

Explorando los cambios sociales silenciosos: la explotación digital de una gran mina de datos históricos

Rosa Congost, **Jordi Regincós**, Rosa Ros y Enric Saguer
Universitat de Girona



**XXX Seminari d'Història Econòmica i Social
DEL TRAÇ AL BYTE
Eines digitals per a l'estudi històric del canvi social**

XXX Seminari d'Història Econòmica i Social
Girona, 4 i 5 de juliol de 2019

DEL TRAÇ AL BYTE

**Eines digitals per a l'estudi històric
del canvi social**

*... i veure al traç de la cultura
... i la cultura, i la cultura, i la cultura*

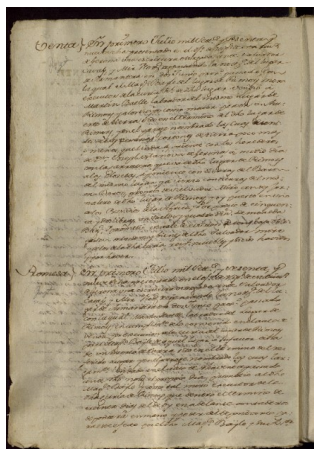
XXX Seminari d'Història Econòmica i Social
Girona, 4 i 5 de juliol de 2019

DEL TRAC

A LA DADA

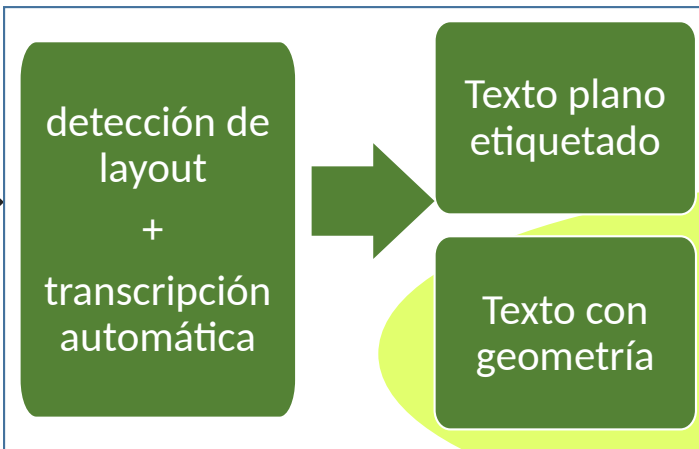
Eines digitals per a l'estudi històric
del canvi social

*... i veure al voltant de la...
... i moxata, ...*

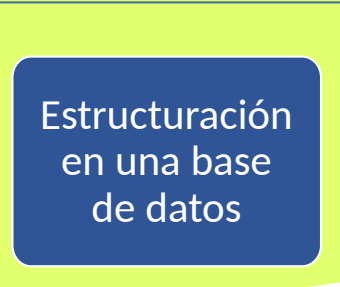
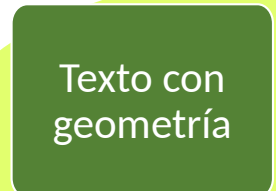


Fase 0
Digitalización
AHG & CRHR

103.300 imágenes



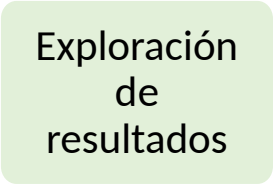
Fase 1
Transcripción
PRHLT & CRHR

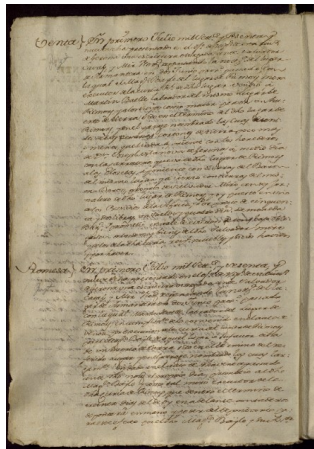


Fase 2
J. Regincós
EPS UdG



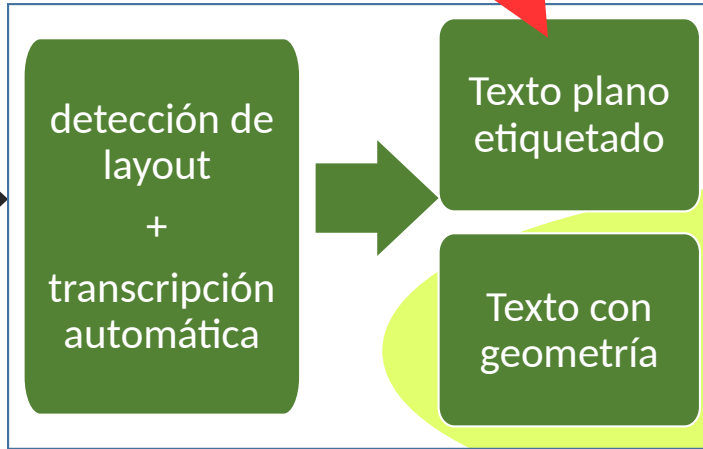
Fase 3
CRHR





Fase 0
Digitalización
AHG & CRHR

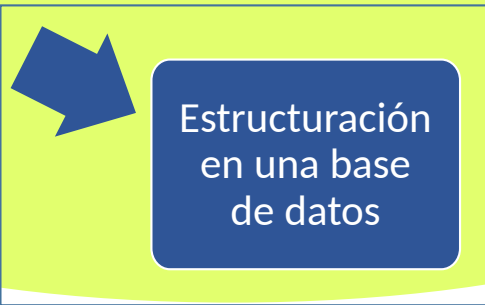
103.300 imágenes



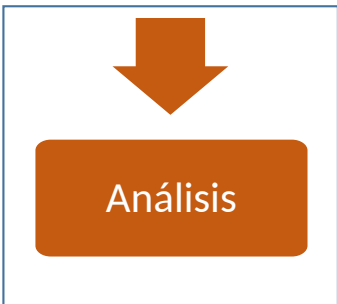
Fase 1
Transcripción
PRHLT & CRHR



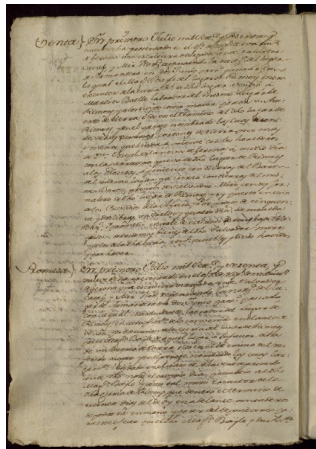
Exploración
de
resultados



Fase 2
J. Regincós
EPS UdG

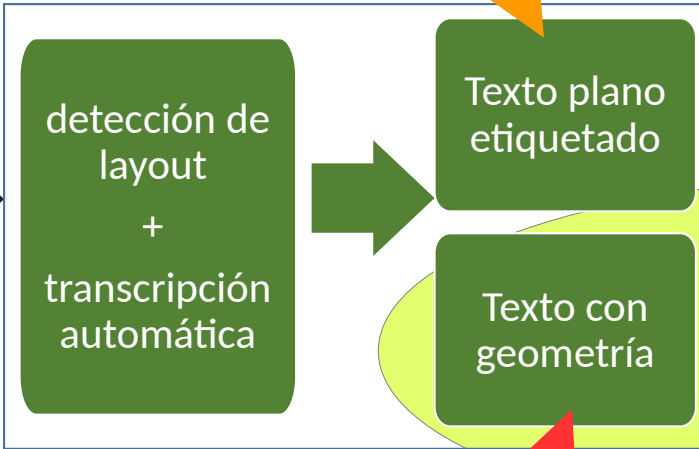
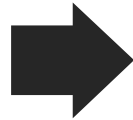


Fase 3
CRHR

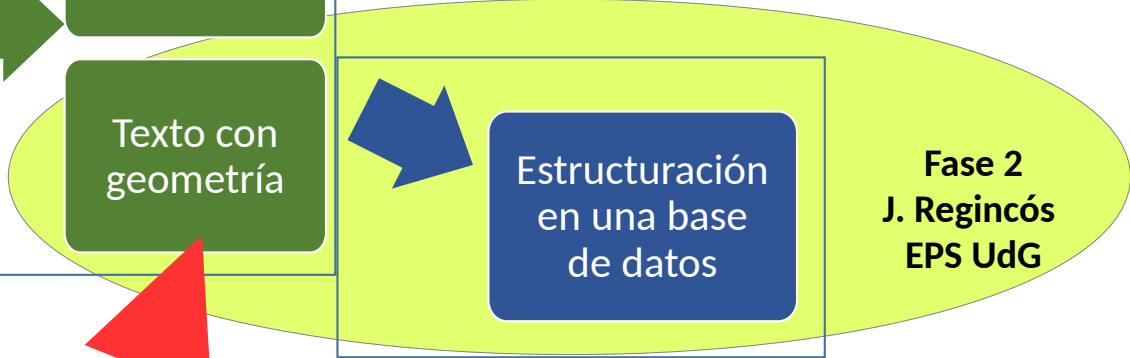


Fase 0
Digitalización
AHG & CRHR

103.300 imágenes



Fase 1
Transcripción
PRHLT & CRHR



Fase 2
J. Regincós
EPS UdG



Exploración de resultados



Fase 3
CRHR

Texto plano etiquetado

```
<TextRegion>
  <tip> Carta de pago </tip>
</TextRegion>
<TextRegion>
  En treinta Abril mil setecientos, sesenta, y ocho, se
  presentó en el Oficio de Hipotecas de esta ciudad de <top>Gerona</top>
  una Escritura otorgada ante <ant>Feliz Veguer</ant> <ofi>Notario</ofi> de
  numero de la Ciudad de <top>Barcelona,</top> en veinte, y nueve de Mar-zo
  del mismo año con la qual <ant>Pedro Flaquer</ant> <ofi>tirador de</ofi>
```

Texto con geometria

```
<TextRegion id="TextRegion_1554219078945_716" custom="readingOrder {index:0;} structure {type:$tip;}">
  <Coords points="102,200 161,200 139,220 158,214 186,215 173,202 264,202 278,199 283,205 288,205 304,216
23,196 418,200 470,172 492,200 497,232 494,243 490,259 492,263 492,275 483,305 83,305 80,216"/>
  <TextLine id="TextLine_1_9" custom="readingOrder {index:0;} structure {type:$tip;}">
    <Coords points="475,224 398,231 309,240 109,240 109,290 311,290 403,281 480,274"/>
    <Baseline points="109,278 311,278 402,269 479,262"/>
    <TextEquiv>
      <Unicode>$tip:Carta $tip:de $tip:pago</Unicode>
    </TextEquiv>
  </TextLine>
</TextRegion>
<TextRegion id="TextRegion_1554219033768_699" custom="readingOrder {index:1;} structure {type:$pac;}">
  <Coords points="584,157 2583,180 2616,1387 2613,1968 624,1968"/>
  <TextLine id="TextLine_20_10" custom="readingOrder {index:0;} structure {type:$pac;}">
    <Coords points="2460,173 2066,193 1548,204 936,216 607,235 610,285 938,266 1549,254 2068,243 2462,223"/>
    <Baseline points="610,273 938,254 2068,231 2462,211"/>
    <TextEquiv>
      <Unicode>En treinta Abril mil setecientos, sesenta, y ocho, se</Unicode>
    </TextEquiv>
  </TextLine>
  <TextLine id="TextLine_19_10" custom="readingOrder {index:1;} structure {type:$pac;}">
    <Coords points="2511,278 2456,297 2027,294 645,333 647,383 2028,344 2465,347 2528,325"/>
    <Baseline points="647,371 2028,332 2463,335 2524,314"/>
    <TextEquiv>
      <Unicode>presentó en el Of.$^o$.Oficio de Hipote.$^s$.Hipotecas de esta ciudad de $top:Gerona</Unicode>
    </TextEquiv>
  </TextLine>
```


La transcripción nos ofrece (i)

- Para cada página
 - Conjunto de regiones
- Para cada región
 - Tipo de región (\$tip, \$not, \$nop, \$par...)
 - Geometría (polígono que la circunscribe)
 - Conjunto de líneas
- Para cada línea
 - Geometría (polígono que la circunscribe)
 - Texto (etiquetado)

- Etiquetas

- Algunas palabras estarán etiquetadas:

- \$tip → Tipo documento
- \$ant → Antropónimo
- \$top → Topónimo
- \$ofi → Oficio
- \$- → palabra partida final/inicio de línea
- \$. → abreviatura
- \$^ → superíndice
- ...

¿Qué buscamos con la base de datos?

¿Qué buscamos con la base de datos? (i)

- Poder almacenar un conjunto de **documentos** (regiones, líneas, palabras) y sus **datos** (notario, notaría, fechas, intervinientes...)
- Mantener la trazabilidad de los datos
 - Poder saber de qué parte de los XML procede cada palabra.
 - Posibilidad de exportar los datos *más o menos* brutos para ser explotados aplicando *Data Mining* o otras tecnologías.

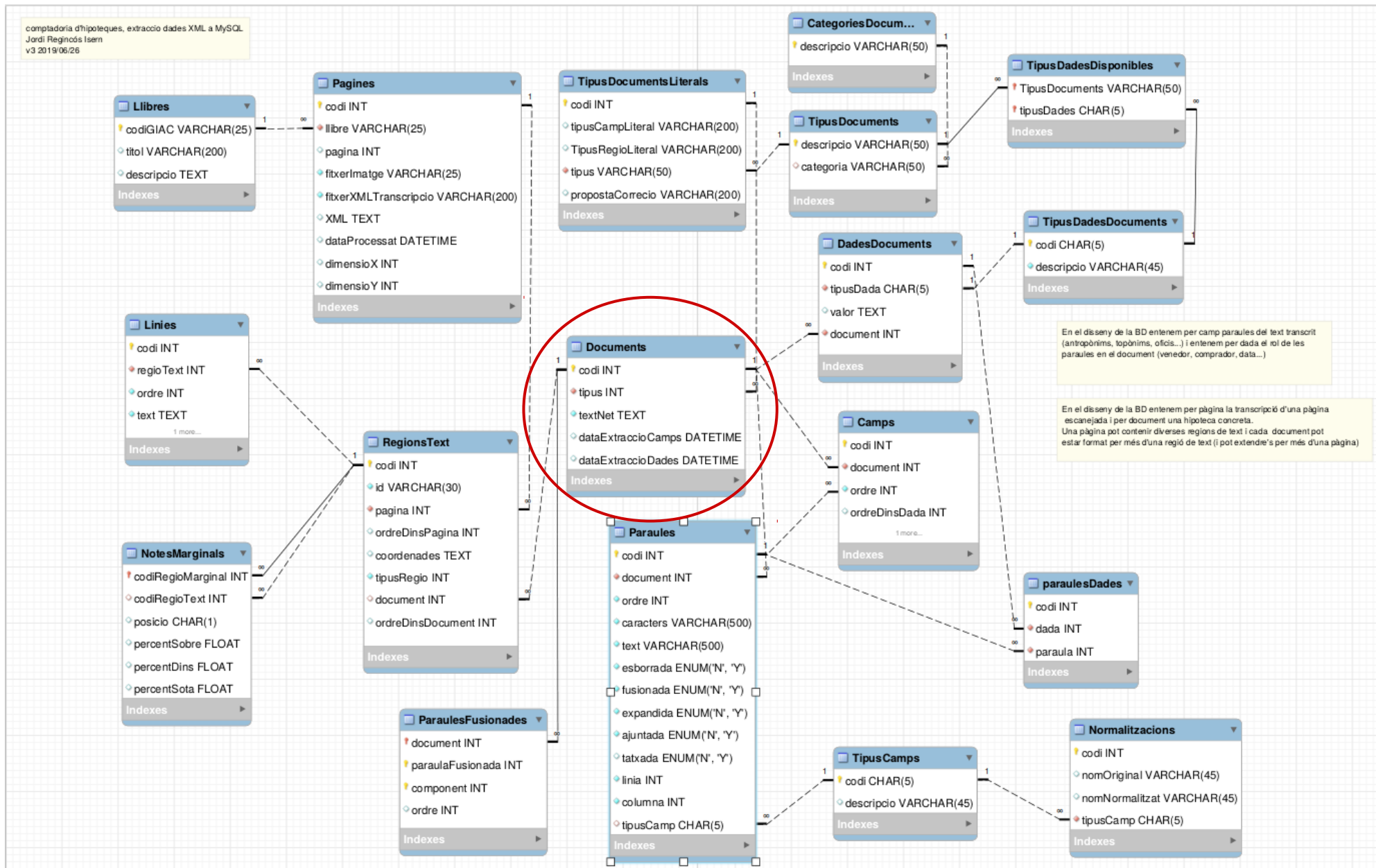
¿Qué buscamos con la base de datos? (ii)

- Almacenar los datos en un formato perdurable
 - Modelo relacional se define el 1970
 - SQL se consolida a partir de los 80.
- Disponer de dos conjuntos de datos
 - Uno con los datos extraídos automáticamente
 - Otro con los datos validados o corregidos.

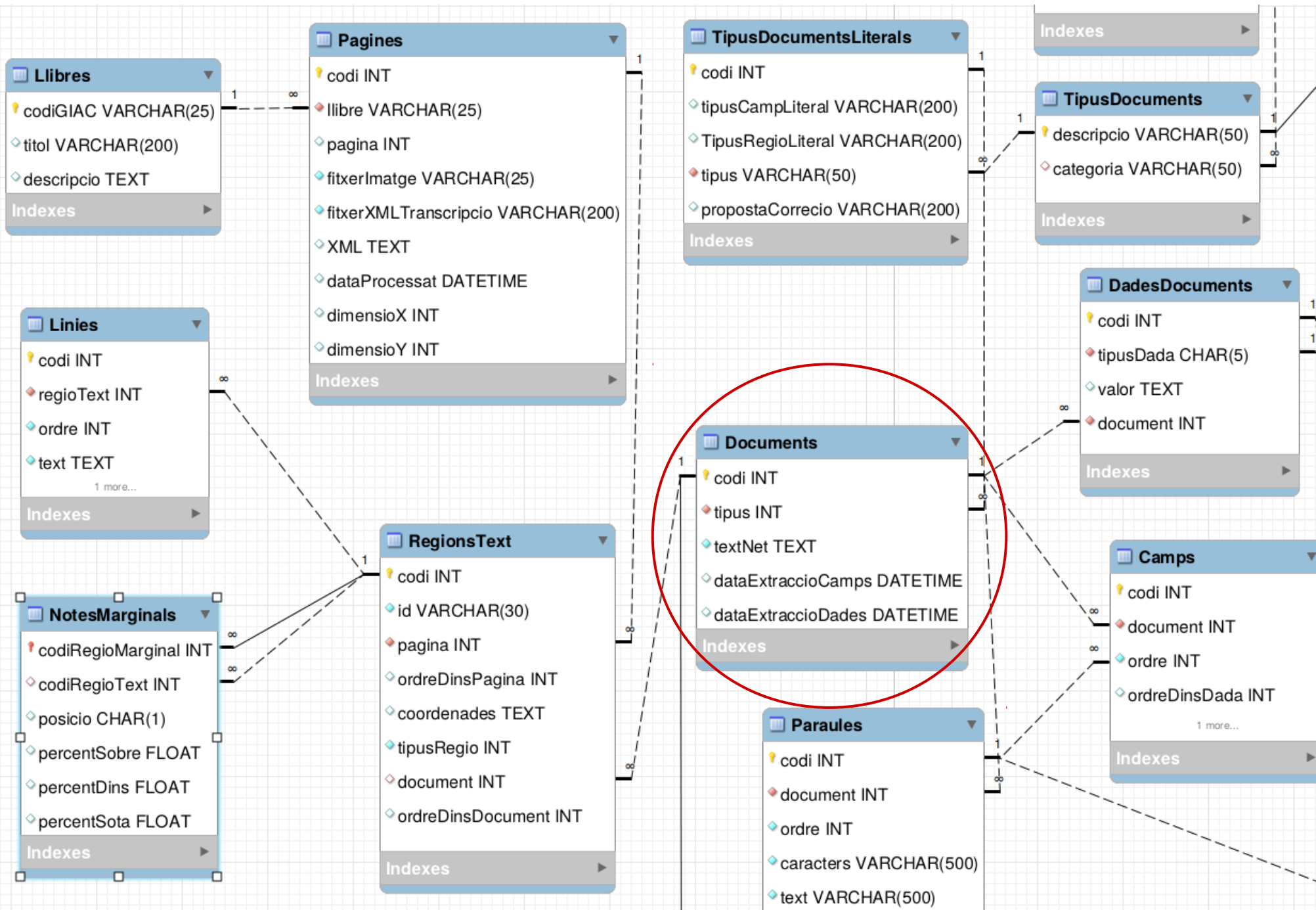
Herramientas utilizadas

- Sistema Gestor de Base de Datos:
 - MySQL (software libre y segundo SGBD más utilizado)
- Lenguaje de programación
 - C++
 - Conector de SQL de MySQL
 - Biblioteca para el proceso de XML TinyXML
- LibreOffice
 - ... para un poco de bricolaje

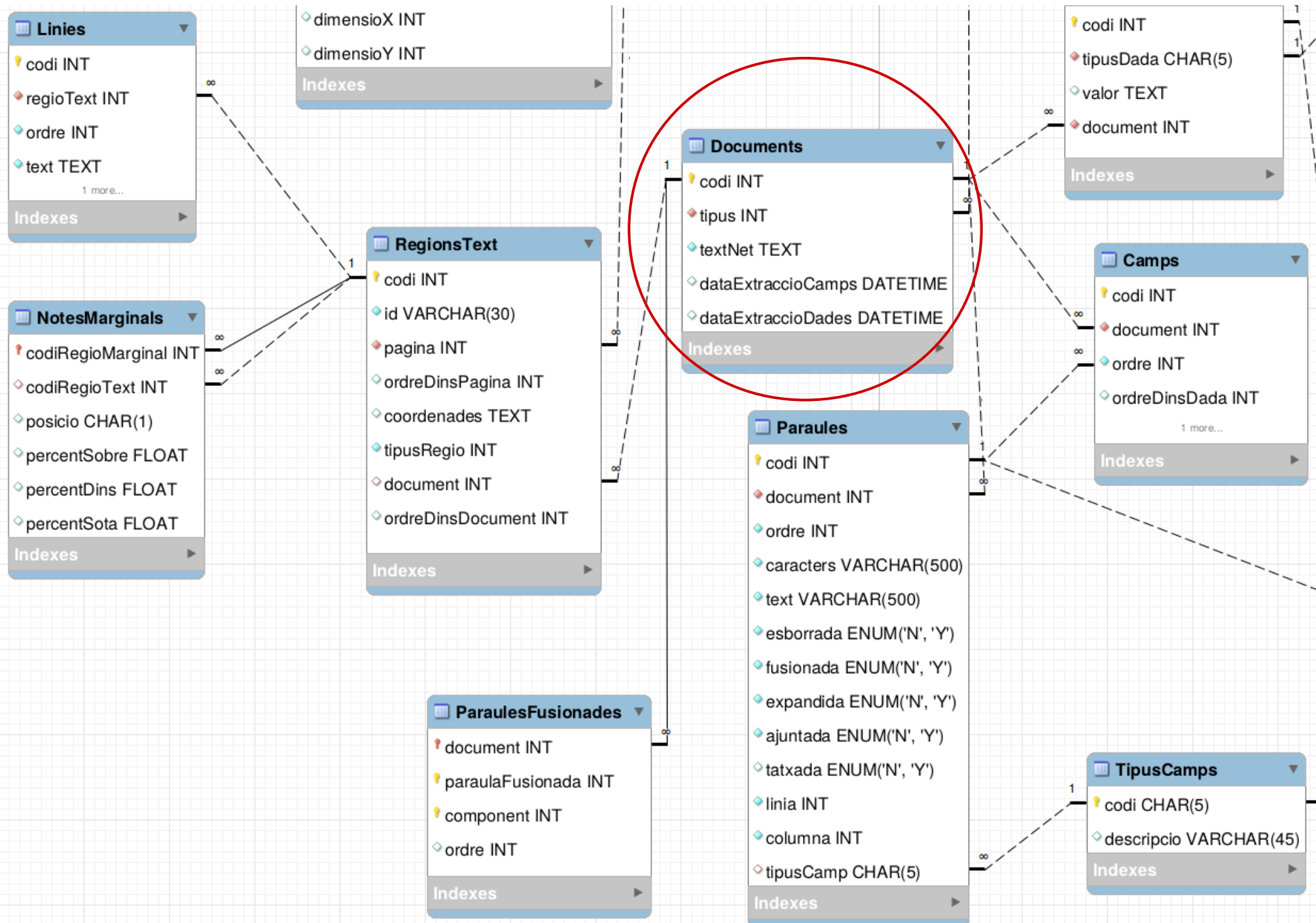
Esquema de la base de Datos



Ampliación esquema de la base de datos (i)



Ampliación esquema de la base de datos (ii)



Reconstrucción de los documentos

Reconstrucción de los documentos

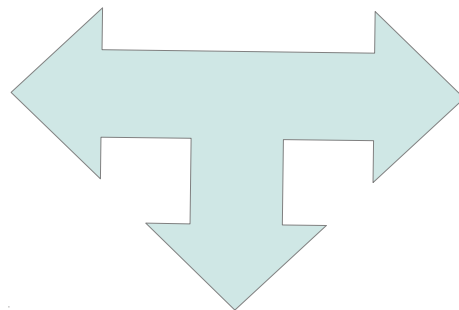
- **Objetivo:** Convertir un conjunto de páginas en un conjunto de documentos:
- Documento:
 - Secuencia ordenada de regiones de texto encabezada por una región TIPO_DOCUMENTO
 - +
 - Secuencia ordenada de notas al margen

Proceso reconstrucción documentos

- 1) Obtener las regiones desde los XMLs
- 2) Organizar las regiones página a página
- 3) Generar los documentos
- 4) Incorporarlos a la base de datos

Para reducir la imprecisión de la transcripción...

Información
textual



Información
geométrica

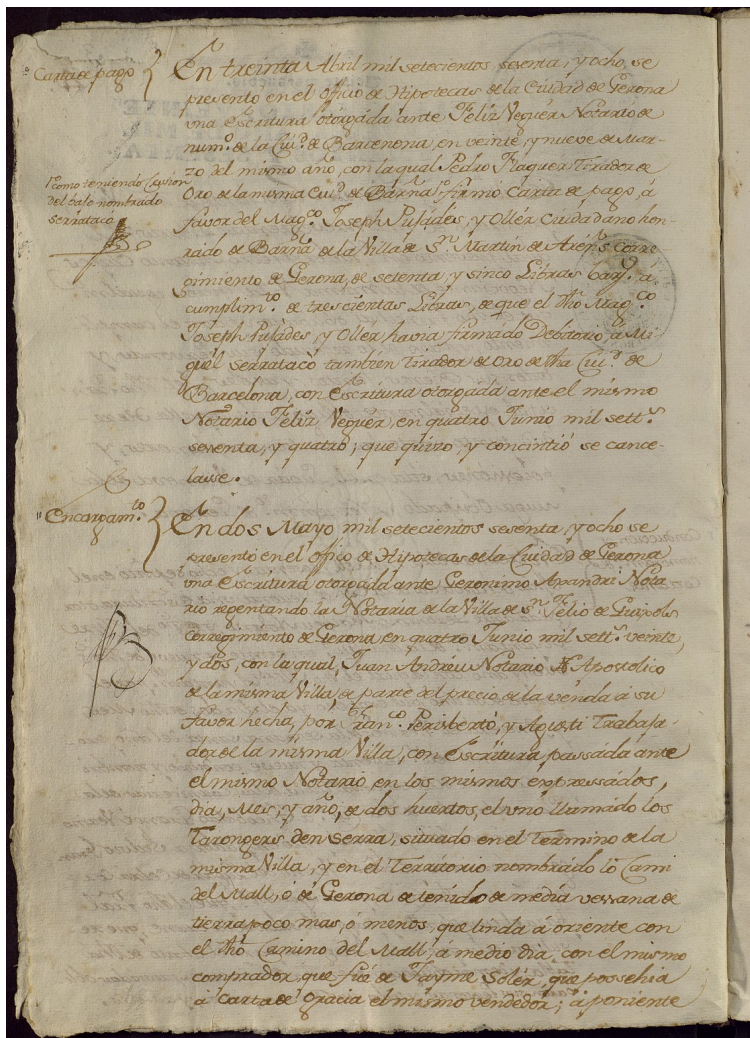
Etiquetas
transcriptor

Proceso reconstrucción documentos

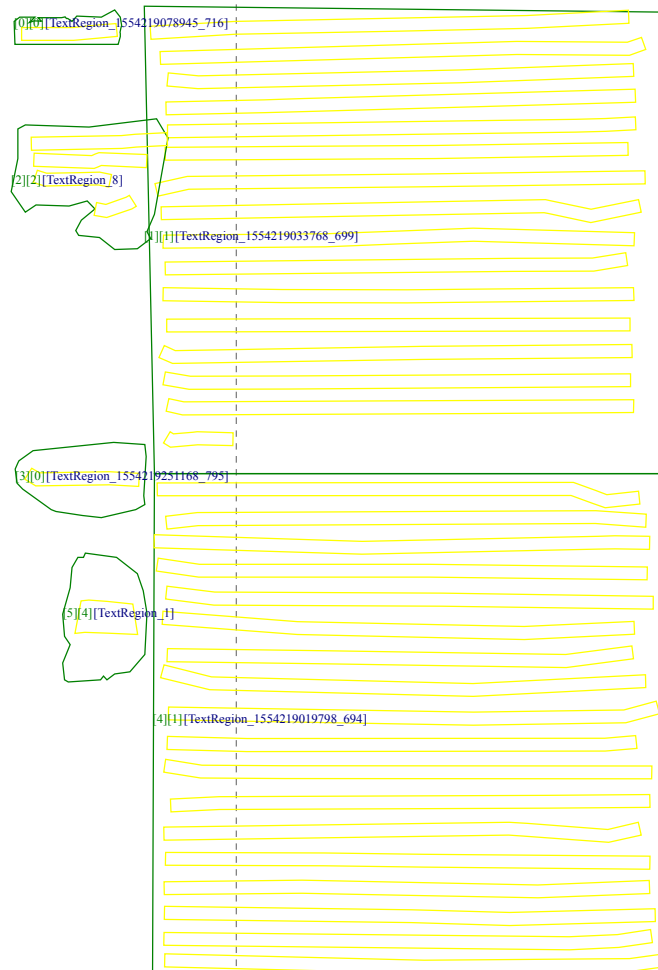
- Obtener las regiones desde los XML
 - Etiquetar regiones
- Organizar las regiones página a página
 - Ordenar vertical/horizontal
 - Enlazar regiones marginales y centrales
 - re-etiquetar regiones
 - Separar marginales TIPO de notas al margen
 - Reordenar vertical/horizontal
- Generar los documentos
 - Construir la secuencia de regiones
 - Procesar la secuencia
 - Extraer texto de las líneas, obtener palabras,...

Obtener regiones

```
</OrderedGroup>
</ReadingOrder>
<TextRegion id="TextRegion_1554219078945_716" custom="readingOrder {index:0;} structure {type:
$tip;}">
  <Coords points="102,200 161,200 139,220 158,214 186,215 173,202 264,202 278,199 283,205 288,205
304,216 323,196 418,200 470,172 492,200 497,232 494,243 490,259 492,263 492,275 483,305 83,305
80,216"/>
  <TextLine id="TextLine_1_9" custom="readingOrder {index:0;} structure {type:$tip;}">
    <Coords points="475,224 398,231 309,240 109,240 109,290 311,290 403,281 480,274"/>
    <Baseline points="109,278 311,278 402,269 479,262"/>
    <TextEquiv>
      <Unicode>$tip:Carta $tip:de $tip:pago</Unicode>
    </TextEquiv>
  </TextLine>
</TextRegion>
<TextRegion id="TextRegion_1554219033768_699" custom="readingOrder {index:1;} structure {type:
$pac;}">
  <Coords points="584.157 2583.180 2616.1387 2613.1968 624.1968"/>
```



2019-06-26 19:37:4

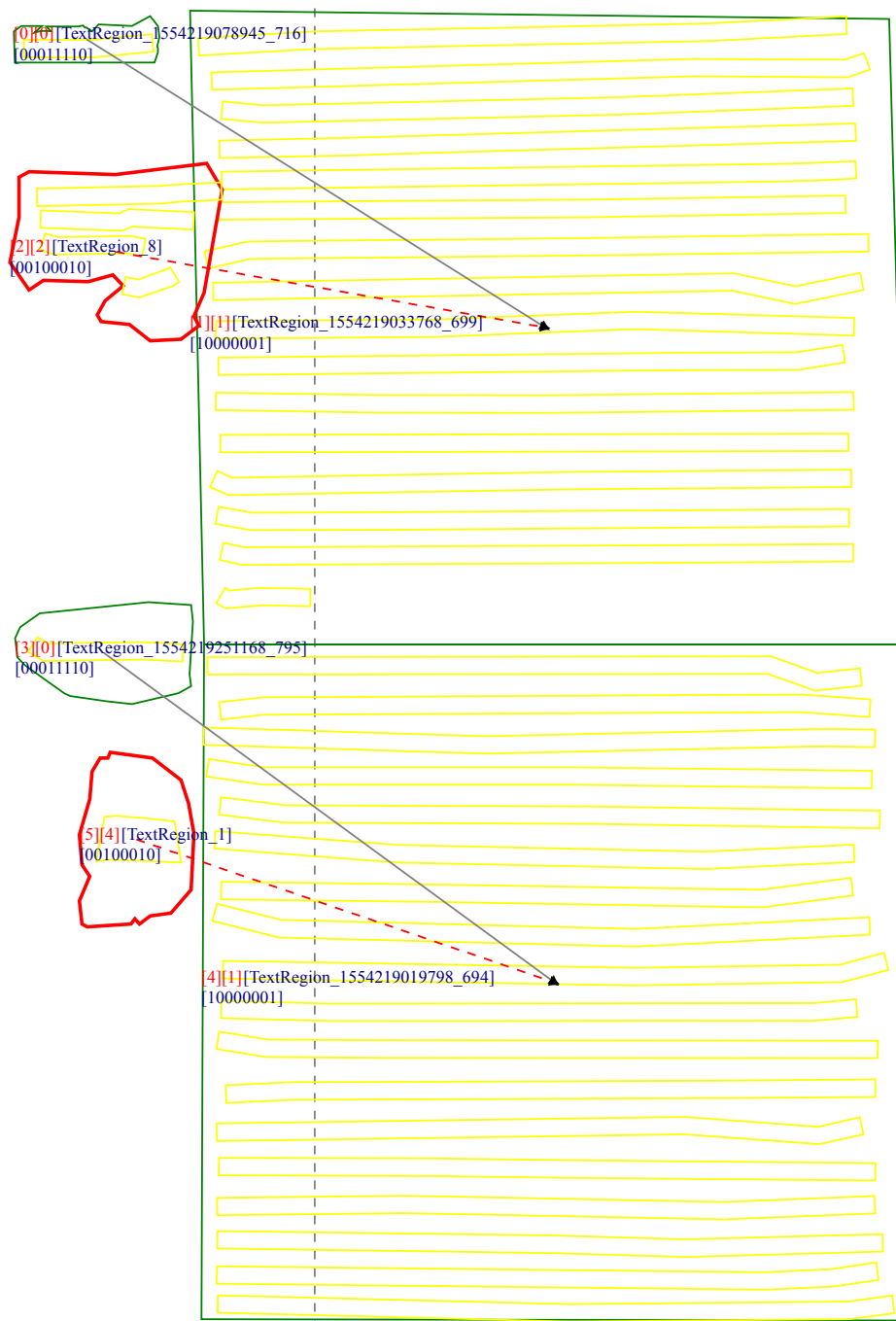
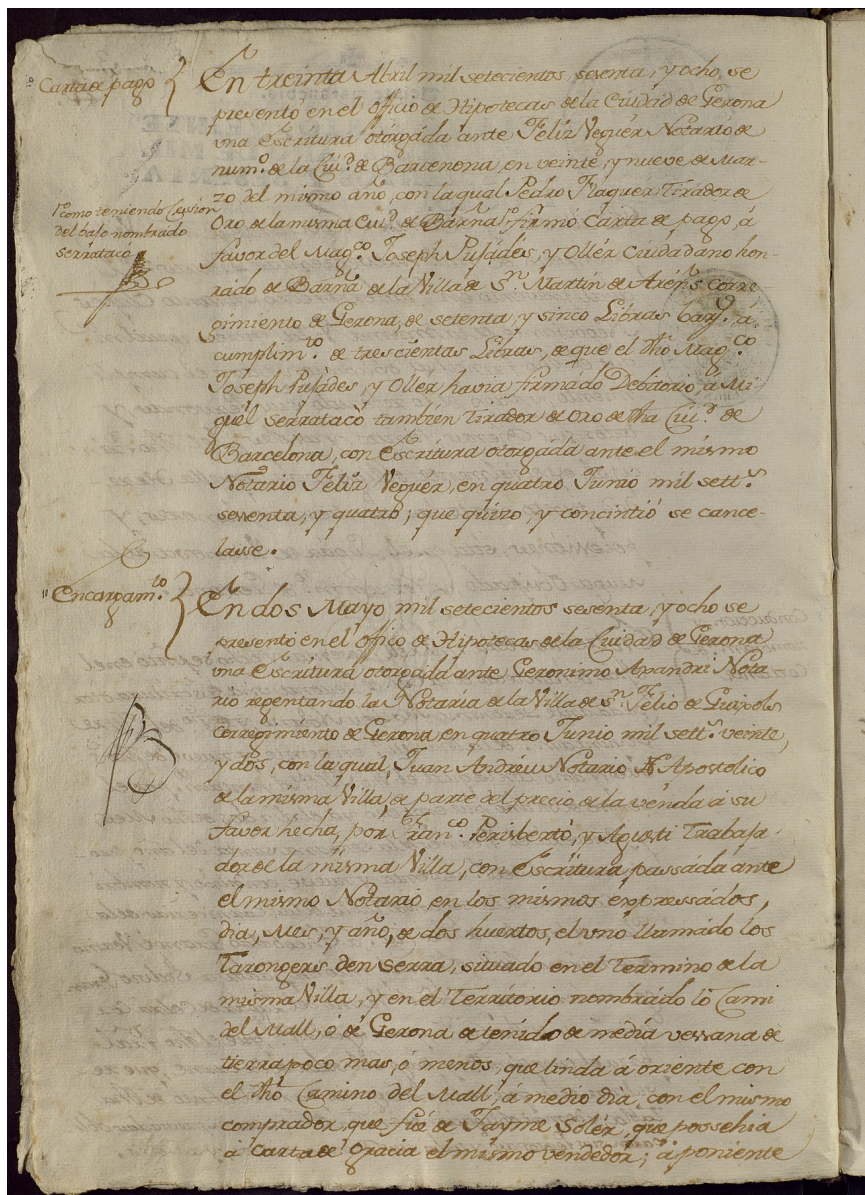


Proceso reconstrucción documentos

- Obtener las regiones desde los XML
 - Etiquetar regiones
- Organizar las regiones página a página
 - Ordenar vertical/horizontal
 - Enlazar regiones marginales y centrales
 - Completar etiquetado de las regiones
 - Separar marginales TIPO de notas al margen
 - Reordenar vertical/horizontal
- Generar los documentos
 - Construir la secuencia de regiones
 - Procesar la secuencia
 - Extraer texto de las líneas, obtener palabras,...

Organizar las regiones página a página

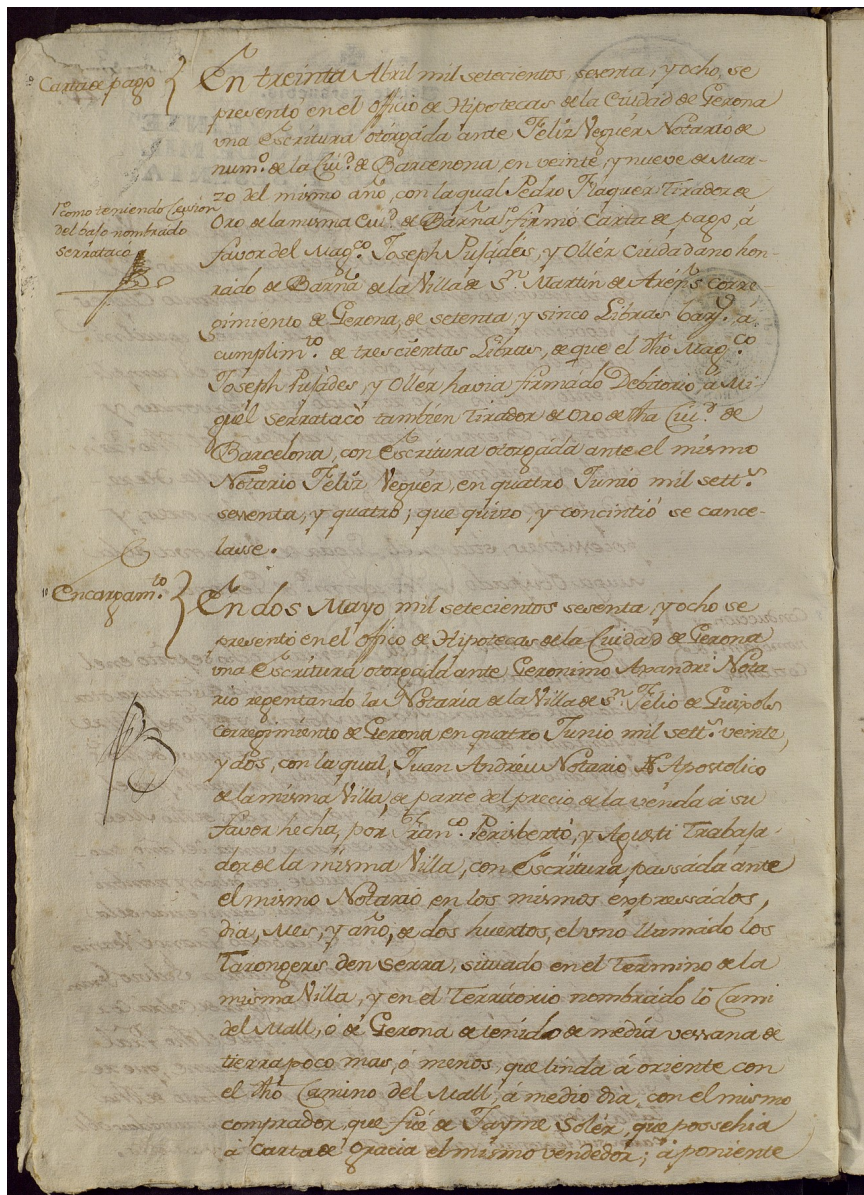
2019-06-26 19:37:14



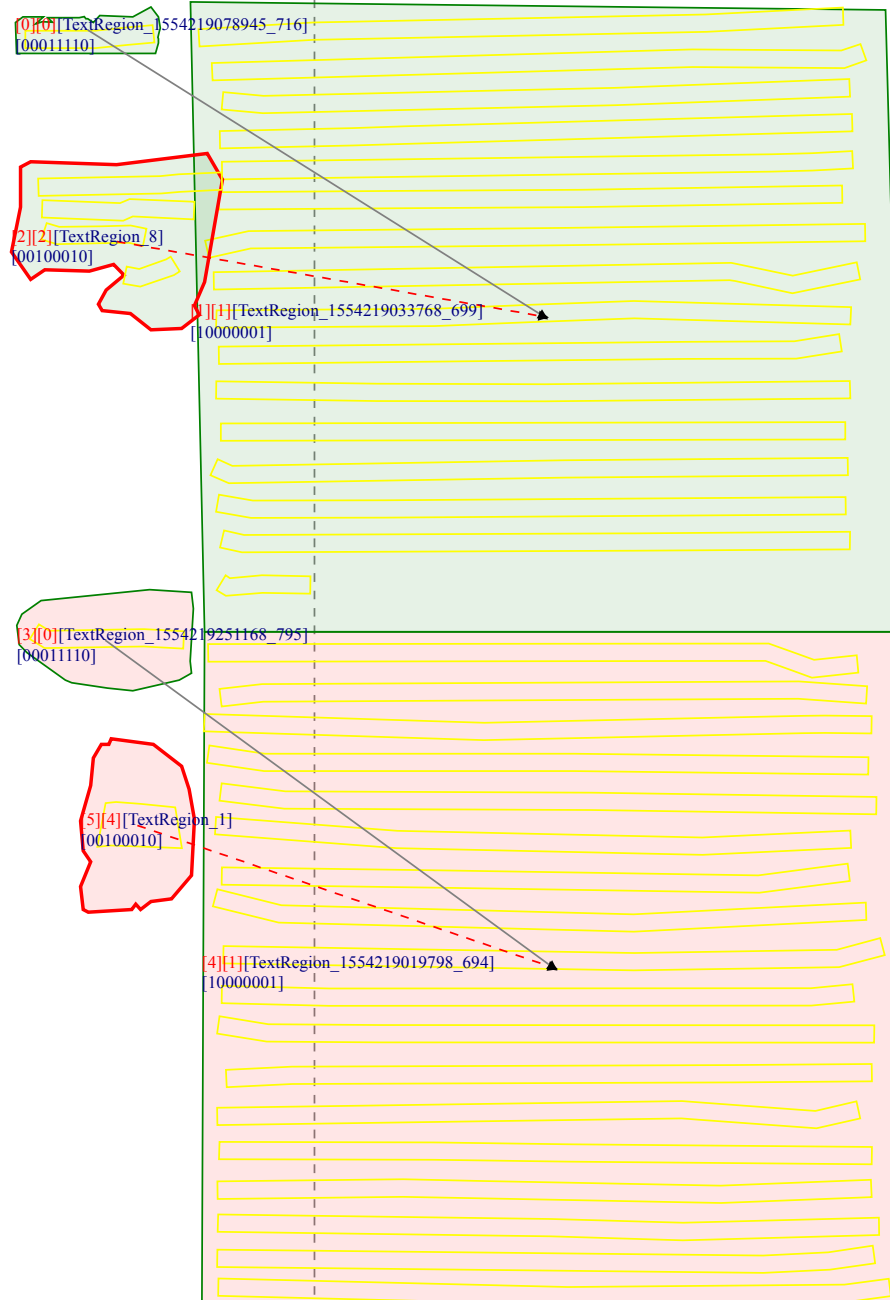
Proceso reconstrucción documentos

- Obtener las regiones desde los XML
 - Etiquetar regiones
- Organizar las regiones página a página
 - Ordenar vertical/horizontal
 - Enlazar regiones marginales y centrales
 - re-etiquetar regiones
 - Separar marginales TIPO de notas al margen
 - Reordenar vertical/horizontal
- **Generar los documentos**
 - Construir la secuencia de regiones
 - Procesar la secuencia
 - Extraer texto de las líneas, obtener palabras,...

Generar los documentos: encadenar regiones



2019-06-26 19:37:14

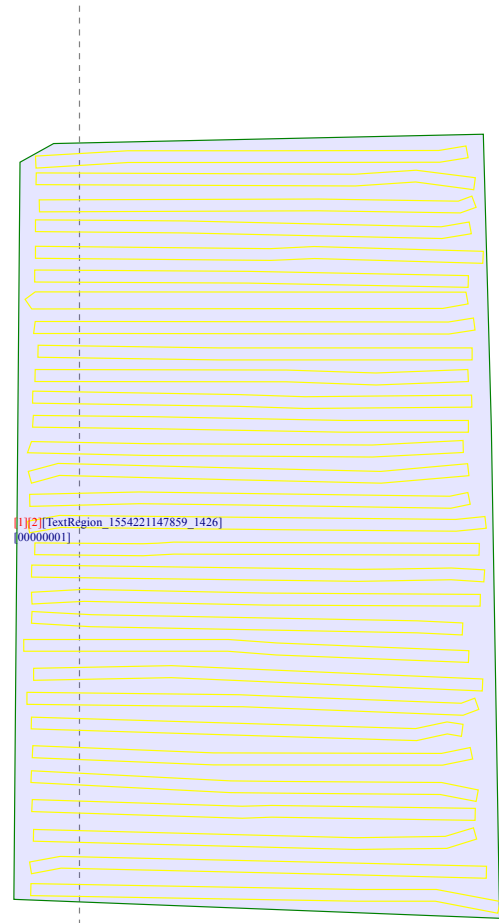


Generar documentos (3 páginas consecutivas)

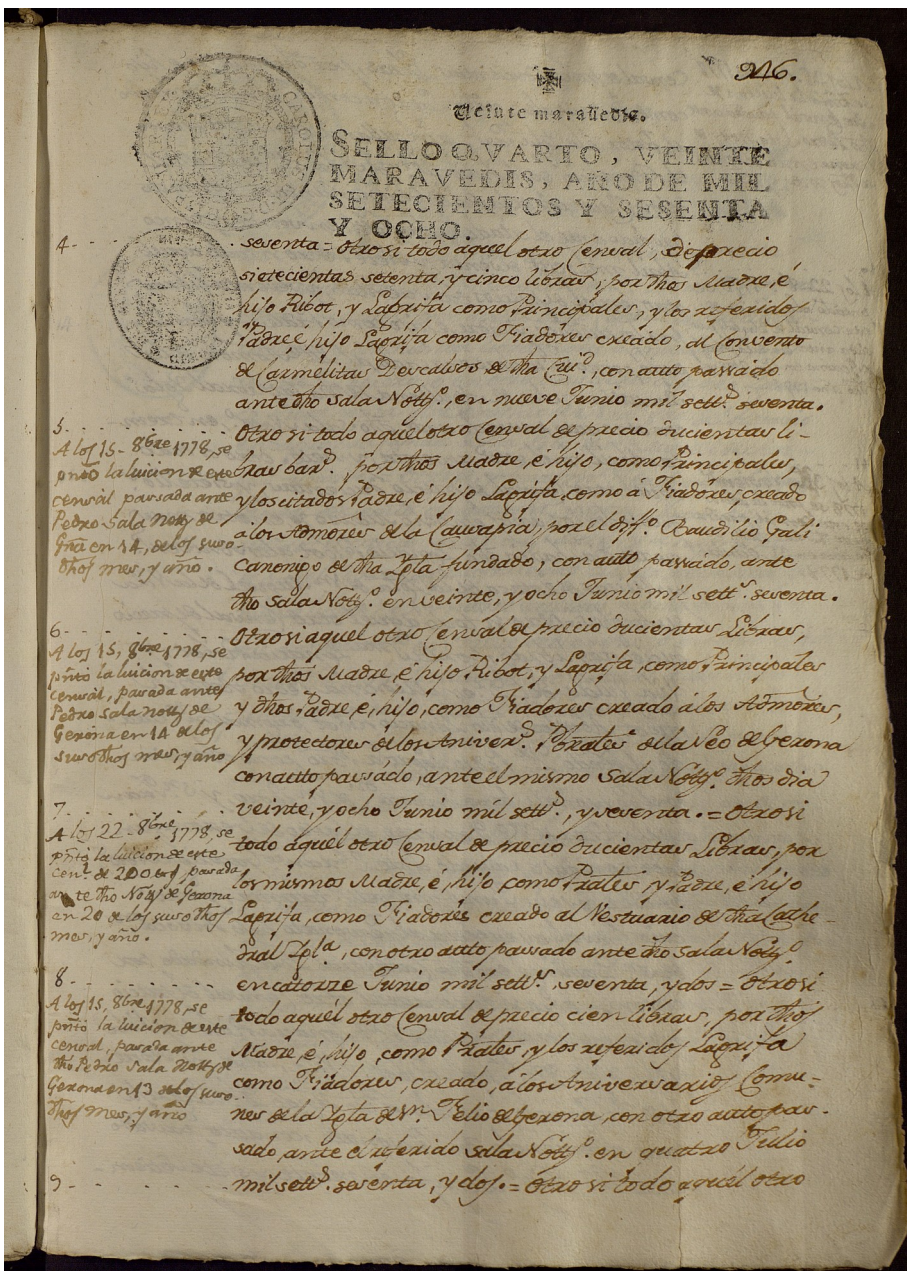
2019-06-26 19:37:14

2019-06-26 19:37:14

2019-06-26 19:37:14



Generar documentos (casos curiosos...)



2019-06-26 19:39:08



Generar documentos: extracción de palabras (i)

```

<TextRegion id="TextRegion_1554219078945_716" custom="readingOrder {index:0;} structure {type:$tip;}">
  <Coords points="102,200 161,200 139,220 158,214 186,215 173,202 264,202 278,199 283,205 288,205 304,216 323,196
418,200 470,172 492,200 497,232 494,243 490,259 492,263 492,275 483,305 83,305 80,216"/>
  <TextLine id="TextLine_1_9" custom="readingOrder {index:0;} structure {type:$tip;}">
    <Coords points="475,224 398,231 309,240 109,240 109,290 311,290 403,281 480,274"/>
    <Baseline points="109,278 311,278 402,269 479,262"/>
    <TextEquiv>
      <Unicode>$tip:Carta $tip:de $tip:pago</Unicode>
    </TextEquiv>
  </TextLine>
</TextRegion>
<TextRegion id="TextRegion_1554219033768_699" custom="readingOrder {index:1;} structure {type:$nac;}">

```

ordre	caracters	text	esborrada	fusionada	expandida	ajuntada	tatxada	linia	columna	tipusCamp
0	\$tip:Carta \$tip:de \$tip:pago	Carta de pago	N	Y	N	N	N	0	0	tip
1	\$tip:Carta	Carta	Y	N	N	N	N	0	0	
2	\$tip:de	de	Y	N	N	N	N	0	11	
3	\$tip:pago	pago	Y	N	N	N	N	0	19	
4			N	N	N	N	N	1	0	

Generar documentos: extracción de palabras (ii)

```

<TextRegion id="TextRegion_1554219033768_699" custom="readingOrder {index:1;} structure {type:$pac;}">
  <Coords points="584,157 2583,180 2616,1387 2613,1968 624,1968"/>
  <TextLine id="TextLine_20_10" custom="readingOrder {index:0;} structure {type:$pac;}">
    <Coords points="2460,173 2066,193 1548,204 936,216 607,235 610,285 938,266 1549,254 2068,243 2462,223"/>
    <Baseline points="610,273 938,254 2068,231 2462,211"/>
    <TextEquiv>
      <Unicode>En treinta Abril mil setecientos, sesenta, y ocho, se</Unicode>
    </TextEquiv>
  </TextLine>
  <TextLine id="TextLine_19_10" custom="readingOrder {index:1;} structure {type:$pac;}">
    <Coords points="2511,278 2456,297 2027,294 645,333 647,383 2028,344 2465,347 2528,325"/>
    <Baseline points="647,371 2028,332 2463,335 2524,314"/>
    <TextEquiv>
      <Unicode>presentó en el Of.$^o$.Oficio de Hipote.$^s$.Hipotecas de esta ciudad de $top:Gerona</Unicode>
    </TextEquiv>
  </TextLine>
</TextRegion>

```

ordre	caracters	text	esborrada	fusionada	expandida	ajuntada	tatxada	linia	columna	tipusCamp
4	En	En	N	N	N	N	N	1	0	
5	treinta	treinta	N	N	N	N	N	1	3	
6	Abril	Abril	N	N	N	N	N	1	11	
7	mil	mil	N	N	N	N	N	1	17	
8	setecientos,	setecientos,	N	N	N	N	N	1	21	
9	sesenta,	sesenta,	N	N	N	N	N	1	34	
10	y	y	N	N	N	N	N	1	43	
11	ocho,	ocho,	N	N	N	N	N	1	45	
12	se	se	N	N	N	N	N	1	51	
13	presentó	presentó	N	N	N	N	N	2	0	
14	en	en	N	N	N	N	N	2	10	
15	el	el	N	N	N	N	N	2	13	
16	Of.\$^o\$.Oficio	Oficio	N	N	Y	N	N	2	16	
17	de	de	N	N	N	N	N	2	31	
18	Hipote.\$^s\$.Hipotecas	Hipotecas	N	N	Y	N	N	2	34	
19	de	de	N	N	N	N	N	2	56	
20	esta	esta	N	N	N	N	N	2	59	
21	ciudad	ciudad	N	N	N	N	N	2	64	
22	de	de	N	N	N	N	N	2	71	
23	\$top:Gerona	Gerona	N	N	N	N	N	2	74	top

Generar documentos: extracción de palabras (iii)

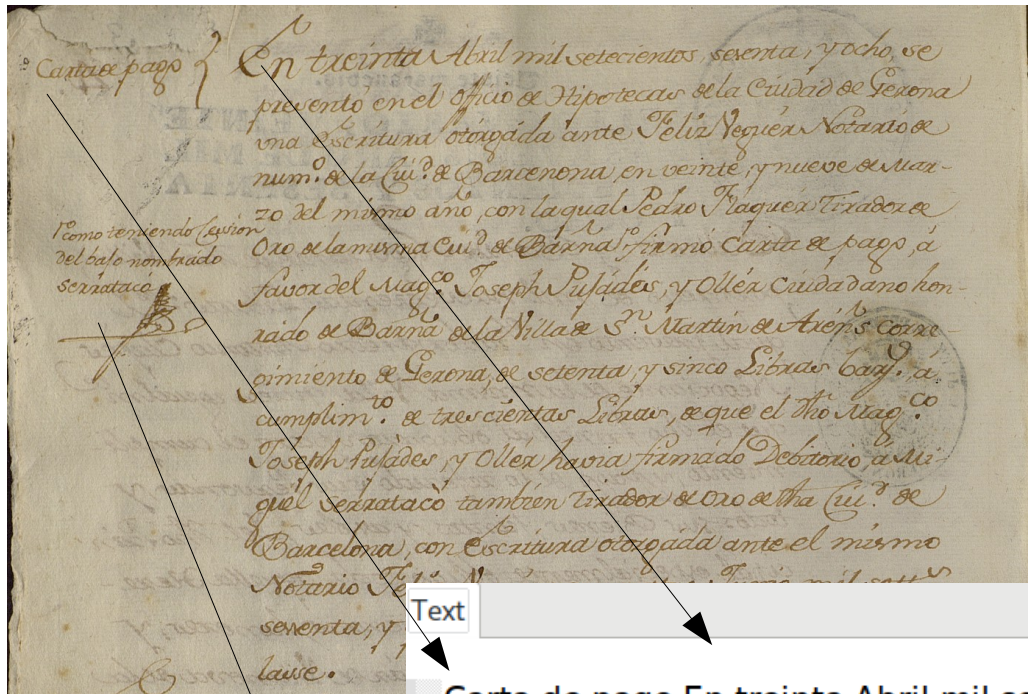
```

<TextLine id="TextLine_18_10" custom="readingOrder {index:2;} structure {type:$pac;}">
  <Coords points="2479,379 1805,404 794,427 678,416 673,466 792,477 1807,454 2481,429"/>
  <Baseline points="675,454 793,465 1807,442 2481,417"/>
  <TextEquiv>
    <Unicode>una Escritura otorgada ante $ant:Feliz $ant:Veguer $ofi:Notario de</Unicode>
  </TextEquiv>
</TextLine>
<TextLine id="TextLine_17_10" custom="readingOrder {index:3;} structure {type:$pac;}">
  <Coords points="2486,479 1203,515 1049,520 668,528 669,578 1051,570 1205,565 2488,529"/>
  <Baseline points="669,566 1051,558 1205,553 2488,517"/>
  <TextEquiv>
    <Unicode>num.$^o$.numero de la Ciu.$^d$.Ciudad de $top:Barcelona, en veinte, y nueve de Mar-$-</Unicode>
  </TextEquiv>

```

ordre	caracters	text	esborrada	fusionada	expandida	ajuntada	tatxada	linia	columna	tipusCamp
24	una	una	N	N	N	N	N	3	0	
25	Escritura	Escritura	N	N	N	N	N	3	4	
26	otorgada	otorgada	N	N	N	N	N	3	14	
27	ante	ante	N	N	N	N	N	3	23	
28	\$ant:Feliz \$ant:Veguer	Feliz Veguer	N	Y	N	N	N	3	28	ant
29	\$ant:Feliz	Feliz	Y	N	N	N	N	3	28	
30	\$ant:Veguer	Veguer	Y	N	N	N	N	3	39	
31	\$ofi:Notario	Notario	N	N	N	N	N	3	51	ofi
32	de	de	N	N	N	N	N	3	64	
33	num.\$^o\$.numero	numero	N	N	Y	N	N	4	0	
34	de	de	N	N	N	N	N	4	16	
35	la	la	N	N	N	N	N	4	19	
36	Ciu.\$^d\$.Ciudad	Ciudad	N	N	Y	N	N	4	22	
37	de	de	N	N	N	N	N	4	38	
38	\$top:Barcelona,	Barcelona,	N	N	N	N	N	4	41	top
39	en	en	N	N	N	N	N	4	57	
40	veinte,	veinte,	N	N	N	N	N	4	60	
41	y	y	N	N	N	N	N	4	68	
42	nueve	nueve	N	N	N	N	N	4	70	
43	de	de	N	N	N	N	N	4	76	
44	Mar-\$-		Y	N	N	N	N	4	79	
45	\$-zo		Y	N	N	N	N	5	0	
46	Mar-\$-\$-zo	Marzo	N	N	N	Y	N	4	79	
47	del	del	N	N	N	N	N	5	5	

Generación de documentos: texto «limpio» completo



Text

Carta de pago En treinta Abril mil setecientos, sesenta, y ocho, se presentó en el Oficio de Hipotecas de esta ciudad de Gerona una Escritura otorgada ante Feliz Veguer Notario de numero de la Ciudad de Barcelona, en veinte, y nueve de Marzo del mismo año, con la qual Pedro Flaquer tirador de Oro de la misma ciudad de Barcelona firmó Carta de pago, á favor del Magnífico Joseph Pujades y Oller Ciudadano honrado de Barcelona de la villa de San Martin de Areñs, corregimiento de Gerona, de setenta, y sinco libras barcelonesas, á cumplim^{to}. de trescientas libras, de que el dicho Magnífico Joseph Pujades, y Oller havia firmado Debitorio á Miguel Serratacá tambien tirador de oro de dicha ciudad de Barcelona, con Escritura otorgada ante el mismo Notario Felix Veguer, en quatro Junio mil Setecientos sesenta, y quatro, que quizo, y concintió se cancelasse.

=====
como teniendo cession del bajo nombrado \$ant:Serratacá [signatura]

Extracción automática de datos

Text

Arriendo En veinte y Seis Abril mil Settecientos sesenta, y se ha presentado en el Oficio de Hipotecas tecas de la ciudad de Gerona una escritura otorgada ante Geronimo Matheu Notario, y del Ilustre de la dicha en veinte y dos de dichos Mes, y año, con la qual el dicho Muy Ilustre testamente hizo Arriendo el que propria de grado, que se ha huerto de la Villa de Pasqua del Señor de este año, hasta el manso Santo del propio mil Setecientos sesenta y nueve para el habasto de la Carniceria publicias de esta misma ciudad, á favor de Juan Serra y Geli Comerciante Camino de la misma ciudad el refecto de diez libras y diez sueldos por cada la qual fueron fiador, á, Juan Moner Comerciante de la misma ciudad, quien igualmente el dicho principal obligaron de el cumplimiento, pago de la referido su Personas y todos sus Bienes junto, y assoles

Doc	página	tipus	Oficina Hipoteques				Nom notari	ofici notari	Ciutat notaria
1	170025120000001-0010	<u>donacion</u>	<u>Gerona</u>	25 <u>otorgada</u>	28 <u>ante</u>	29	<u>Pedro Pages</u>	30 <u>Notario,</u>	33 <u>Barcelona</u>
2	170025120000001-0010	PENDENT	<u>Gerona</u>	25 <u>otorgada</u>	33 <u>ante</u>	34	<u>Antonio Virell</u>	35 ?	70 <u>Canet</u>
3	170025120000001-0012	carta de pago	<u>Gerona</u>	27 <u>otorgada</u>	30 <u>ante</u>	31	<u>Antonio Vendrell del Turró</u>	32 <u>Notario</u>	37 <u>Canet de Mar</u>
4	170025120000001-0012	PENDENT	<u>Gerona</u>	26 <u>otorgada</u>	29 <u>ante</u>	30	<u>Miguel Prim</u>	31 <u>Ciudadano</u>	34 <u>Gerona</u>
5	170025120000001-0014	<u>arriendo</u>	<u>Gerona</u>	25 <u>otorgada</u>	28 <u>ante</u>	29	<u>Geronimo Matheu</u>	30 <u>Notario,</u>	33 <u>Gerona</u>
6	170025120000001-0014	<u>concordia y deutorio</u>	<u>Gerona</u>	24 <u>otorgada</u>	31 <u>ante</u>	32	<u>Francisco Lagrifa</u>	33 <u>Notario</u>	36 <u>Gerona</u>

Algunas magnitudes

- 4 libros
 - 9.000 páginas (8.949)
 - 35.000 regiones (35.197)
 - 11.000 **documentos** (11.262)
 - 325.000 líneas (325.528)
 - 3.000.000 palabras (2.990.687)
- Tiempo para leer páginas y alimentar la BD
 - Ejemplo Libro 0002 (2.355 páginas, 2.947 documentos, 9.388 regiones, 90.183 líneas, 809.888 palabras)

```
** Temps mesurats ****  
Llegir fitxers: 3.46794 segons  
Generar documents: 0.596446 segons  
Processant documents generats: 1.99899 segons  
Escriure a la Base de Dades: 187.47 segons  
*****
```

Algunas magnitudes (ii)

- Extraer notarios, ciudad notaría, etc...
 - Menos de un minuto por libro
 - Se extrae automáticamente nombre notario y notaría en aproximadamente un 99% de los documentos.

Futuro cercano

- Afinar programa generación de documentos
 - Quedan algunos flecos en las palabras cortadas a final de línea y en la detección de regiones tipo o regiones nota al margen.
- Extracción automática de datos en tipologías de documento concretos (normalización, personas...)
- Explorar técnicas de *data-mining* sobre texto completo para intentar superar las imprecisiones inherentes al transcriptor.
- Desarrollo de una interfaz *historian-like* para consultar y corregir/validar el contenido de la base de datos.

Bonus slide...

- En la reunión del proyecto de investigación de ayer se habló de estudiar el papel de la mujer y los niños en los cambios sociales... Y en concreto de las viudas.
- ... hay viudas en el registro de hipotecas?
 - Buscamos en la Base de Datos en cuantos documentos de cada tipo aparece «viuda» en el texto completo (**consulta libre**), agrupándolo por tipo de documento (**consulta dato normalizado**)

```

SELECT tdl.tipus AS `Tipus`,
       COUNT(IF(textNet LIKE '%viuda%',1,NULL)) AS `Con Viuda`,
       ROUND(COUNT(IF(textNet LIKE '%viuda%',1,NULL))/COUNT(*)*100,2) AS `% Con Viuda`
FROM Documents JOIN TipusDocumentsLiterals tdl ON Documents.tipus=tdl.codi
GROUP BY tdl.tipus
HAVING COUNT(IF(textNet LIKE '%viuda%',1,NULL))>=10
ORDER BY `Con Viuda` DESC;

```

Tipus	Con Viuda	% Con Viuda
Capitulos Matrimoniales	364	31,52
venta	250	11,64
censal	94	7,35
establecimiento	92	8,65
PENDIENT	74	12,4
debitorio	70	8,93
arriendo	67	8,28
acollamiento	62	8,95
donacion	59	22,61
inventario	55	47,83
testamento	54	26,09
carta de pago	52	14,9
concordia	47	22,82
luicion	35	13,41
reventa	26	10,2
fundacion	19	27,54
reduccion	13	44,83

Jordi.Regincos@udg.edu