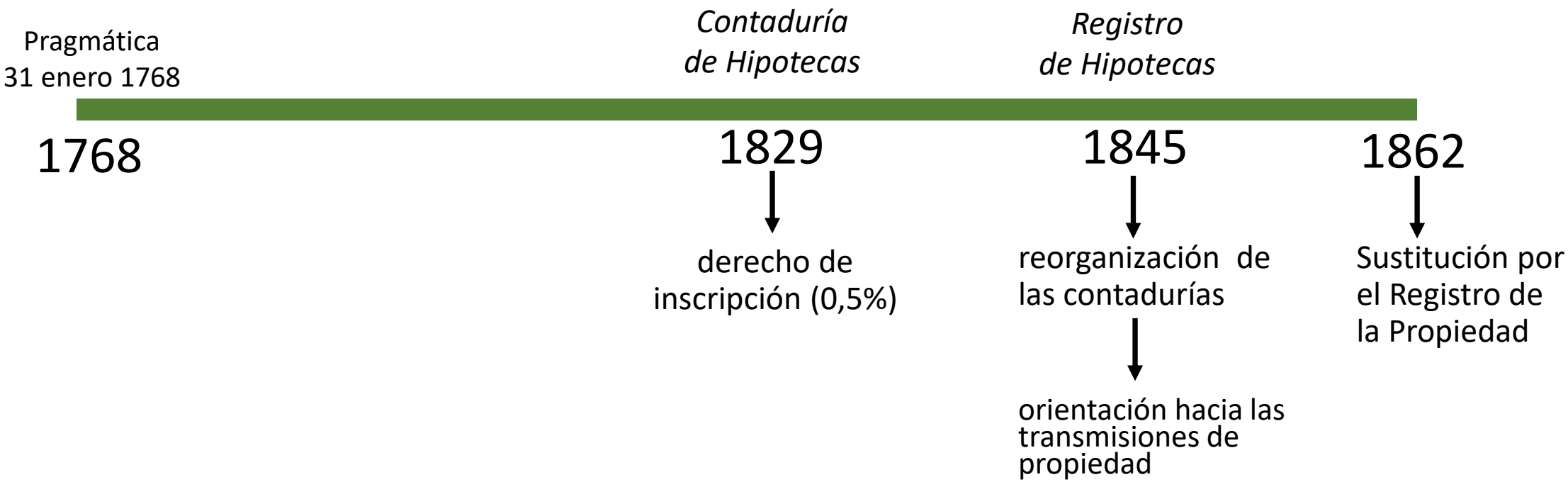
Corregimiento

Distritos hipotecarios 1768



Distritos hipotecarios 1780

Transformaciones importantes en el Oficio de Hipotecas



Objetivo general

Estudio de los procesos de cambio social desde mediados del siglo XVIII hasta mediados del siglo XIX

procesos de cambio social previos a la crisis del antiguo régimen

Emergencia de un grupo de trabajadores ‘relativamente’ enriquecidos

MENESTRALS

Consolidación de un grupo de propietarios de masos, con orígenes campesinos, que se convierte en clase dominante

HISENDATS

Objetivos analíticos

1. Análisis de trayectorias individuales (enfoque prosopográfico)

   - ejemplo 1: Hipótesis sobre la pérdida de cualificación (▽ skill premium, ▽ cuota catastral) y posible descenso social de los **albañiles**
     - identificar y aislar a los albañiles
     - seguir sus pautas de movilidad intergeneracional (capítulos matrimoniales)
     - seguir sus movimientos en el mercado inmobiliario y del crédito

   - ejemplo 2: Hipótesis sobre la acumulación patrimonial de los **grandes arrendatarios** (**masovers**)
     - identificarlos y aislarlos
     - reconstruir su actividad en el mercado inmobiliario
     - analizar su movilidad

Objetivos analíticos

2. Análisis del funcionamiento de los **mercados** (de la tierra, del crédito, matrimonial,...)

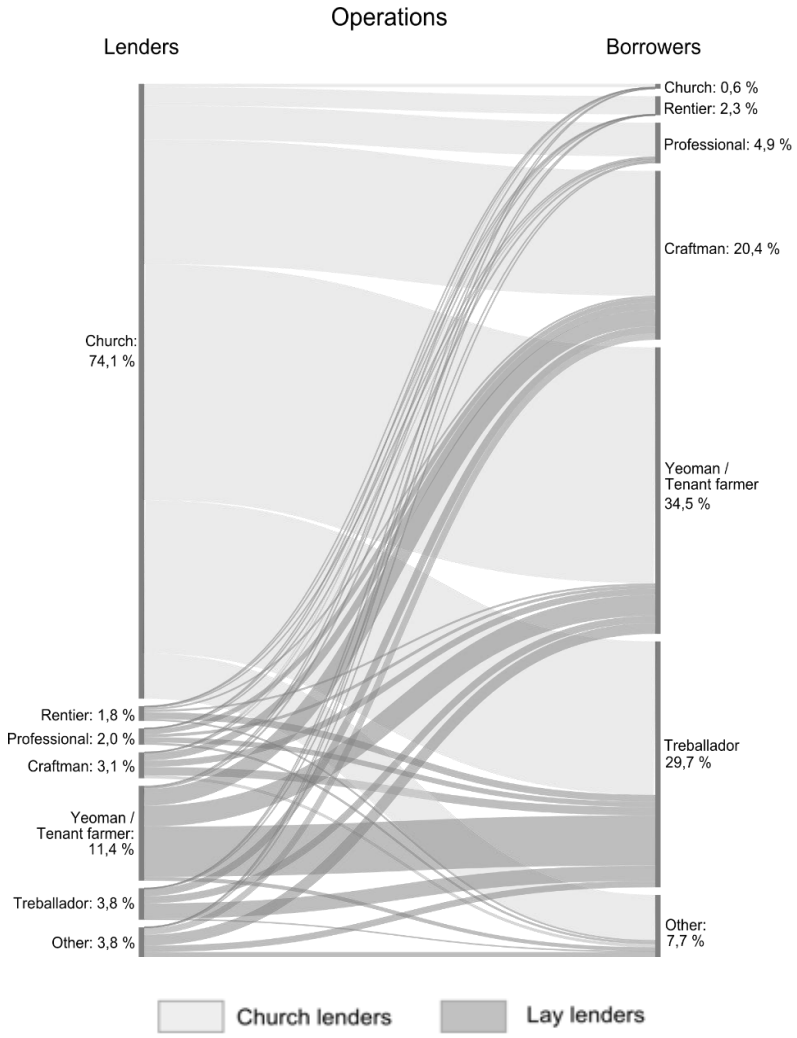
ejemplo: el mercado formal del **crédito**

- aislar las tipologías documentales referidas a fórmulas de préstamo (censal, debitorio, obligación, venta a carta de gracia,...)
- identificar a acreedores y deudores (a escala individual y social)
- medir los flujos de crédito: volumen, geografía,...

Deeds involving credit operations (1768-73)

Type of deed	Number of deeds (A)	per cent of A	Number of deeds of *treballadors* as borrowers (B)	B/A x 100	per cent of B
Annuities	3,580	41.1	1,061	29.6	39.0
Obligations	2,085	24.0	605	29.0	22.2
Transferred annuities	1,881	21.6	687	36.5	25.3
Sales *a carta de gràcia*	1,154	13.3	367	31.8	13.5
Total	8,700	100	2.720	31.3	100

Socio-professional status of lenders and borrowers in annuities (1768-73)



Congost, Garcia-Orallo & Saguer (2022) “Seeing Credit and Property Rights from Below. The Experience of Catalan Smallholders in the Eighteenth Century”, en prensa

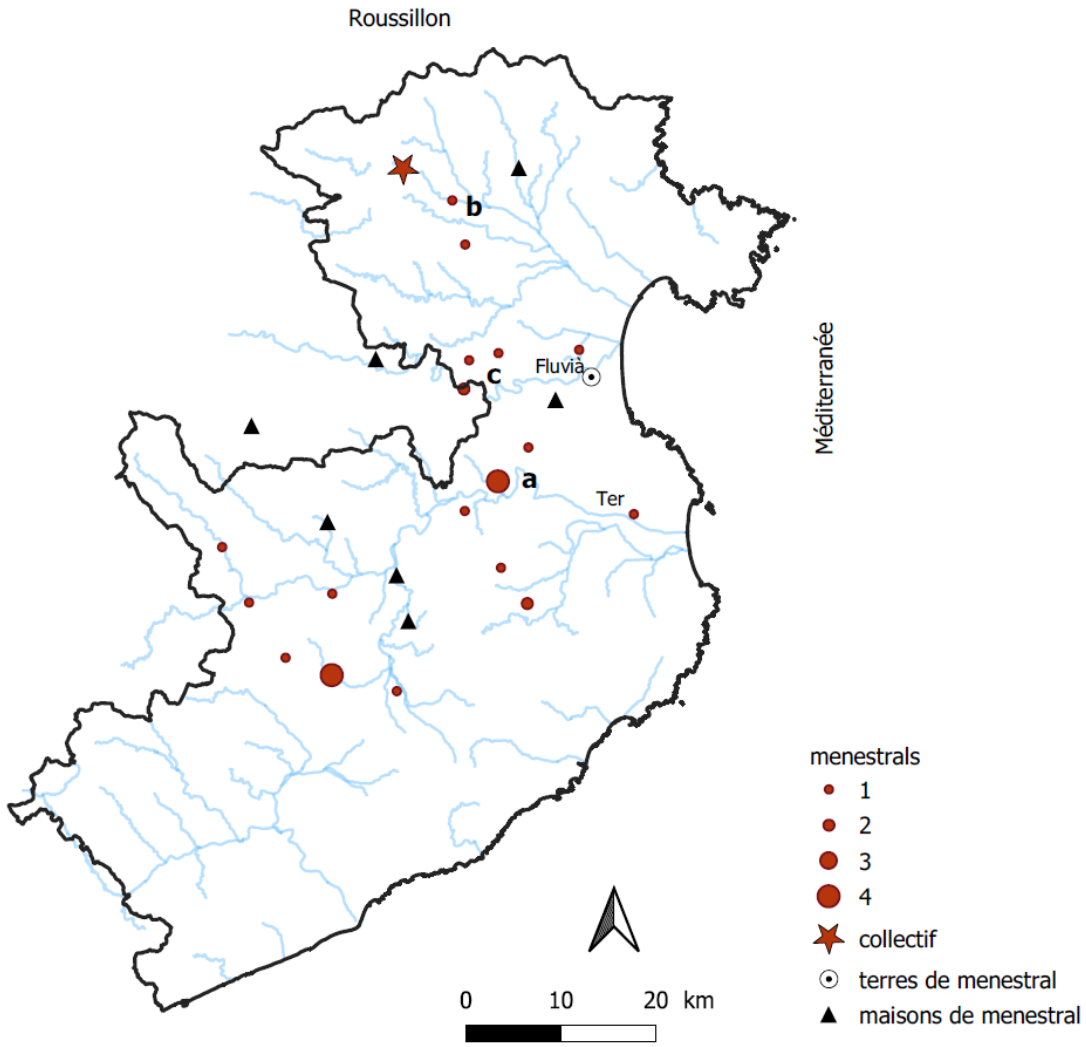
Objetivos analíticos

3. Análisis de **procesos colectivos** de cambio social

ejemplo: Hipótesis sobre la **emergencia de un nuevo grupo social** (*menestrals*) surgido de las filas de los *treballadors*

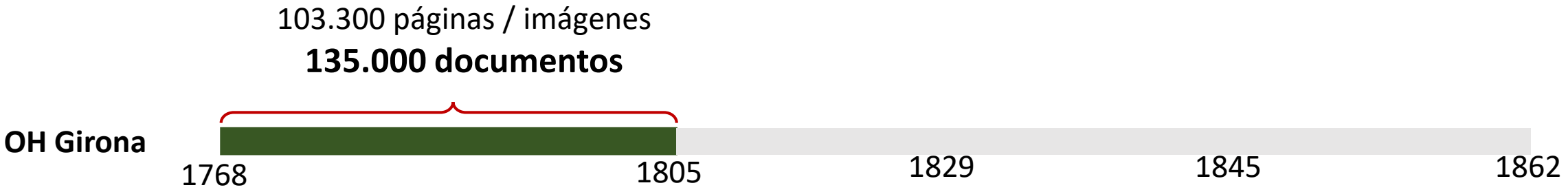
- identificar los orígenes sociales de los que a finales del s. XVIII, se autoidentifican como *menestrals*
- delimitar sus condiciones de vida material (mercado inmobiliario, inventarios,...) y cronología de los procesos de acumulación
- analizar sus estrategias matrimoniales (con quién se casan, importe dote recibida)
- ...

Localisation des références aux *menestrals*, 1768-1774



Congost & Saguer (2022) « Sur les traces des petites gens de la Catalogne rurale (XVIIIe siècle) », en prensa

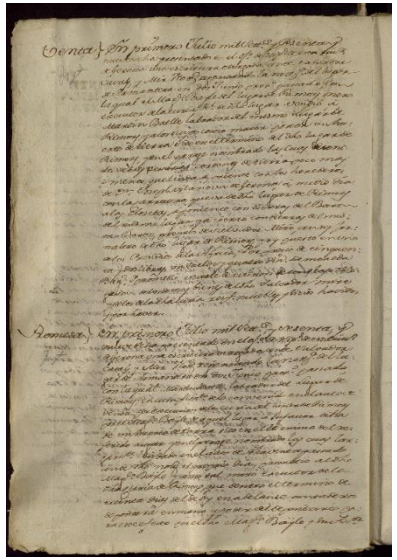
Objetivo operativo a medio plazo: OH Girona, 1768-1805



Por qué hasta 1805?

- relativa homogeneidad geográfica (excepto 1768-1774)
  >1806 - recuperación libros OH de Figueres
- mantenimiento de una única tipología libros (registros generales)
- relativa continuidad en el tipo de letra manuscrita (≈)
- estimación vaciado manual: 25 años x 6 estudiantes x 15 horas/semana

Etapas del proceso



Fase 0
Digitalización

103.300 imágenes

detección de layout
+
transcripción automática

Texto con geometria
(xml)

Fase 1
Transcripción
PRHLT & CRHR

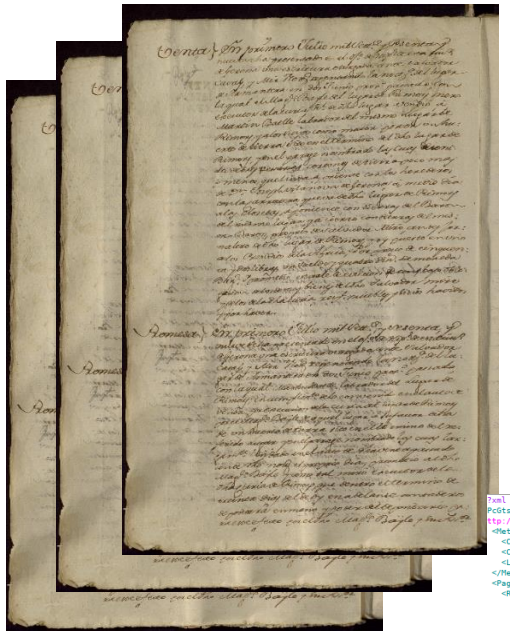
Estructuración en una base de datos

Fase 2

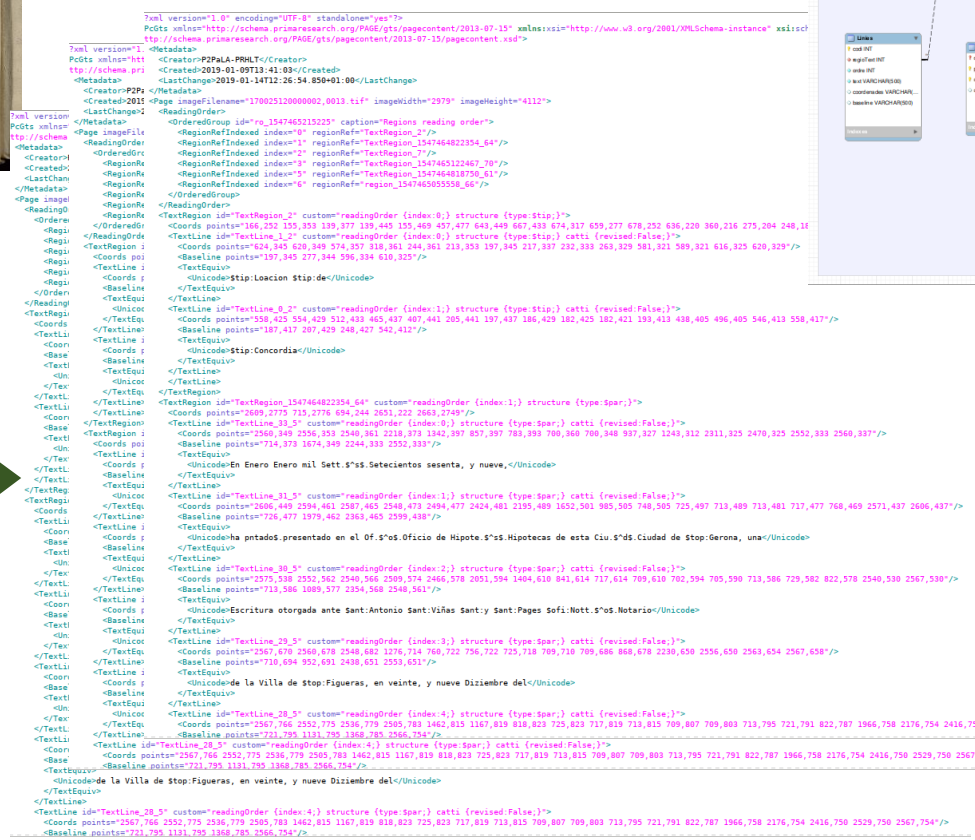
Análisis

Fase 3

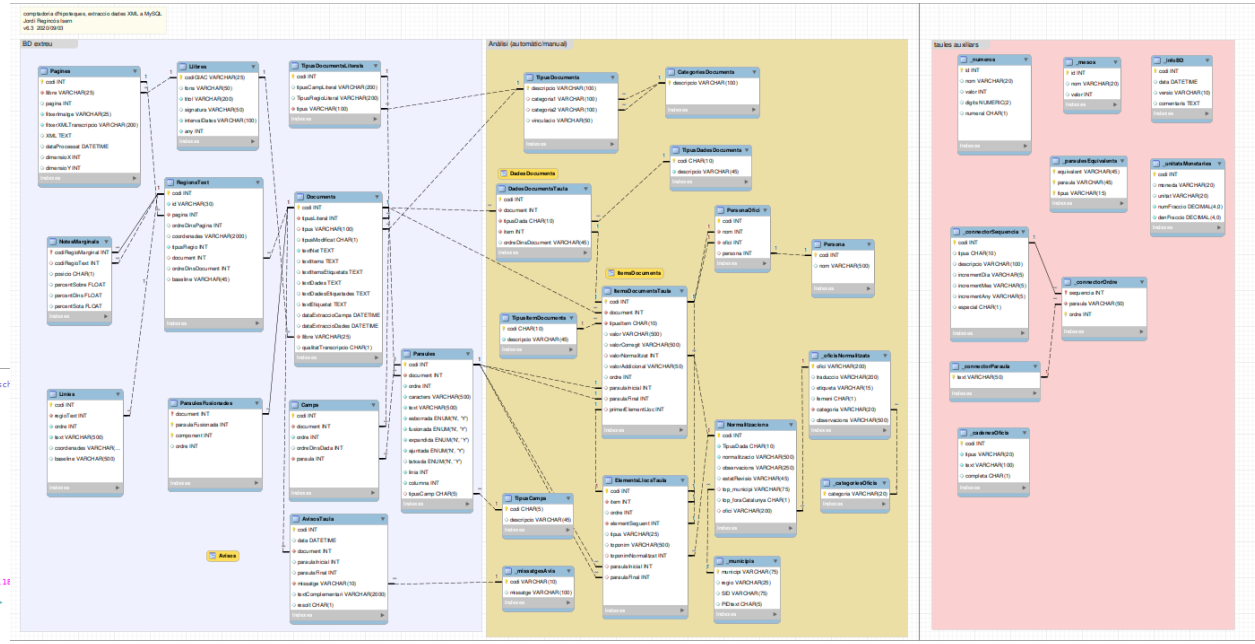
Etapas del proceso: resultados



Fase 1



Fase 2



Fase 1 (resultado)

archivo PAGE

```
<?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 2560,170 2563,171 2631,200 2644,148"/>
        <Baseline points="2500,177 2561,157 2623,183 2632,146"/>
        <TextEquiv>
          <Unicode>$pag:54</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
      <Coords points="755,157 1718,145 2107,153 2662,204 2630,1008 688,988"/>
      <TextLine id="TextLine_7_23" custom="readingOrder {index:0;} structure {type:$pac;}">
        <Coords points="2641,200 2371,194 1756,174 789,188 784,238 1753,224 2369,244 2640,250"/>
        <Baseline points="796,227 1754,212 2370,232 2641,238"/>
        <TextEquiv>
          <Unicode>$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>
```

Fase 1 (resultado)

archivo PAGE

```
<?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>
```

geometría

```
<TextLine id="TextLine_0_1" custom="readingOrder {index:0;} structure {type:$pag;}">
  <Coords points="2631,145 2562,117 2488,141 2503,188 2560,170 2563,171 2631,200 2644,148"/>
  <Baseline points="2500,177 2561,157 2623,183 2632,146"/>
```

```
        <Unicode>$pag:54</Unicode>
      </TextEquiv>
    </TextLine>
  </TextRegion>
  <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
    <Coords points="755,157 1718,145 2107,153 2662,204 2630,1008 688,988"/>
    <TextLine id="TextLine_7_23" custom="readingOrder {index:0;} structure {type:$pac;}">
      <Coords points="2641,200 2371,194 1756,174 789,188 784,238 1753,224 2369,244 2640,250"/>
      <Baseline points="796,227 1754,212 2370,232 2641,238"/>
      <TextEquiv>
        <Unicode>$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>
```

Fase 1 (resultado)

archivo PAGE

```
<?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 2560,170 2563,171 2631,200 2644,148"/>
        <Baseline points="2500,177 2561,157 2623,183 2632,146"/>
        <TextEquiv>
          <Unicode>$pag:54</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
      <Coords points="755,157 1718,145 2107,153 2662,204 2630,1008 688,988"/>
```

transcripción (por líneas)

```
<TextEquiv>
  <Unicode>$ofi:Marineros de dha$.dicha Villa; Por precio de cinco cientos</Unicode>
</TextEquiv>
```

```
      <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>
```

Fase 1 (resultado)

archivo PAGE

```
<?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 2560,170 2563,171 2631,200 2644,148"/>
        <Baseline points="2500,177 2561,157 2623,183 2632,146"/>
        <TextEquiv>
          <Unicode>$pag:54</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
      <Coords points="755,157 1718,145 2107,153 2662,204 2630,1008 688,988"/>
      <TextLine id="TextLine_7_23" custom="readingOrder {index:0;} structure {type:$pag;}">
        <Coords points="2641,200 2371,194 1756,174 789,188 784,238 1753,224 2369,244 2640,250"/>
        <Baseline points="...232 2641,238"/>
        <TextEquiv>
          <Unicode>$ofi:Marineros 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>
```

etiquetas

Fase 1 (resultado)

archivo PAGE

```
<?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 2560,170 2563,171 2631,200 2644,148"/>
        <Baseline points="2500,177 2561,157 2623,183 2632,146"/>
        <TextEquiv>
          <Unicode>$pag:54</Unicode>
        </TextEquiv>
      </TextLine>
    </TextRegion>
    <TextRegion id="TextRegion_1554290914948_2926" custom="readingOrder {index:1;} structure {type:$pac;}">
      <Coords points="755,157 1718,145 2107,153 2662,204 2630,1008 688,988"/>
      <TextLine id="TextLine_7_23" custom="readingOrder {index:0;} structure {type:$pac;}">
        <Coords points="2641,200 2371,194 1756,174 789,188 784,238 1753,224 2369,244 2640,250"/>
        <Baseline points="796,227 1754,212 2370,232 2641,238"/>
        <TextEquiv>
          <Unicode>$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="2584,28                                                                                                    ,338"/>
        <Baseline points="815,3
        <TextEquiv>
          <Unicode>noventa y un
        </TextEquiv>
```

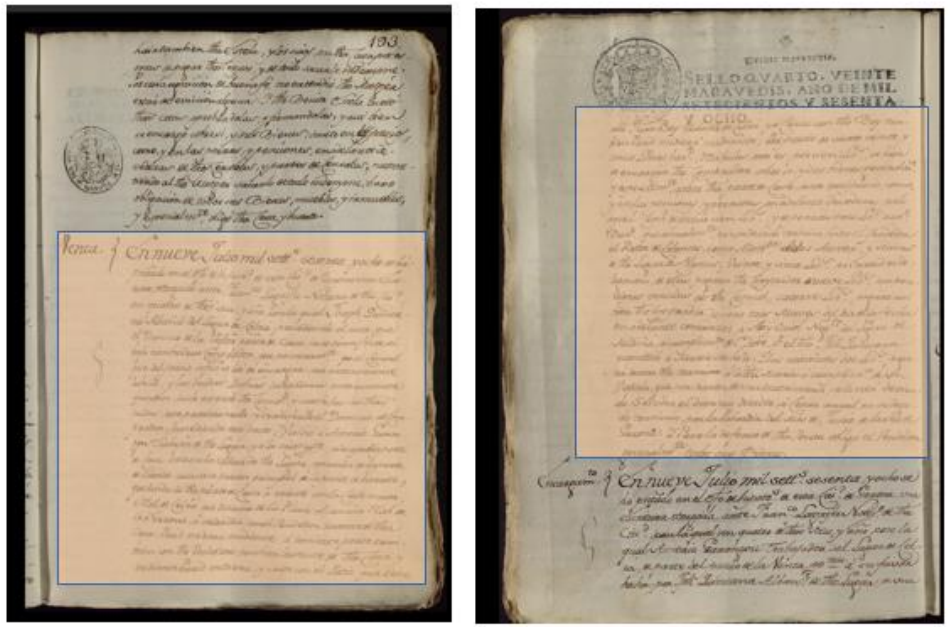
libras bar.^{s}.barcelonesa,

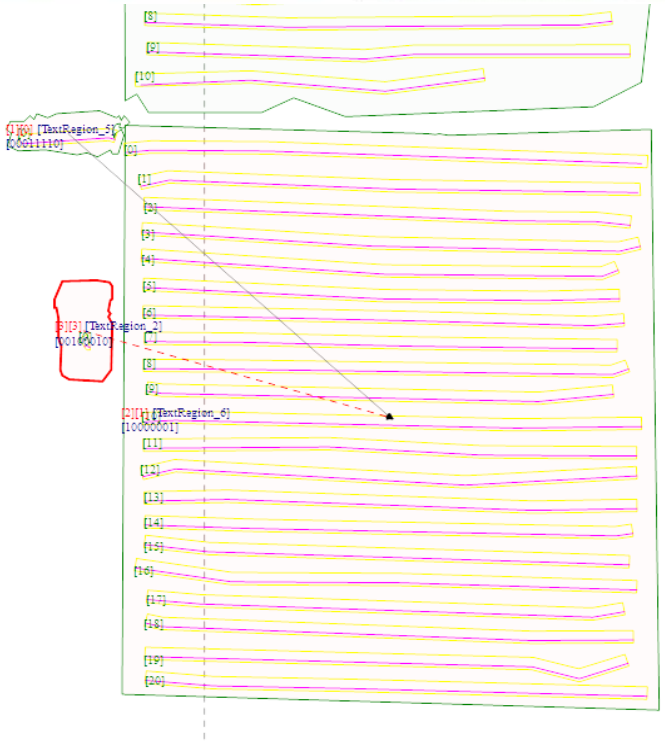
despliguegue abreviaturas

Fase 2 (detalle)

1. Reconstrucción de los asientos

- unificación áreas de texto correspondientes al mismo asiento
- vinculación de las áreas marginales al asiento



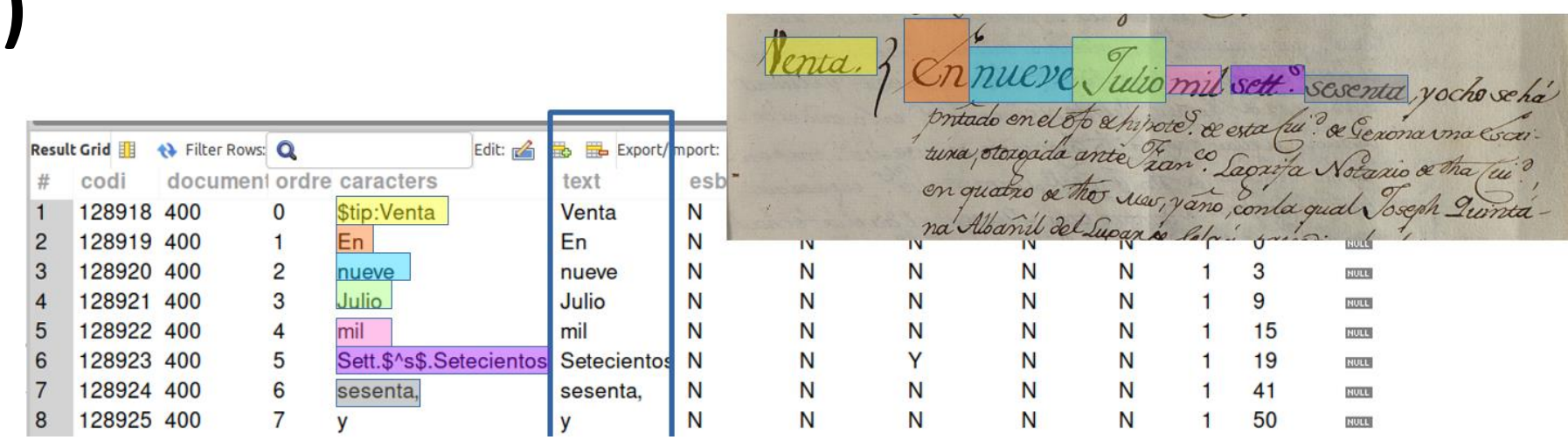


Fase 2 (detalle)

1. Reconstrucción de los asientos

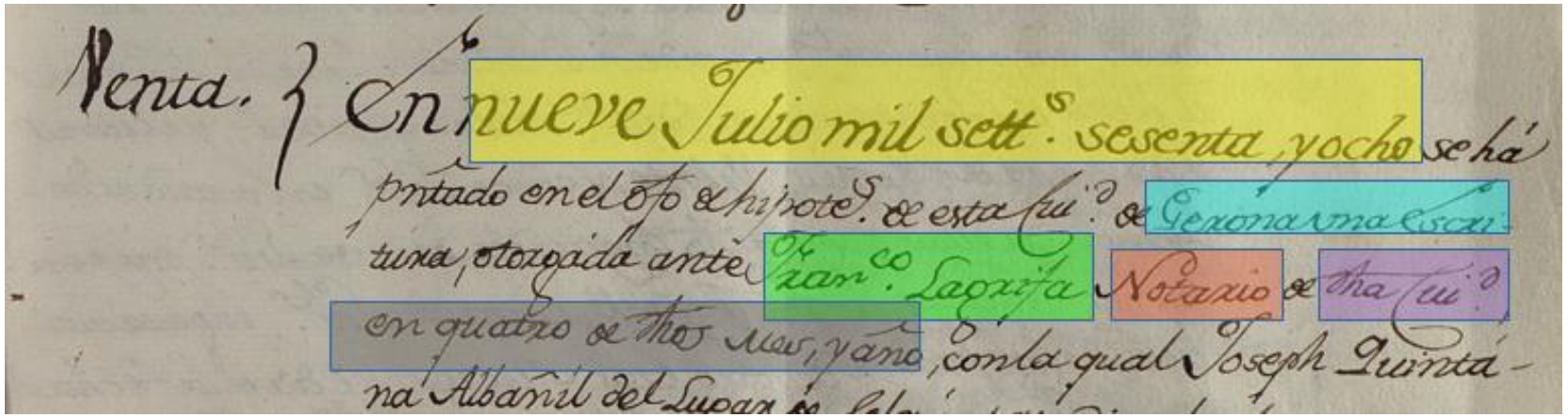
2. Generación tabla de PALABRAS

- unificación de particiones



Fase 2 (detalle)

1. Reconstrucción de los asientos
2. Generación tabla de PALABRAS
3. Generación tabla de ÍTEMS

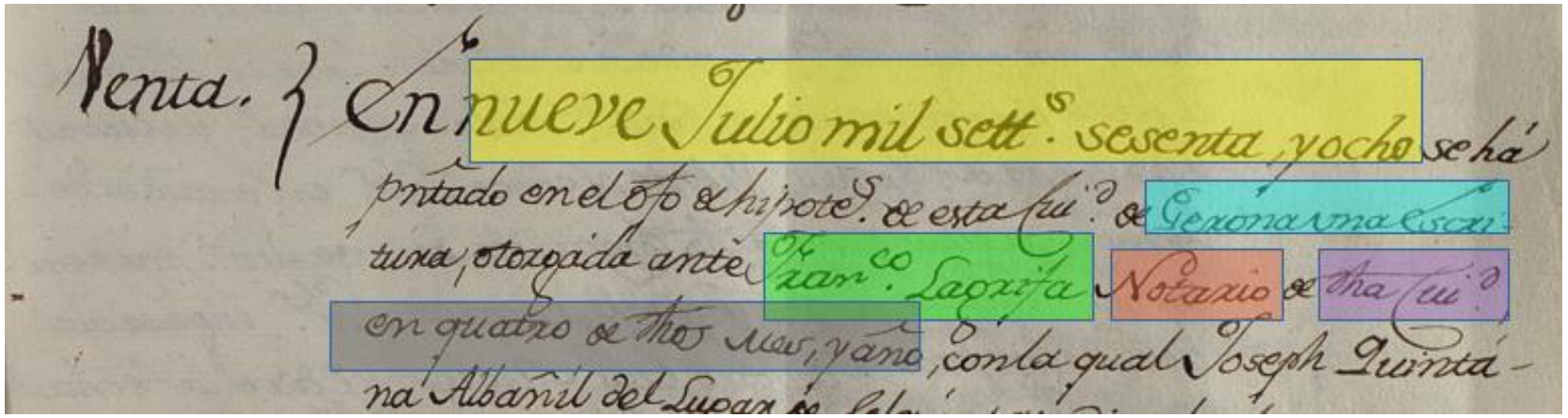


#	codi	documen	tipusItem	valor	valorAddiciona	ordre	paraulaInicia	paraulaFinal
1	1098	400	dataAbs	1768/07/09	NULL	0	2	8
2	655656	400	llocAbs	gerona		0	21	21
3	162639	400	antroponim	francisco lagrifa	NULL	0	28	28
4	1033119	400	ofici	notario		0	31	31
5	1559156	400	llocRel	gerona		0	32	33
6	1099	400	dataRel	1768/07/04	0/0/-|0	0	36	41
7	162640	400	antroponim	joseph quinta	NULL	0	45	45
8	1033120	400	ofici	albanil		0	49	49
9	655657	400	llocAbs	celra		0	53	53

- solo **palabras escogidas**: fechas, valores, antropónimos, topónimos, oficios, fórmula de tratamiento
- reunión de **términos compuestos** (antropónimos, topónimos,...)
- **conversión** de:
  - valores en cifras
  - fechas a formato aaaa/mm/dd
  - referencias relativas de fechas a formato aaaa/mm/dd
  - referencias relativas de topónimos (en dicho lugar)

Fase 2 (detalle)

1. Reconstrucción de los asientos
2. Generación tabla de PALABRAS
3. Generación tabla de ÍTEMS
3. Generación tabla de DATOS



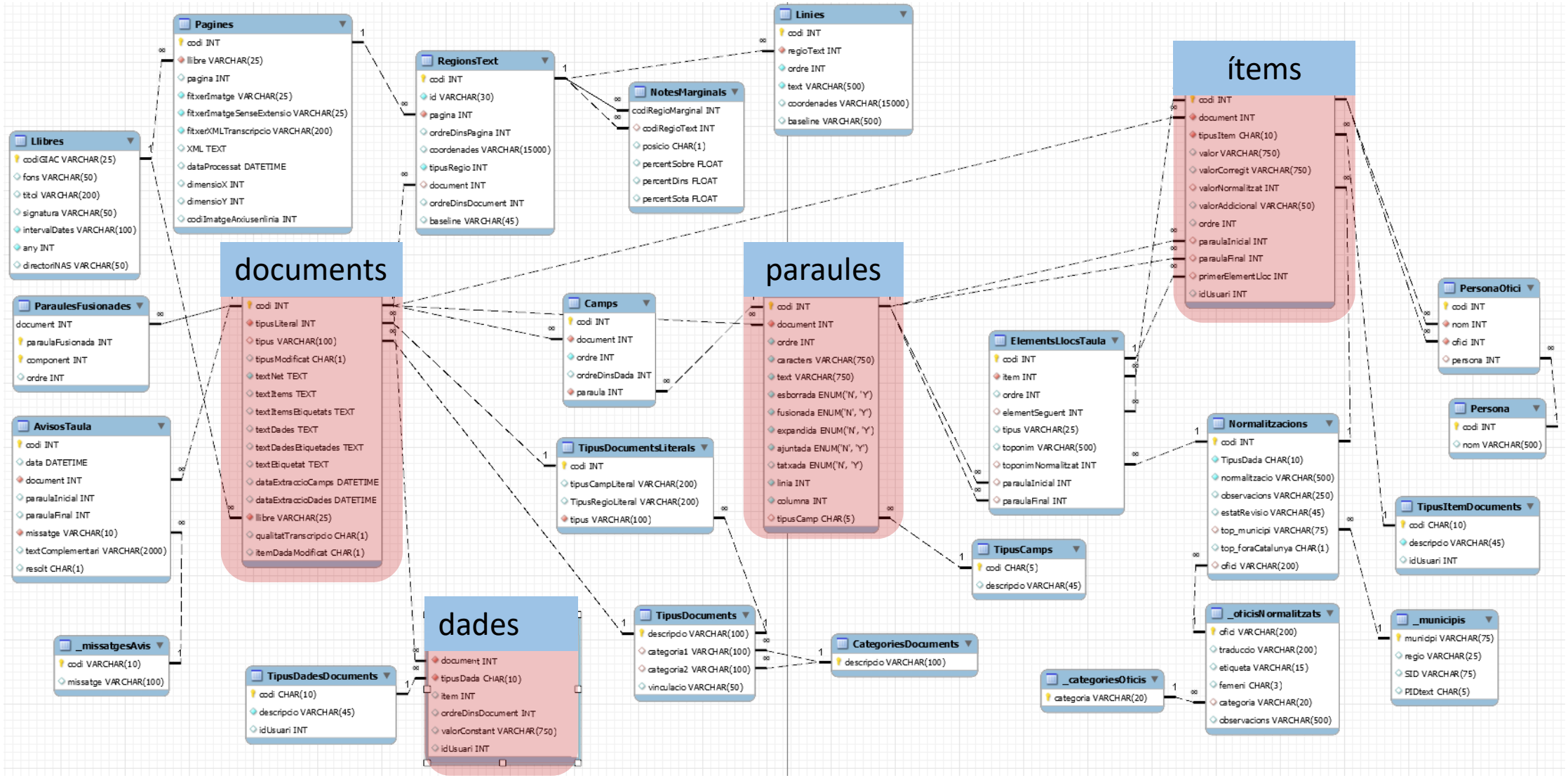
#	codi	documenI	tipusDada	valor	paraulaInicia	paraulaFinaI	ordreI
1	1	400	dataReg	1768/07/09	2	8	1
2	2	400	ciutatReg	gerona	21	21	1
3	3	400	notari	francisco lagrifa	28	28	1
4	4	400	ofiNotari	notario	31	31	1
5	5	400	notaria	gerona	32	33	1
6	6	400	dataActa	1768/07/04	36	41	1

- Determinación de la FUNCIÓN realitzada por cada ítem dentro del documento

  A partir de criterios posicionales y expresiones clave
  Crucial = formato variablemente estructurado de los asientos

- Estado actual = incompleto / pendiente de desarrollo para todas las tipologías documentales

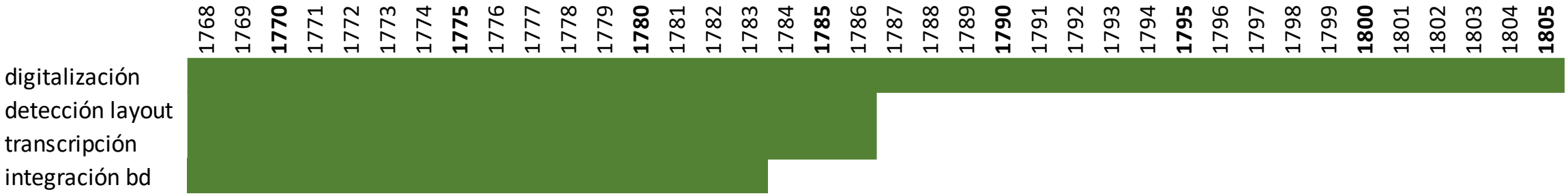
Fase 2: estructura de la base de datos



Acceso web a la base de datos (PHP)

- prototipo de uso interno
- apertura prevista a medio plazo
- permite consultas simples
- devuelve el asiento entero
- conecta con la imagen (servidor Red de archivos de Catalunya)
- permite corrección individual de los ítems

Estado actual



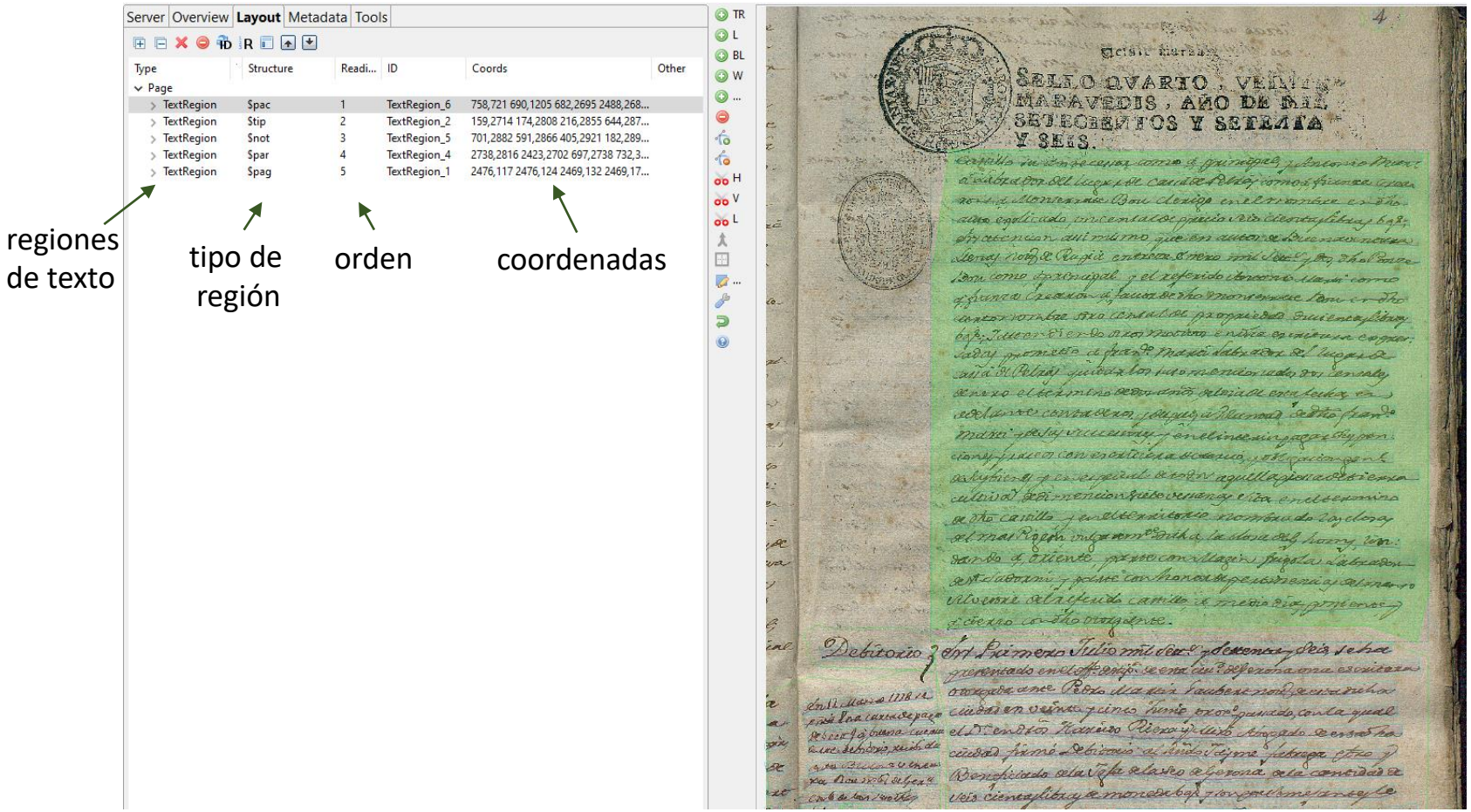
Contenido de la base de datos

	n
libros	34
Imágenes	60.838
documentos	78.294
palabras	24.884.194
ítems	2.028.082
datos	530.352

% imágenes transcritas: 67%
% imágenes integradas BD: 59%

Dificultades: el coste de revisión

- Opción por una transcripción completa (# word spotting)
  - mayor exigència de revisión en la fase de detección del layout



Dificultades: el coste de revisión

Tareas de revisión del layout en los últimos libros

libro	imágenes (n)	horas (h)	n/h
36	1.575	64,3	24,5
37	1.122	48,4	38,8
38	1.122	48,4	23,2
39	1.520	53	28,7
40	1.131	29	39,0
	6.470	243,1	26,6

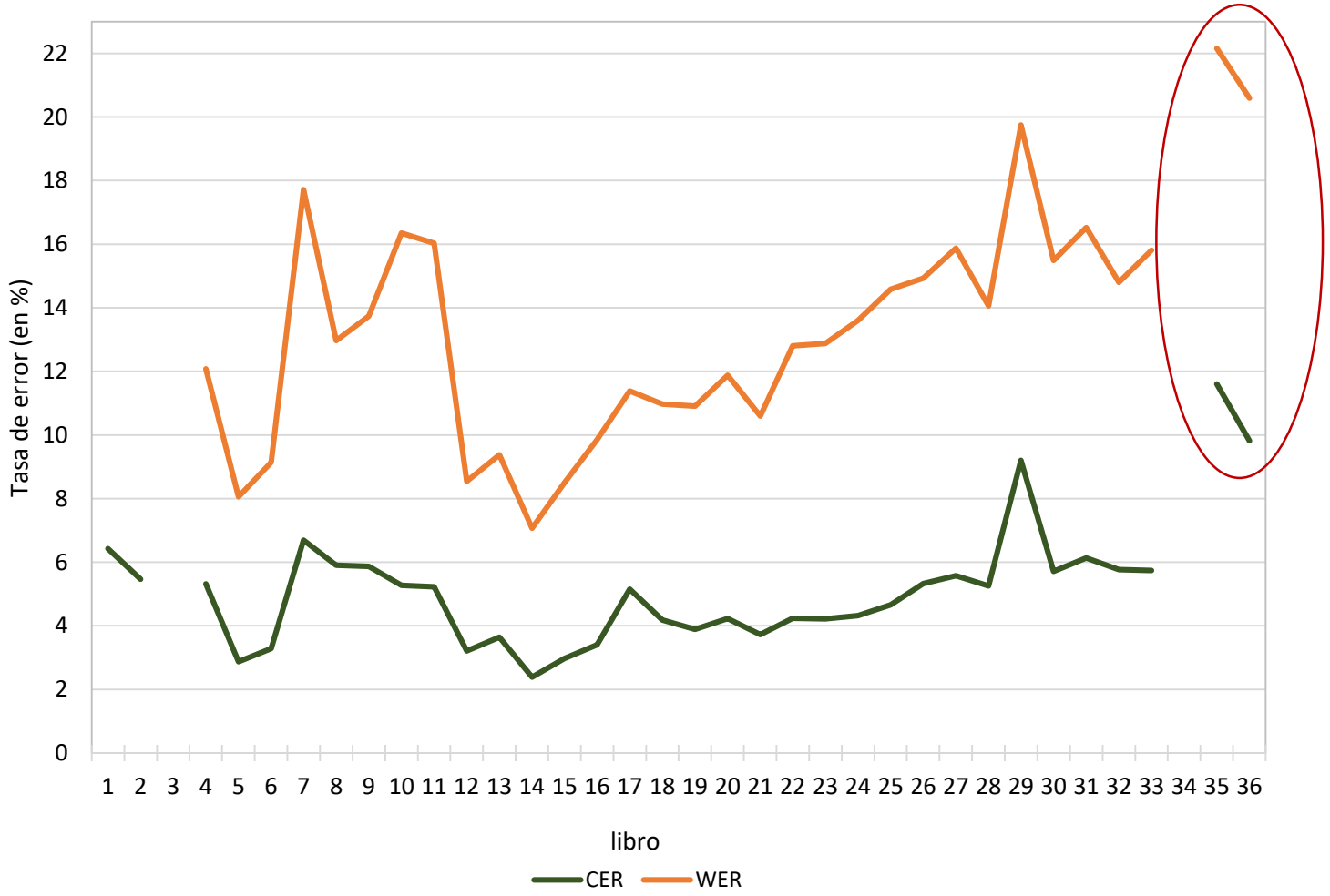
Coste estimado para el conjunto del proyecto

- Imágenes: 103.300
- velocidad media de revisión: 26,6 imágenes/hora
- Estimación total horas necesarias ≈ 3.900

Otros costes:

- entrenamiento del sistema (y eventual reentrenamiento)
- control de calidad (25 imágenes/libro)

Calidad de la transcripción



Métricas genéricas

CER: Tasa de error en carácteres
WER: Tasa de error en palabras

no indican la calidad de la transcripción en términos específicos (fechas, topónimos, valores numéricos, antropónimos,...)

Calidad en los términos críticos

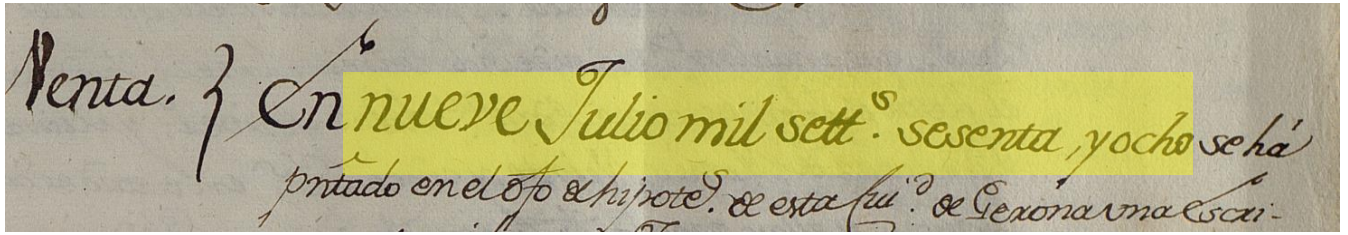
Ensayo realizado sobre una muestra de documentos referidos al sector de la construcción (albañiles, carpinteros y ladrilleros)

- 6.246 documentos
- revisión exhaustiva y determinación del tipo de error en:
  - fecha de registro
  - fecha de escritura del documento notarial
  - antropónimo
  - lugar de residencia (topónimo)
  - categoria sociolaboral (sesgo → criterio de selección)

Calidad en los términos críticos: fecha de registro

Errores en la fecha de registro

tipo de error		n
ERRORES DE TRANSCRIPCIÓN		
	A. Transcripción incorrecta	
	error transcripción mes	48
	error transcripción año	29
	error transcripción general	34
	B. Transcripció incompleta	
	falta dia	1
	falta mes	9
	falta año	42
ERRORES DE CONVERSIÓN		
	error de conversión dia	2
	error de conversión mes	5
	error de conversión año	2
	error de conversión general	5
	Total errores	177
	Total registros	6246
	% error	2,8%



- localización constante en la primera línea de cada asiento
- tasa de error muy baja
- fácilmente enmendable por la estricta ordenación de los asientos

no atribuibles a la transcripción

conversión del formato literal a formato aaaa/mm/dd

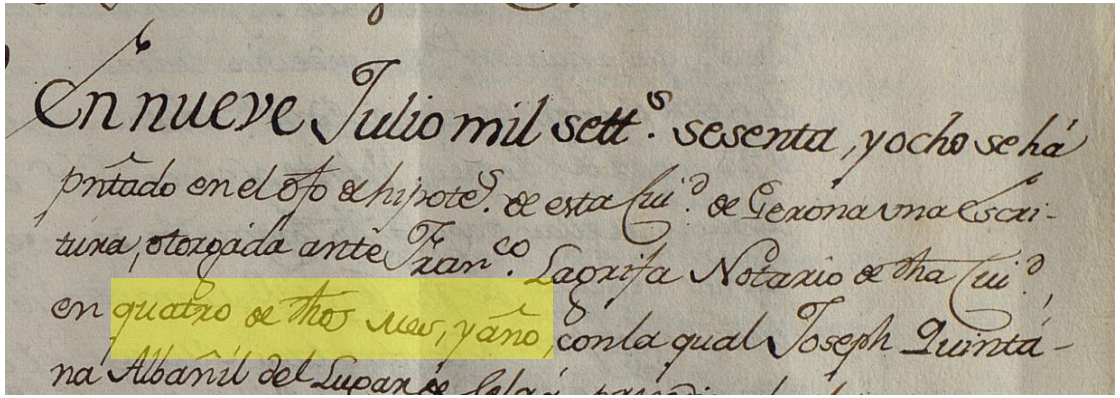
#	codi	documen	tipusDada	valor	paraulaInicia	paraulaFina
1	1	400	dataReg	1768/07/09	2	8
2	2	400	ciutatReg	gerona	21	21
3	3	400	notari	francisco lagrifa	28	28
4	4	400	ofiNotari	notario	31	31
5	5	400	notaria	gerona	32	33
6	6	400	dataActa	1768/07/04	36	41

Calidad en los términos críticos: fecha notarial

	tipo de error	
error transcripción	sin error de transcripción	5858
	error transcripció	1
	error transcripción año	23
	error transcripción mes	152
	error transcripción dia	13
	error transcripción general	35
	no detectada	5
	no transcrita	9
	transcripción incompleta falta año	6
	transcripción incompleta falta mes	50
	transcripción incompleta falta dia	7
error conversión	error conversión general	17
	error de conversión año	55
	error de conversión dia	1
	error de conversión mes	14
	suma	6246
	% sin error	92,7%
	% con error de transcripción o conversión	6,2%

- tasa de error baja, pero un poco menos

Δ errores de conversión
por transformación de la fecha relativa



#	codi	documen	tipusDada	valor	paraulaInicia	paraulaFinal
1	1	400	dataReg	1768/07/09	2	8
2	2	400	ciutatReg	gerona	21	21
3	3	400	notari	francisco lagrifa	28	28
4	4	400	ofiNotari	notario	31	31
5	5	400	notaria	gerona	32	33
6	6	400	dataActa	1768/07/04	36	41

Calidad en los términos críticos: antropónimos

error transcripción \ error de etiquetado	sin error	inclusión términos adyacentes	incompleto	incorrecto	no detectado	suma
sin error	**6309**	68	99	9	23	6508
incompleto	232	6	59	8	3	308
incorrecto	575	9	46	15	8	653
no transcrito	3		6	3	2	14
suma	7119	83	210	35	36	7483

% sin error	84%
% con error de transcripción	13%
% con error de etiquetado	5%

tasa de error moderada

Calidad en los términos críticos: oficios

error transcripción \ error de etiquetado	sin error	inclusión términos adyacentes	incompleta	incorrecta	no detectado	suma
sin error	**7239**	88	6	1	76	7410
incompleta	30		3		4	37
incorrecta	29	1		1	3	34
no transcrito					2	2
suma	7298	89	9	2	85	7483

% sin error	97%
% con error de transcripción	1%
% con error de etiquetado	2%

tasa de error muy baja

Calidad en los términos críticos: topónimos

error transcripción \ error de etiquetado	sin error	inclusión términos adyacentes	incompleta	incorrecta	lugar relativo no detectado	no detectado	suma
sin error	**3925**	11	184	223	2736	56	7135
incompleta	78		19	1	8	5	111
incorrecta	153		20	8	24	17	222
no transcrito	12				3		15
suma	4168	11	223	232	2771	78	7483

% sin error	52%
% con error de transcripción	5%
% con error de etiquetado	44%

tasa de error transcripción = baja

tasa error etiquetado = muy elevada

principal error = no detección del lugar relativo (error de fase 2)

Interés = análisis espacial → vinculación a Sistemas de Información Geográfica

Otro ensayo de evaluación de la identificación de topónimos

Sobre 67.426 documentos

- Objetivo: extracción de la **notaría** donde se ha realizado cada escritura
  - Operación relativamente simple → estructura constante = 2º topónimo

    *Arriendo En $dataAbs[1770/02/09] se ha presentado en el Oficio de Hipotecas de esta ciudad de **$llocAbs[Girona]** una escritura otorgada ante $antroponim[juan bautista morell y milsocos] $ofici[notario] de la villa de **$llocAbs[Figueres]** en $dataRel[1770/01/14]...*

  - Elevada frecuencia de **referencias relativas**

    *Establecimiento En $dataAbs[1768/06/02] se ha pntado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una Escritura otorgada ante $antroponim[francisco casanoves y garriga] $ofici[notario] **de dicha Ciudad**, en $dataRel[1768/05/28]...*

    Resultado del algoritmo de conversión de topónimos relativos

    *Establecimiento En $dataAbs[1768/06/02] se ha pntado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una Escritura otorgada ante $antroponim[francisco casanoves y garriga] $ofici[notario] de **$llocRel[Girona],** en $dataRel[1768/05/28]...*

Otro ensayo de evaluación de la identificación de topónimos

- Complicaciones: aparición de **regencias** → referencia a más de una sede notarial

Venta En $dataAbs[1768/08/25] ve se ha presentado en el Oficio de Hipotecas de esta Ciudad de $llocAbs[Girona] una escritura otorgada ante $antroponim[pedro puig] ***$ofici[notario] de la Villa de $llocAbs[Figueres]*** ***regentante*** ***la Notaria del Castillo de $llocAbs[Siurana]***

palabra clave → refinamiento del algoritmo

- Incluye la **normalización** de los topónimos → exigència para uso en GIS

Nomenclàtor oficial de toponímia major de Catalunya (2009) → coordenadas UTM

Otro ensayo de evaluación de la identificación de topónimos

Resultado identificación de la notaría de cada documento

		n		%	
Total documentos			67.426		100%
Topónimo notaría identificado en la transcripción			56.501		84%
	topónimo correcto y normalizado *	18.736		28%	
	topónimo corregido o normalizado*	36.964		55%	
	identificación incorrecta	801		1%	
Topónimo notaría no identificado en la transcripción			10.925		16%
	reconstruido con búsquedas selectivas y reemplazos masivos	10.281		15%	
	no reconstruidos	644		1%	

* Incluye topónimos relativos reconstruidos

Sólo el 28% de notarías identificadas sin necesidad de corrección

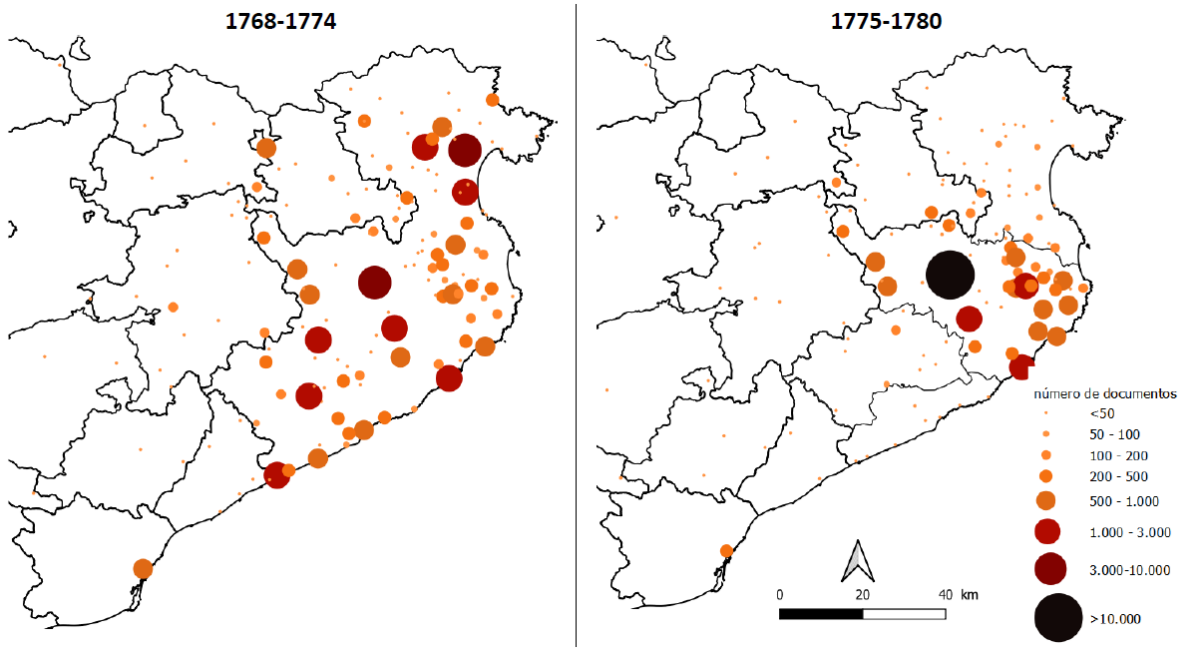
La mayor parte (55%) requirió corrección en la transcripción y/o normalización (no distinguido)

Se pudo reconstruir el 15% de topónimos no identificados

Uso de consultas de búsqueda y corrección en lote

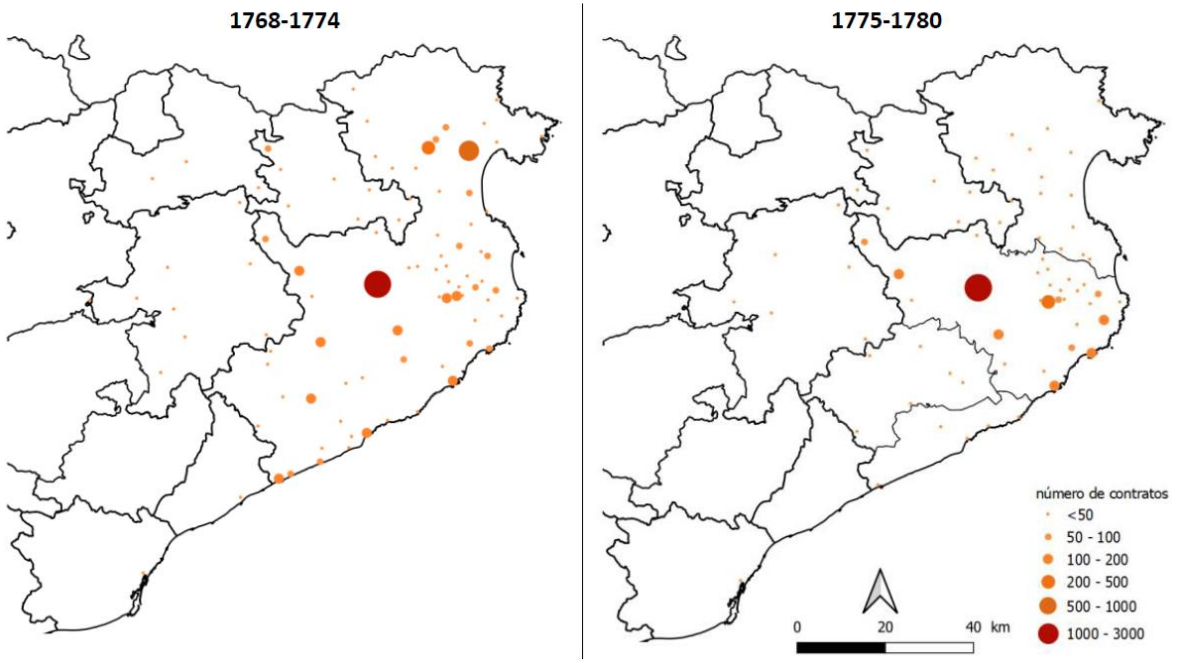
Alimentan tabla de normalizaciones

coste razonable



Actividad notarial (general)

- Mayor parte de la contratación realizada dentro del distrito hipotecario
- Moderada concentración en la ciudad de Girona
- Red tupida de notarías secundarias



Notarías donde se escrituraron censos consignativos

- Escasez de préstamos contratados fuera de las notarías del distrito hipotecario
- Mayor peso relativo de la ciudad de Girona
- Menor importancia de las notarías secundarias

↓

Relación con las principales instituciones prestamistas (religiosas)

Los otros topónimos

Elementos de complejidad

Diversidad de **funciones**:

- residencia de cada uno de los actores de la escritura
- naturaleza de los mismos
- localización de las fincas objeto de transacción
- notas de traslado a otros oficios de hipotecas

Diversidad **tipológica**:

- lugares, vecindarios, parroquias, pueblos o municipios
- condado, obispado, distrito…
- microtopónimos para las lindes de los campos

Cadenas jerárquicas multitoponímicas

establecieron á $antroponim[julian vila] $ofici[carretero] del ***vezindado de las Barracas Parroquia de dicho Lugar de $llocAbs[Celrà]***

Procedimiento de solución

Determinable mediante **palabras clave:**

natural de
residente en
habitante de
presente en
oy en
…

Determinable porque se explicita el tipo o categoria de cada topónimo

Importante para la desambiguación
Reducción a una unidad tipológica operativa

municipio

Resultado: cartografía de los flujos de crédito

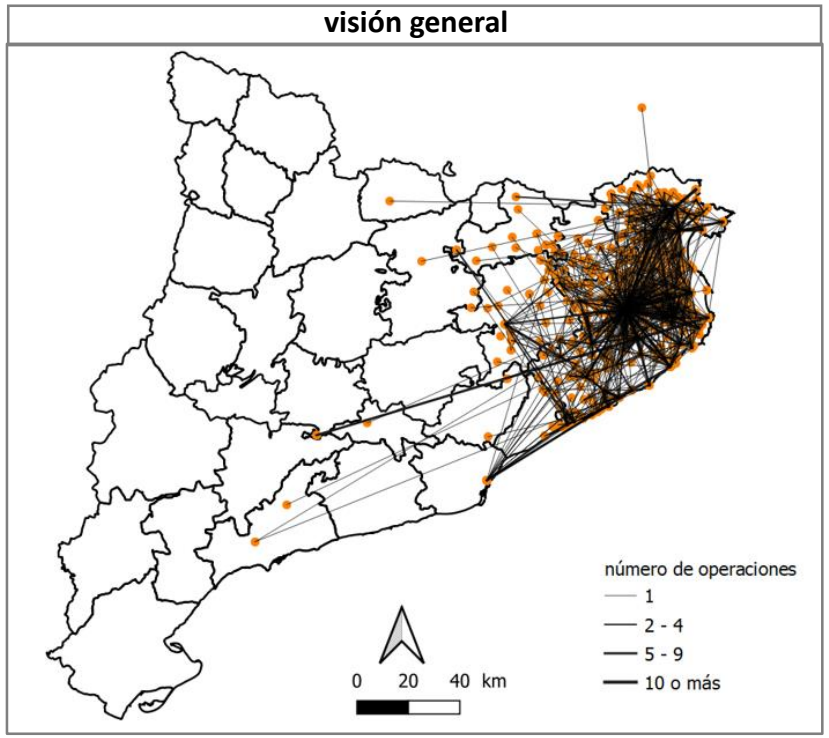
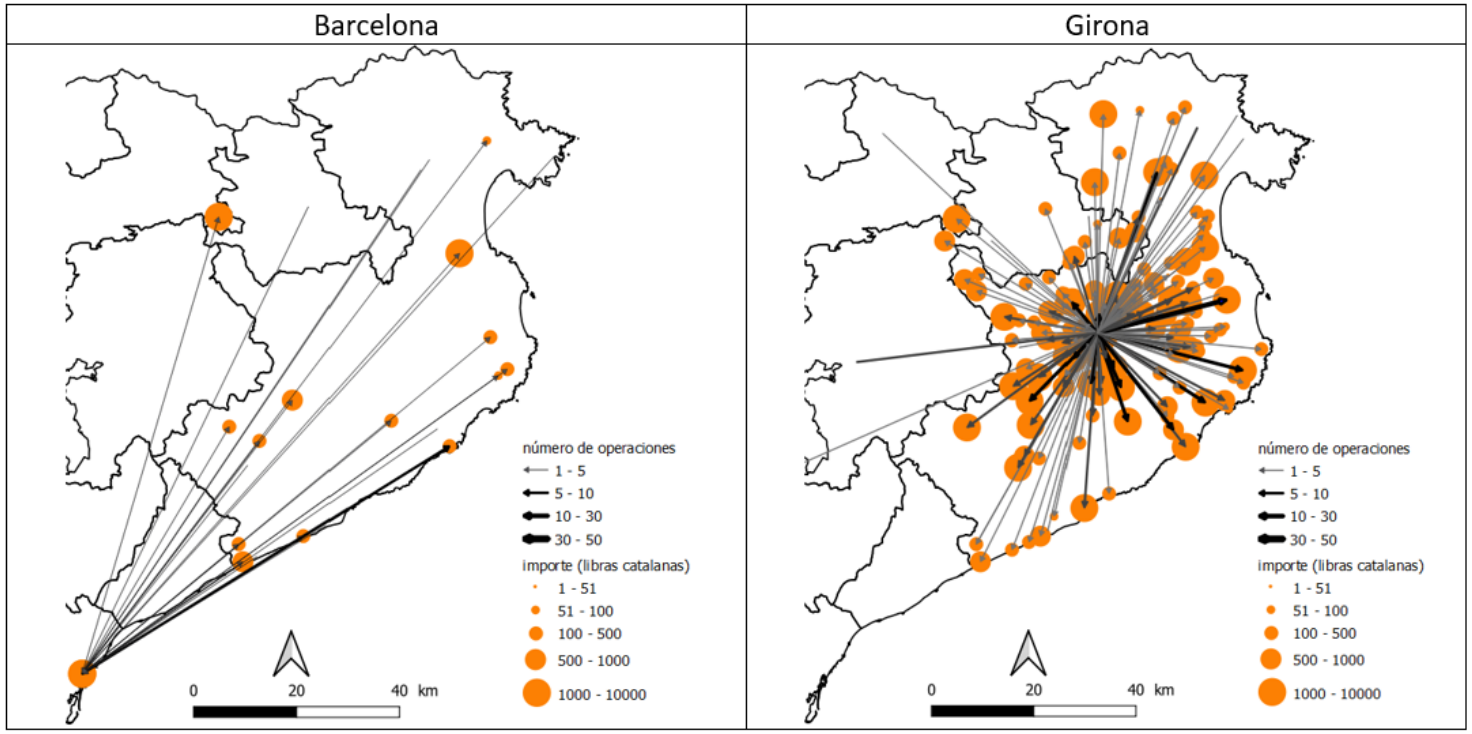
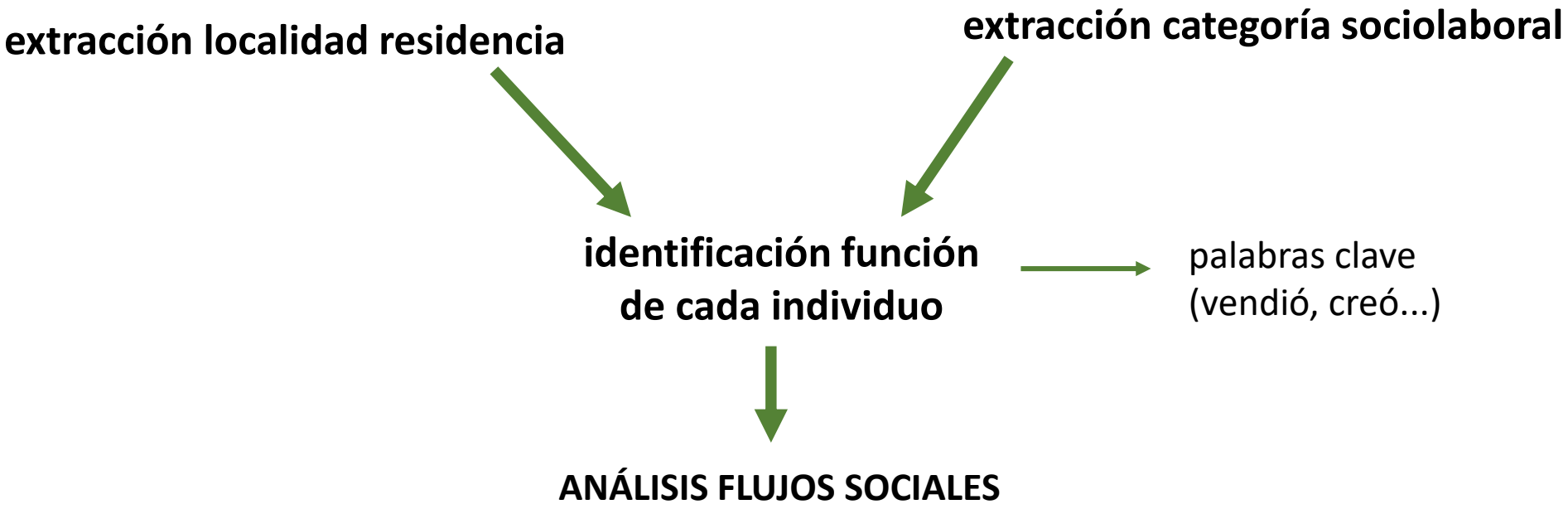


Figura 5. Cartografía de los flujos de crédito, según lugar de residencia de censalistas (1768-1773)

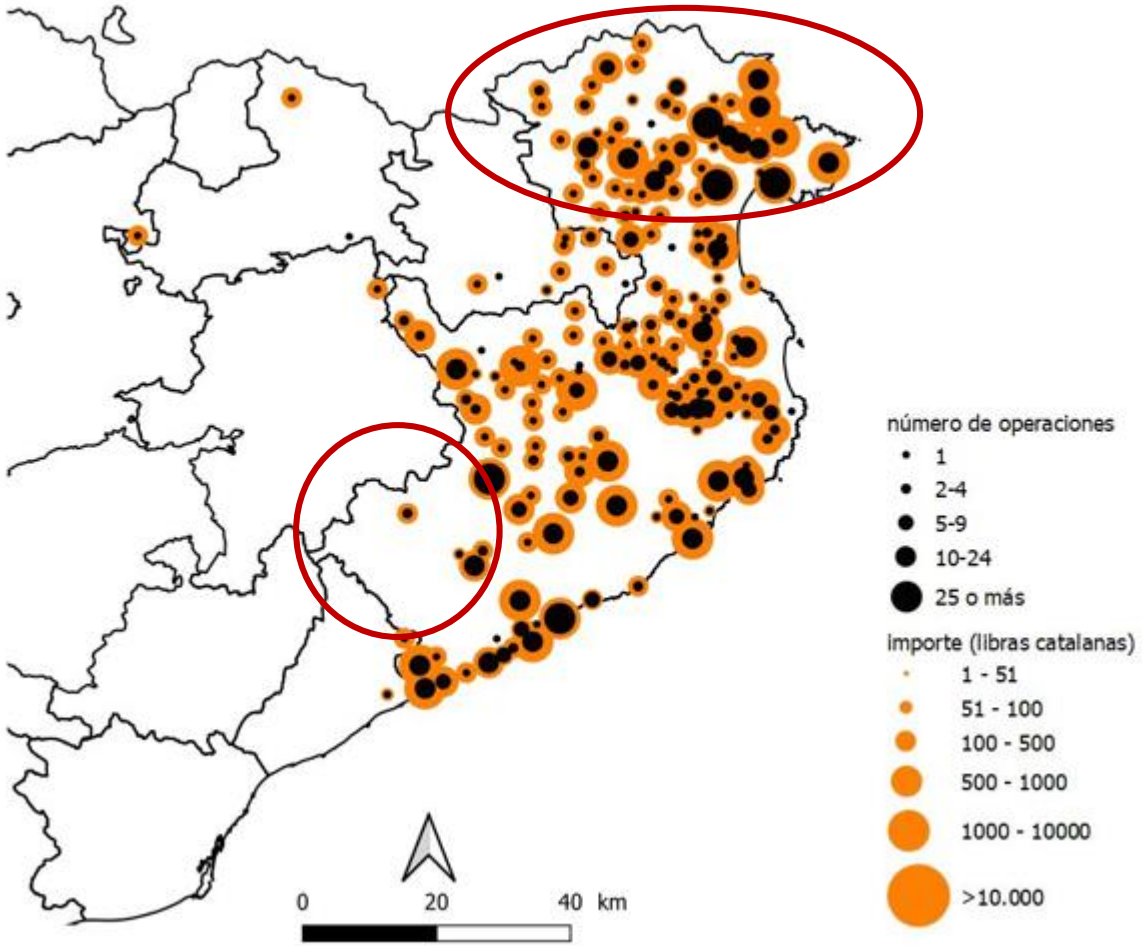


Avances en la cartografía social



Acceso de los ‘treballadors’ al crédito hipotecario

Figura 6: Localización de los trabajadores que accedieron a crédito entre 1768 y 1773



posibilidad de visualizar la actuación de los grupos sociales con menor rastro documental (*treballadors*)

receptores del 29% operaciones de crédito

peso demográfico grupo = 33%

volumen capital = 14%

Acceso amplio al crédito hipotecario

cartografía = acceso no homogéneo

- zona vitícola Empordà = intenso
- zona la Selva-Montseny = débil

Conclusión

- Expectativas en la aplicación de la transcripción automática a la investigación histórica = elevadas
  - probablemente modificaran las formas de trabajar con documentación manuscrita
  - exigirán nuevas competencias → habilidad manejo grandes cantidades de documentación transcrita

- Estadio actual → implica asumir:

  1. Costes elevados en la revisión del proceso de transcripción (layout)
  2. Inevitabilidad de cierta tasa de error → menor precisión y exactitud en los resultados
  3. Necesidad de procedimientos de revisión, corrección y normalización de los datos
     - manuales o asistidos, en buena medida
     - esfuerzo laboral importante

- El volumen de datos generados permite minimizar el impacto de los errores derivados del proceso de transcripción