DAS–2022

La Rochelle, May 2022

Efective Crowdsourcing in the EDT Project with Probabilistic Indexes (PrIx's)

Joan Andreu Sánchez, Enrique Vidal, Vicente Bosch

tranSkriptorium *tS*

Presentation available at: https://www.prhlt.upv.es/~evidal/tmp/edtTsPresenDAS22.pdf

Index

○ 1 *The EDT project and Manuscript Collections* ▷ 2

2 Probabilistic Indexing (PrIx) ▷ 9

3 Initial Results on EDT Datasets ▷ 12

4 Crowdsourcing Production of Additional GT ▷ 14

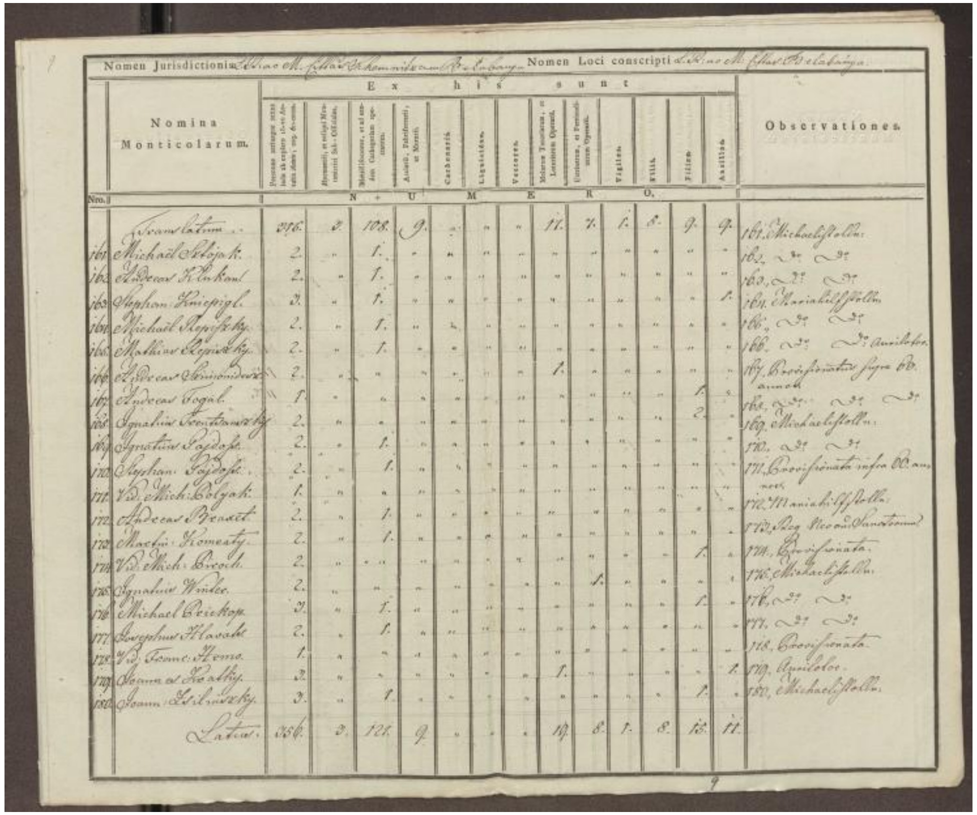
5 Model Retraining and Final Results ▷ 16

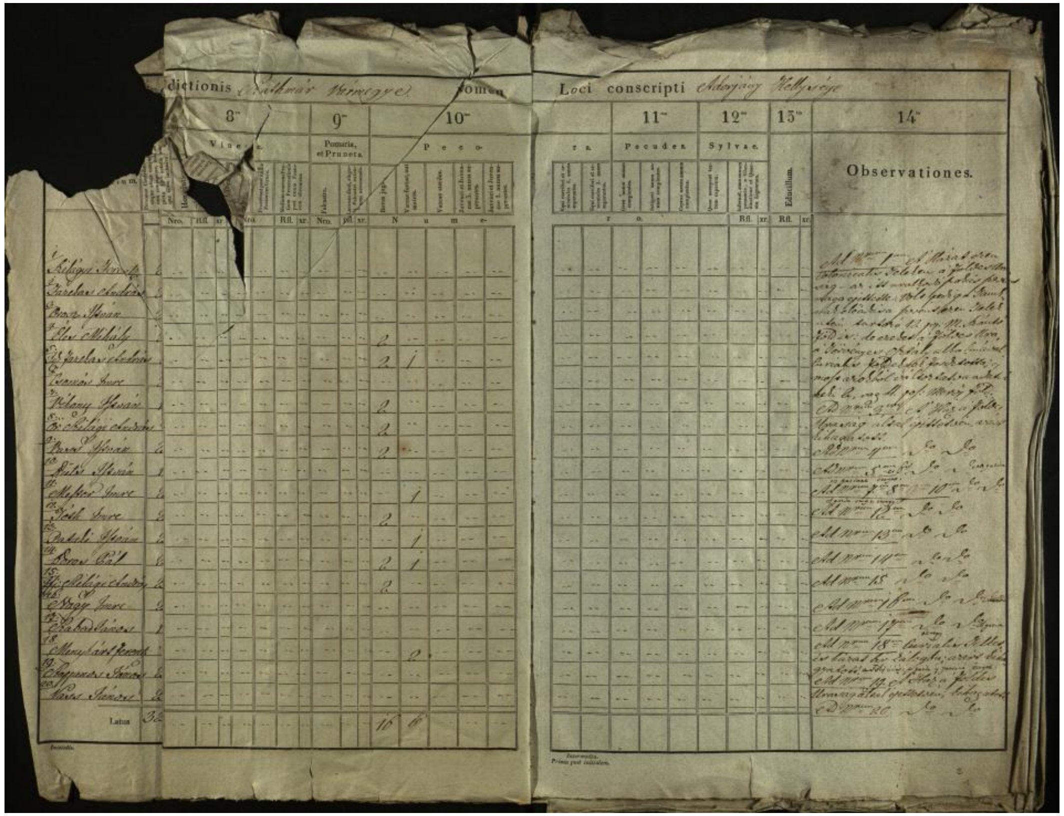
6 Conclusion ▷ 18

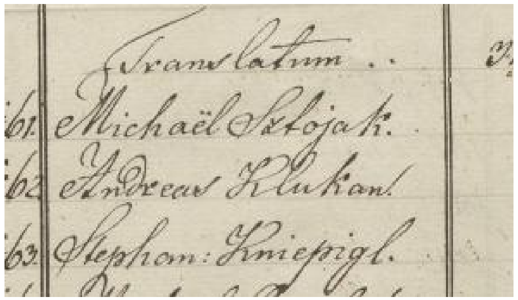
The EDT Project

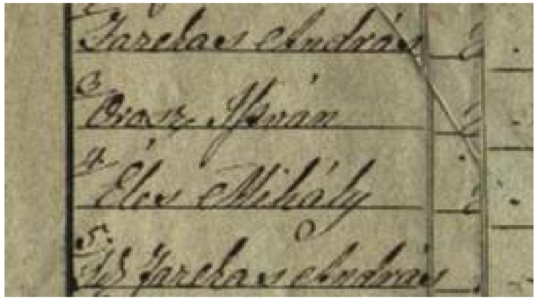
- The main aims of the Europen project *“European Digital Treasures* (EDT): *Management of centennial archives in the 21st century”* were to provide major visibility, outreach and use of digital heritage.
- Manuscripts hosted by National Archives of five countries were considered: *Hungary, Norway, Portugal, Spain* and *Malta*.
- The *Probabilisitic Indexing* (PrIx) framework was adopted to meet some of the EDT goals.
- PrIx is a Machine Learning technology and therefore needs adequate amounts of manually transcribed images to train the required optical and language models.
- In this work we explore new *crowdsourcing* techniques based on a PrIx platform to produce adequate training data in a cost-effective way.

EDT-Hungary Manuscripts



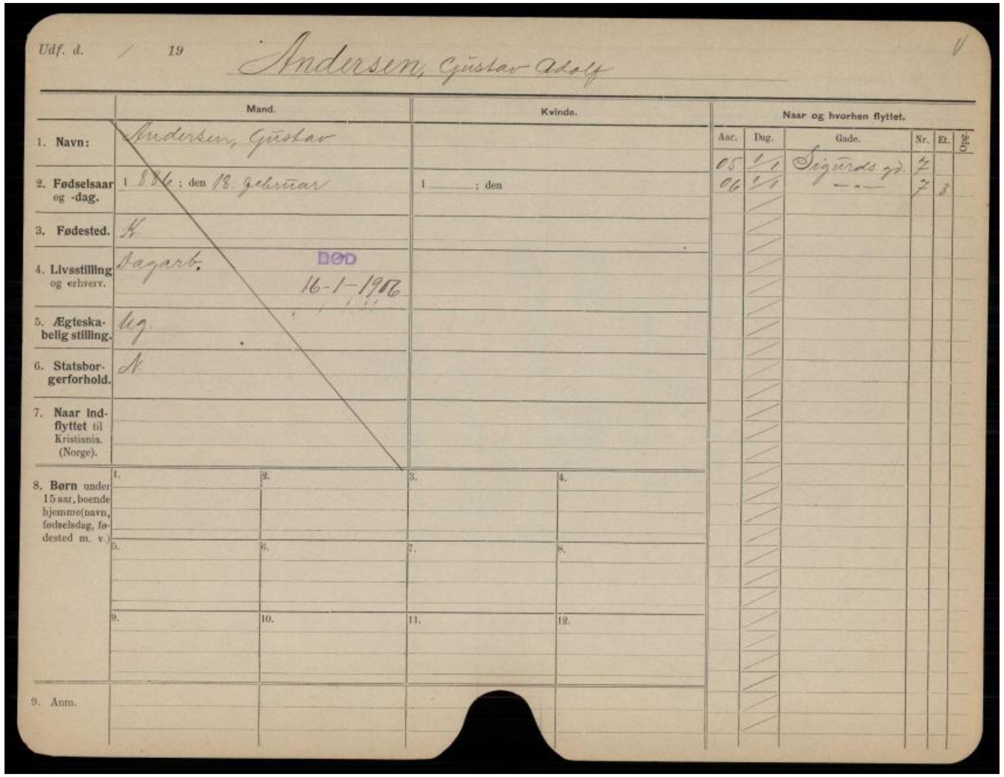
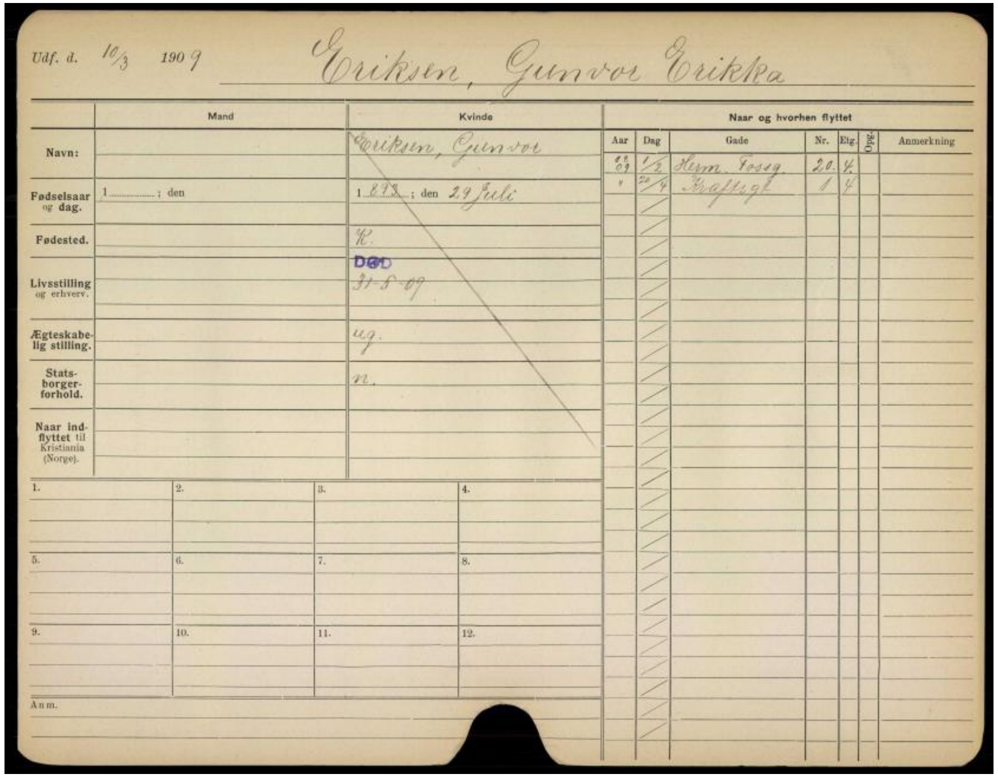
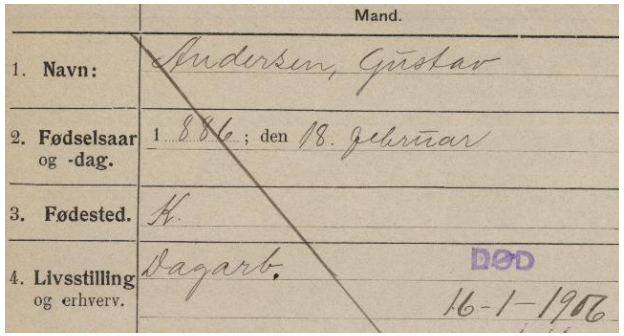


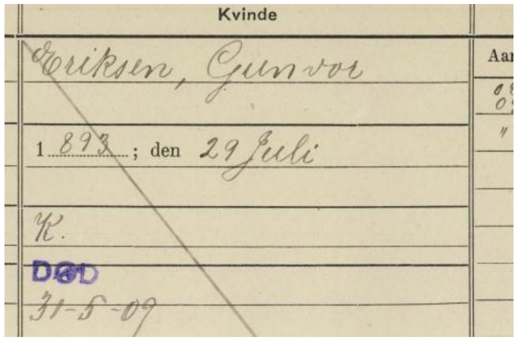




Only the handwritten names in the left column of each table are of interest. Abbreviations are plenty, but the system is expected to index only the expanded and modern versions of these words.

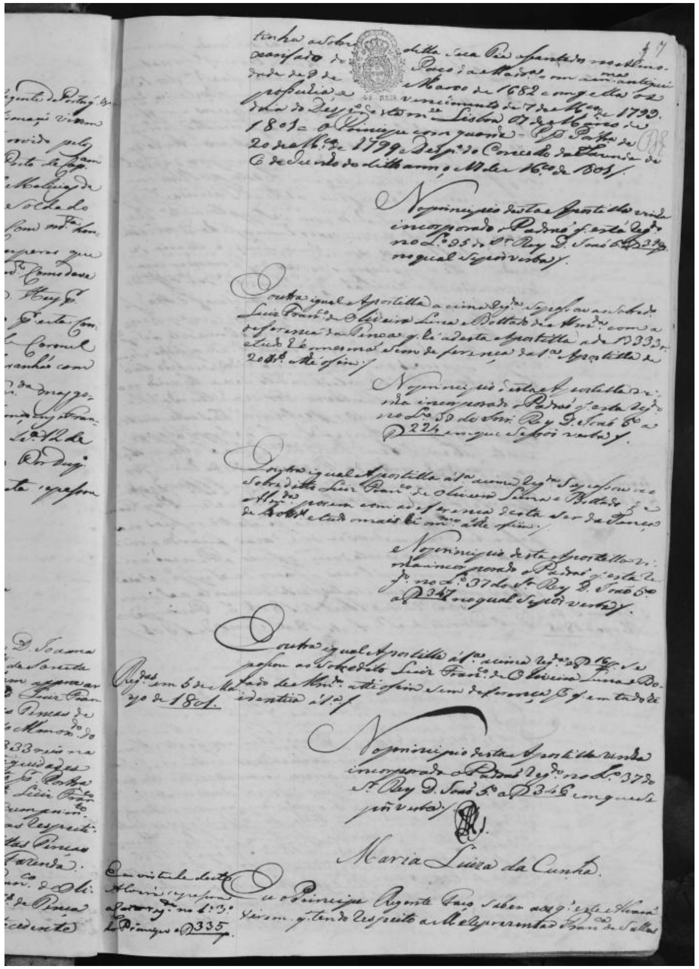
EDT-Norway Manuscripts

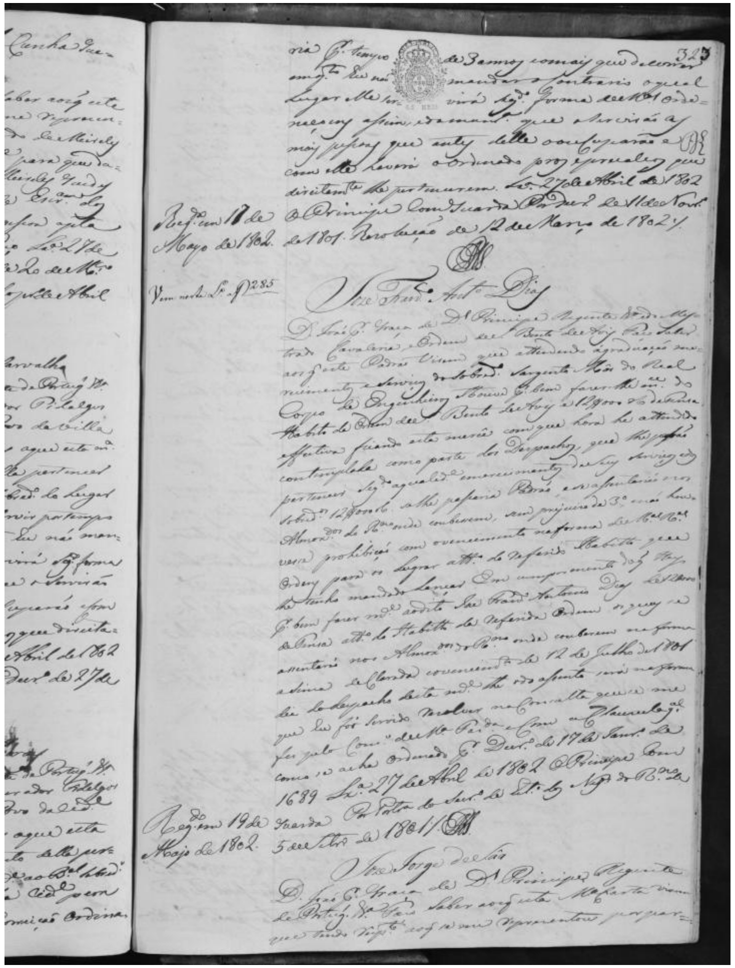




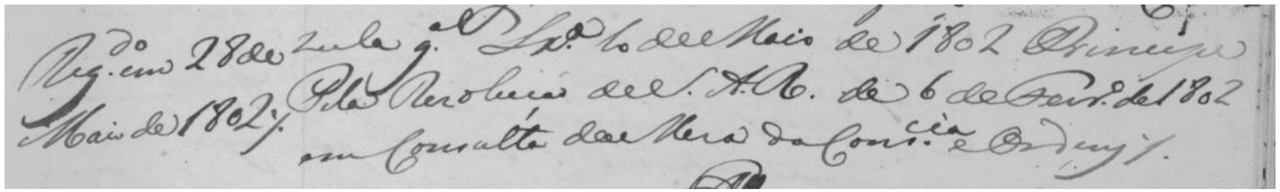


Both printed and handwritten fields are of interest. Note the variable, mostly random position of important data headed by the stamped abbreviation “DØD” (Date Of Death).

EDT-Portugal Manuscripts

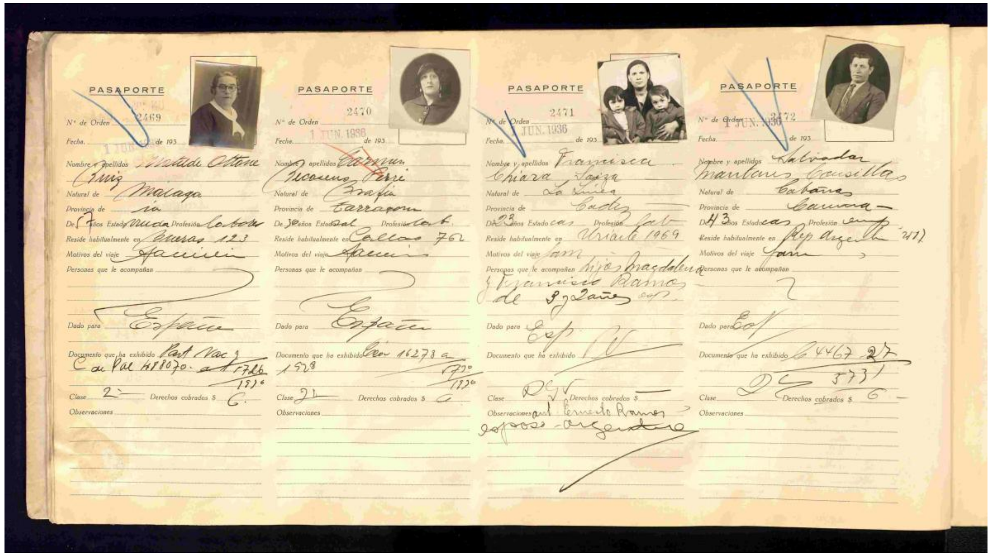


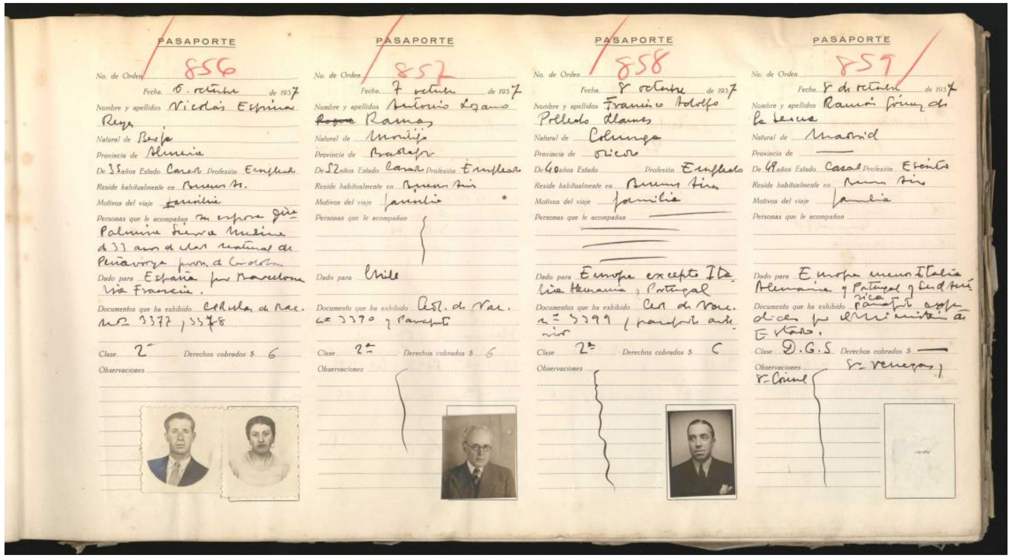




Most of the writing is heavyly abbreviated freehand body text, with important layout variations.

EDT-Spain Manuscripts

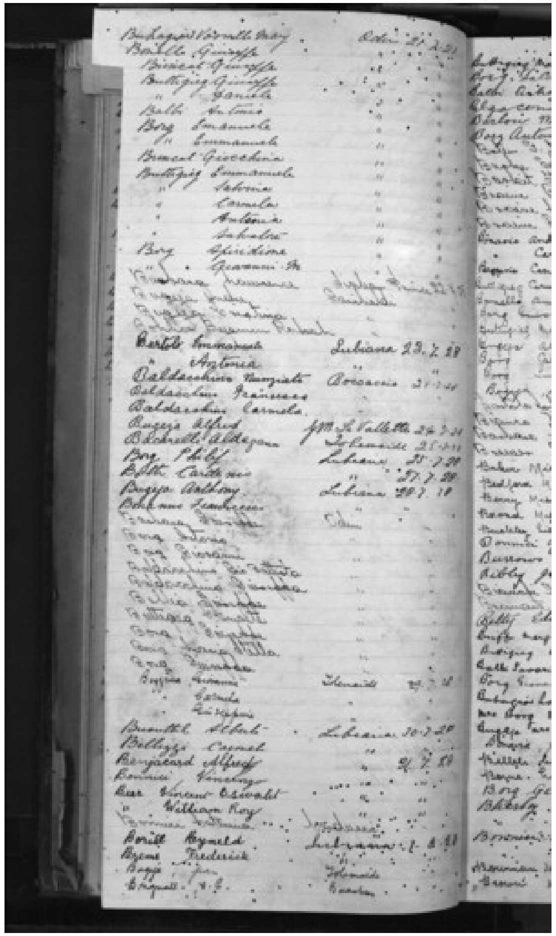
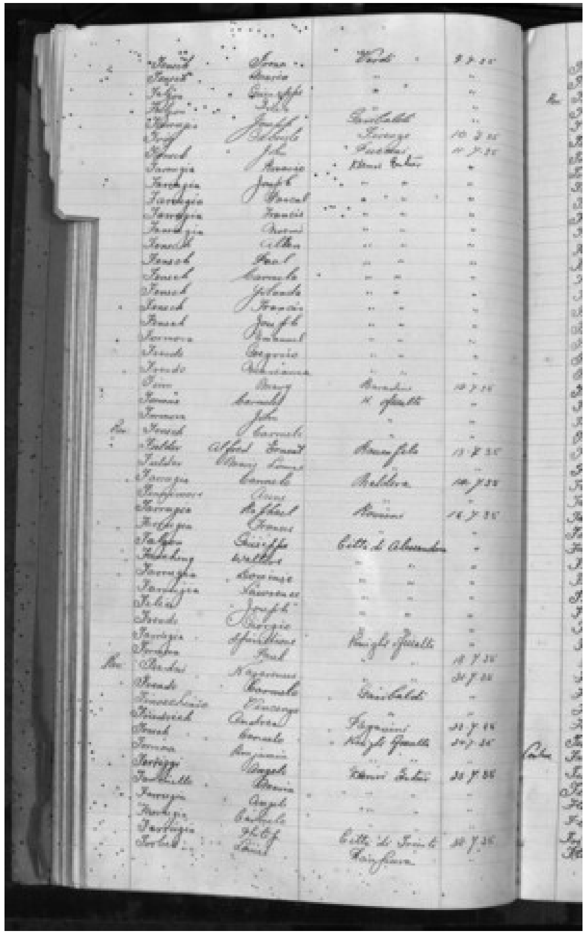
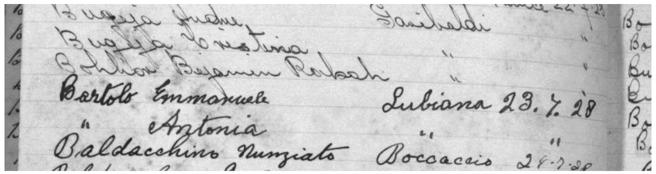
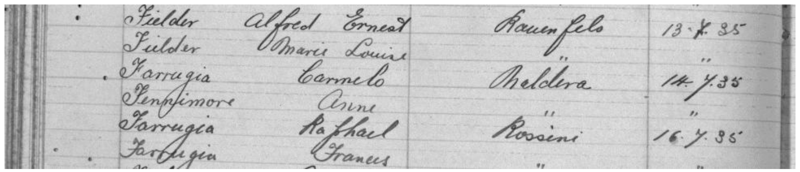





Nº de Orden
Fecha. de 193
Nombre y apellidos Amalia Fernandez Alvarez
Natural de S. Martin
Provincia de Lugo
De 27 años Estado solt Profesión lab
Reside habitualmente en Callao 1332
Motivos del viaje
Nº de Orden
Fecha. de 193
Nombre y apellidos Mauricio Serrano Rico
Natural de Yanguas
Provincia de Soria
De 22 años Estado solt Profesión emp
Reside habitualmente en Caseros 4706
Motivos del viaje

Each image contains four visa records (forms), with a fairly regular layout. The handwritten parts contain thousands of scrawled, often invented abbreviations, but the system is expected to index only the expanded and modern versions of these abbreviations.

EDT-Malta Manuscripts

The four handwritten columns in each page are of interest. But only the text in each page is relevant (not the parts belonging to adjacent pages). Note the severe warping exhibited by most images.

Abbreviations and Semantic Tagging

In some collections, many words, such as *"Franciscus"*, can are spelled in many abbreviated and/or unconventional forms; but the system is expected to provide the same, unique hypothesis (*"Franciscus"* in this example) for all the spellings.

Token	Frequency	Token	Frequency	Token	Frequency
Fr	1	**Fran.**	104	**Francisc**	3
Fr.	46	**Franc**	8	**Francisci**	1
Fr:	19	**Franc.**	7	**Franciscus**	47
Fra.	2	**Franc:**	5	**Frank**	3
Fran	10	**Francis**	2	**Franz**	4

In other collections words are tagged according to their *"semantic"* roles; for instance "Juan" can be a *given name*, a *surname* or *place name*. For EDT-Spain, the folloing tagging was adopted:

<print>	**<surname>**	**<civilstate>**	**<job>**
<date>	**<state>**	**<residence>**	**<age>**
<gname>	**<country>**	**<place>**	

Examples of GT transcripts:

Nombre<print> y<print> Apellidos:<print> Juan<name> Juan<surname>

Natural<print> de:<print> San<place> Juan<place>

Index

1 The EDT project and Manuscript Collections ▷ 2

○ 2 *Probabilistic Indexing (PrIx)* ▷ 9

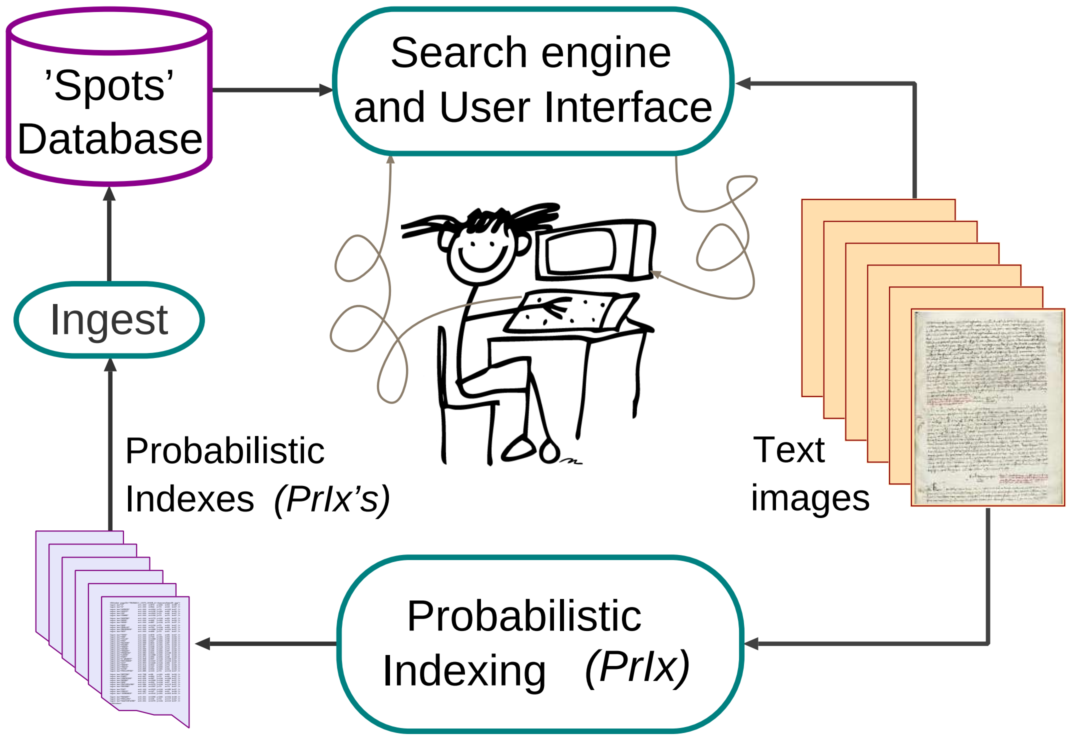
3 Initial Results on EDT Datasets ▷ 12

4 Crowdsourcing Production of Additional GT ▷ 14

5 Model Retraining and Final Results ▷ 16

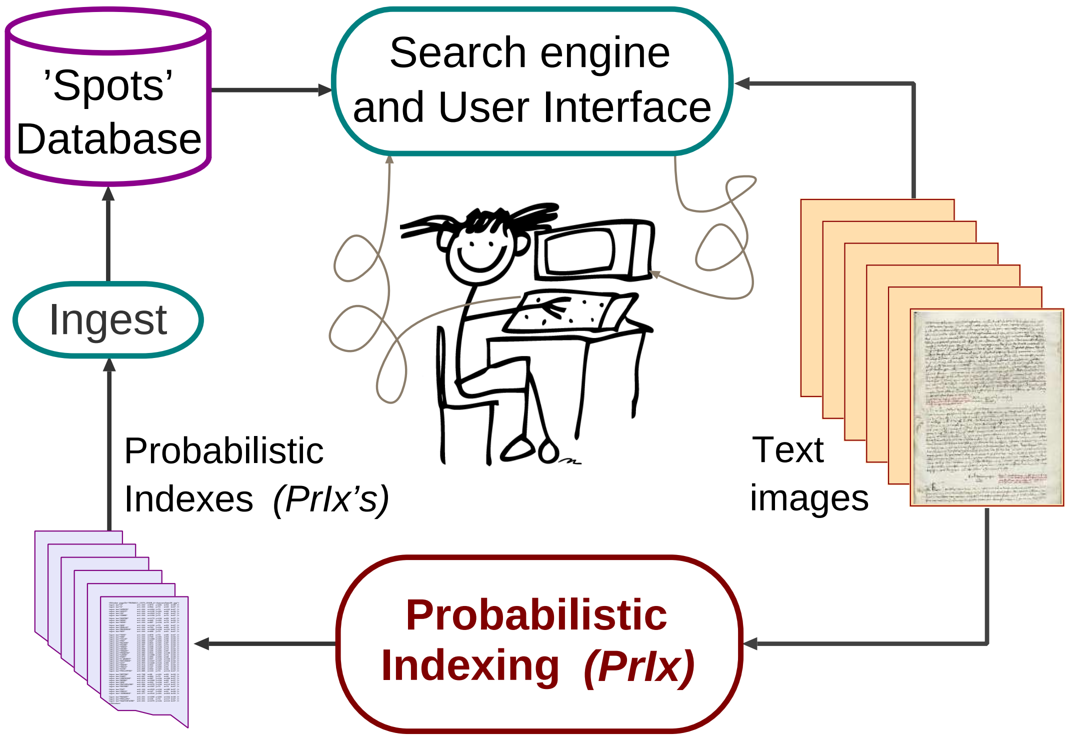
6 Conclusion ▷ 18

Probabilistic Text Image Indexing and Search: System Diagram



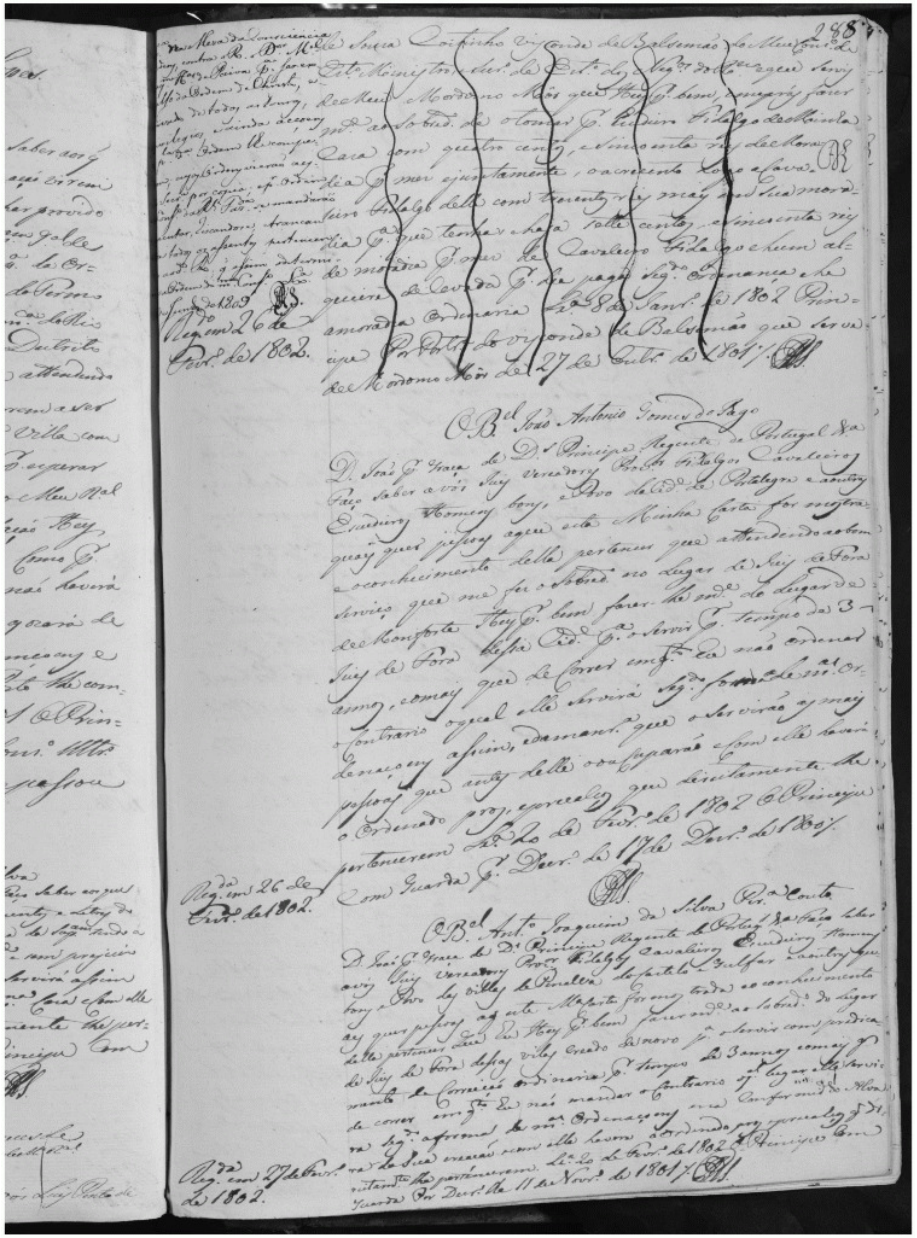
- *"Probabilistic Indexing (PrIx)"*: Off-line pre-computation of PrIx's
- *"Ingest"*: Off-line creation of the actual database. Typically a simple and computationally cheap process
- *'Search engine and GUI"*: On-line user query analisys, find the requested information and present the retrived images. Short response times needed.

Probabilistic Text Image Indexing and Search: System Diagram



- *"Probabilistic Indexing (PrIx)"*: Off-line pre-computation of PrIx's
  - It needs heavy (*off-line*) computing – but it allows extremely fast on-line query responses, even for huge manuscript collections.
  - Most complex component. Based on *contextual word (or char string) recognition*, which require models *trained* from transcribed images (as in HTR)

A Page Image from EDT-Portugal



A Real Example of PrIx

Spots marked in colors according to their Relevance Probabilities: low=red, high=green.

Index

1 The EDT project and Manuscript Collections ▷ 2

2 Probabilistic Indexing (PrIx) ▷ 9

○ 3 *Initial Results on EDT Datasets* ▷ 12

4 Crowdsourcing Production of Additional GT ▷ 14

5 Model Retraining and Final Results ▷ 16

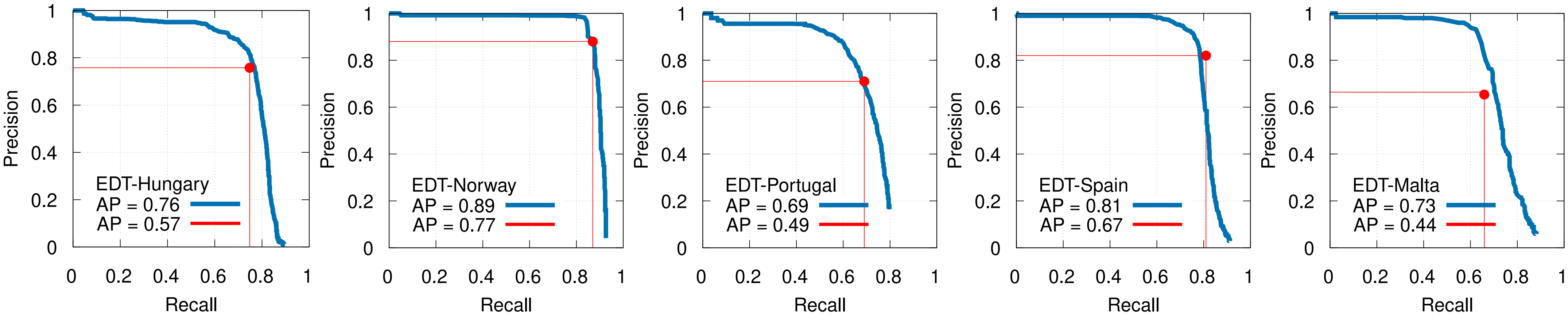
6 Conclusion ▷ 18

The EDT Initial GT-Annotated Datasets and Results

Archive experts selected adequate images of each collection and produced initial GT

DATASETS	EDT-Hung	EDT-Norw	EDT-Port	EDT-Spain	EDT-Malta
Running words	17 687	12 978	13 285	27 214	11 481
Train + Validation	16 284	12 650	12 469	26 157	11 033
Test	1 403	328	816	1 057	448

RESULTS		EDT-Hung	EDT-Norw	EDT-Port	EDT-Spain	EDT-Malta
HTR	WER (%)	25.7	13.4	33.9	20.6	34.8
	AP	0.57	0.77	0.49	0.67	0.44
PrIx	AP	0.76	0.89	0.69	0.81	0.73



Recall-Precision curves for the EDT datasets. PrIx in blue, HTR in red

Index

1 The EDT project and Manuscript Collections ▷ 2

2 Probabilistic Indexing (PrIx) ▷ 9

3 Initial Results on EDT Datasets ▷ 12

○ 4 *Crowdsourcing Production of Additional GT* ▷ 14

5 Model Retraining and Final Results ▷ 16

6 Conclusion ▷ 18

Crowdsourcing Production of Additional GT for Model Re-training

- For each *complete* collection, the models trained with the Initial GT were used to compute PrIx's which were ingested into the corresponding *search platforms*,
- *Validating* and *editing* capabilities were added to these search platforms to allow croudsourcing production of additional GT,
- Archives recruited volunteers who were instructed to visit appropriate parts of each collection and *validate* and/or *edit* spots as needed,
- Validated or edited word spots were used, along with surrounding spots, to assamble reliable training text-lines to *re-train* the *PrIx optical & language models*.

		EDT-Hung	EDT-Norw	EDT-Port	EDT-Spain	EDT-Malta
PrIndxd	Images	36 396	11 837	747	811	12 908
	Running Words*	1 408 290	528 019	321 218	199 520	2 641 440
	Lexicon Size*	143 508	22 806	14 675	13 526	148 115
CrwdSrc	Visited images	6 942	2 390	152	603	391
	Reviwed Word Spots	303 036	38 402	26 913	110 331	78 834
	Edited Words	131 821	1 612	7 293	24 824	5 597
DataSet†	Running Words	272 904	50 655	31 662	27 214	60 876
	Training + Validation	271 501	50 327	30 846	26 157	60 428
	Test (same as Initial)	1 403	328	816	1 057	448

*Running words and lexicon size estimated as in *Toselli et al., ICDAR-2019*; †Including the initial GT

Index

1 The EDT project and Manuscript Collections ▷ 2

2 Probabilistic Indexing (PrIx) ▷ 9

3 Initial Results on EDT Datasets ▷ 12

4 Crowdsourcing Production of Additional GT ▷ 14

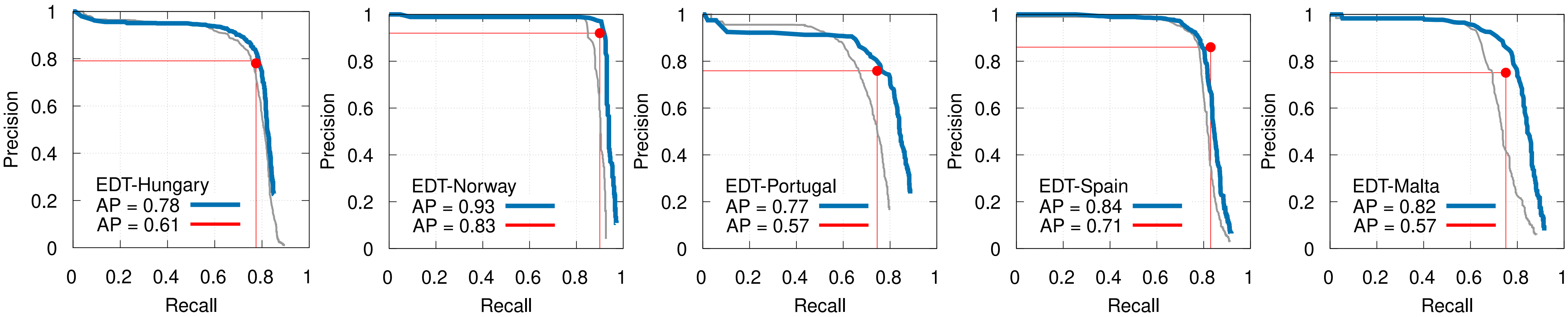
◦ 5 *Model Retraining and Final Results* ▷ 16

6 Conclusion ▷ 18

Final Results After Model Retraining

HTR WER (%) and AP and PrIx AP obtained for each collections after re-training with crowdsourcing GT data, and relative improvements (%).

		EDT-Hung	EDT-Norw	EDT-Port	EDT-Spain	EDT-Malta
HTR	WER	24.3	10.4	28.4	18.7	25.9
	AP	0.61	0.83	0.57	0.71	0.57
	Improvement	7.0	7.8	16.3	6.0	29.5
PrIx	AP	0.78	0.93	0.77	0.84	0.82
	Improvement	2.6	4.5	11.6	3.7	12.3



R-P curves after re-training with crowdsourcing GT data. In grey R-P curves before re-training.

Index

1 The EDT project and Manuscript Collections ▷ 2

2 Probabilistic Indexing (PrIx) ▷ 9

3 Initial Results on EDT Datasets ▷ 12

4 Crowdsourcing Production of Additional GT ▷ 14

5 Model Retraining and Final Results ▷ 16

○ 6 *Conclusion* ▷ 18

Conclusion

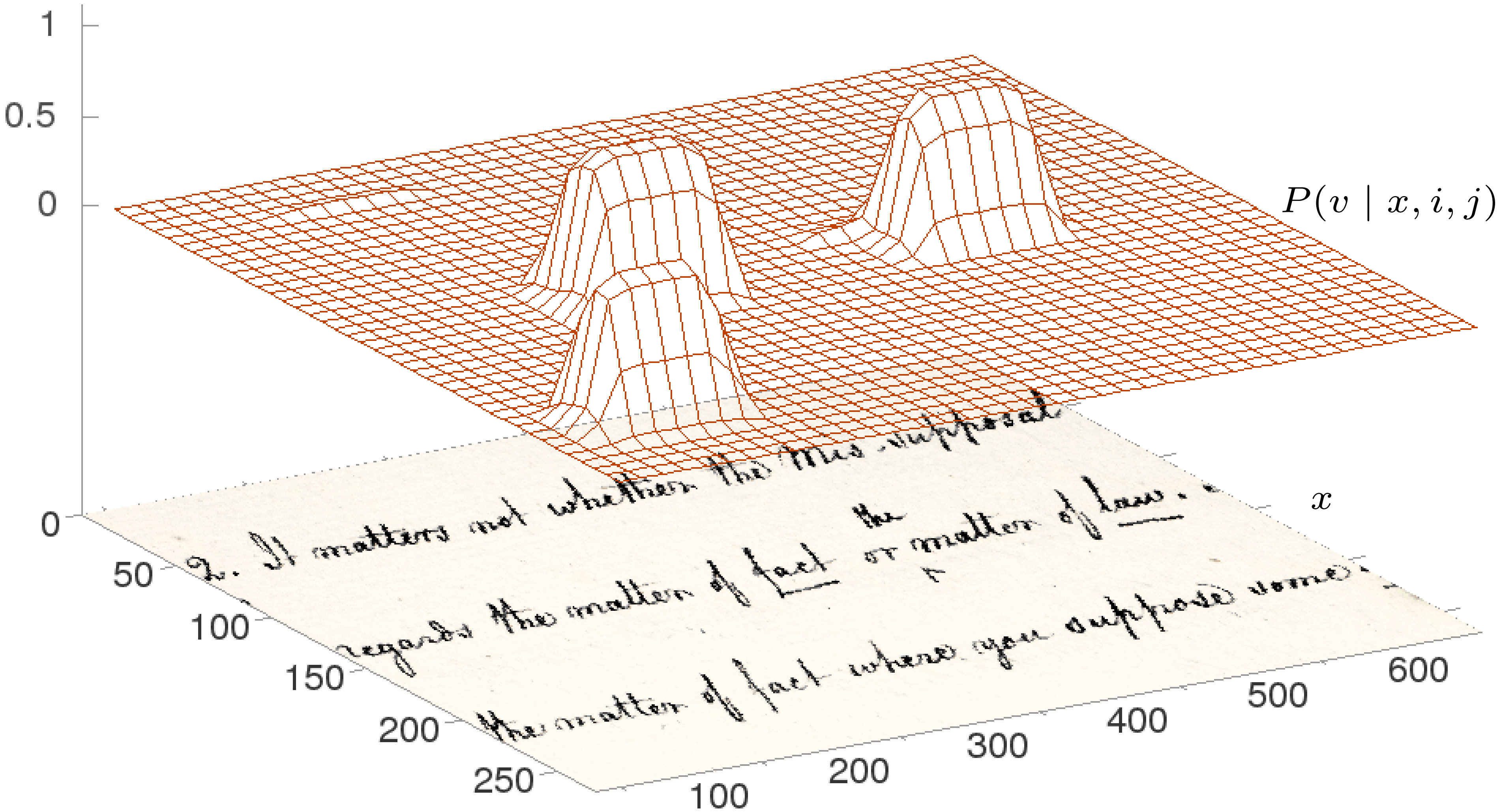
- *Probabilistic Indexing* (PrIx) is a *mature technology* which allows very effective *free-text searching* in large collections of handwritten text images
- It has been very *successfully* applied to all EDT manuscript collections with very *different characteristics* and challenges
- Effective capabilities to collect additional GT training data through croudsourcing has been easily added to PrIx platforms initially aimed at only at textual information search
- Re-training the PrIx models with the additional data has resulted in moderate but significant performance improvements

PrIx's are ***not*** transcripts, but they provide a much more *robust representations* of *textual contents* of images which enable very effective search for textual information in large collections of essentially untranscribed images

Thanks for your attention!

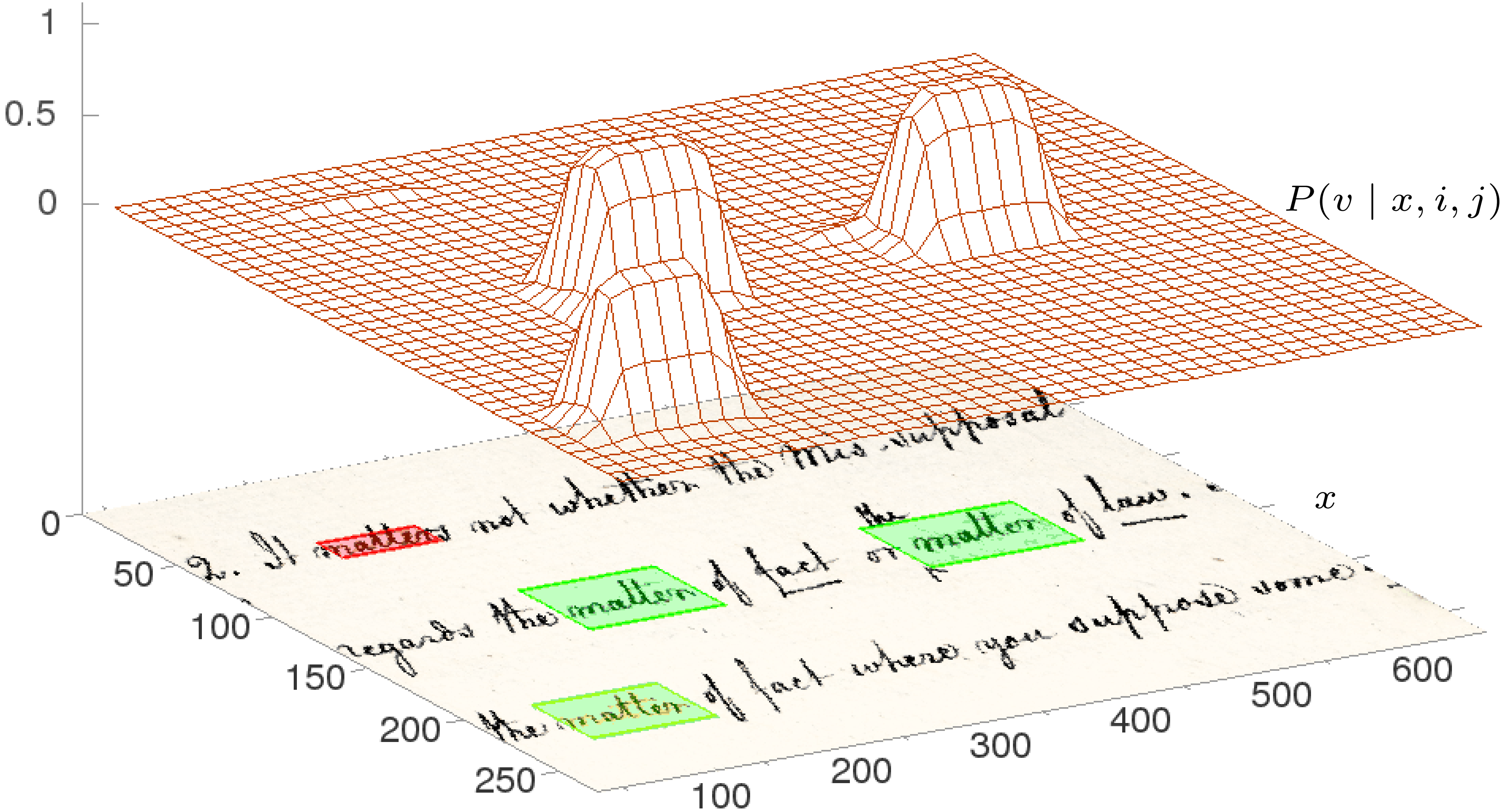
(additional details below)

PrIx fundamentals: Pixel-level Posteriorgram



Pixel-level posterior probabilities $P(v \mid x, i, j)$ for a text image x and word $v =$"**`matter`**", computed using an *accurate, contextual* (n-gram based) *word classifier*. This helped to achieve very good posteriors: low in a region of x around $(i=100, j=60)$, where a very similar (but *different*) word, "**`matters`**", is written; high for the other three correct words.

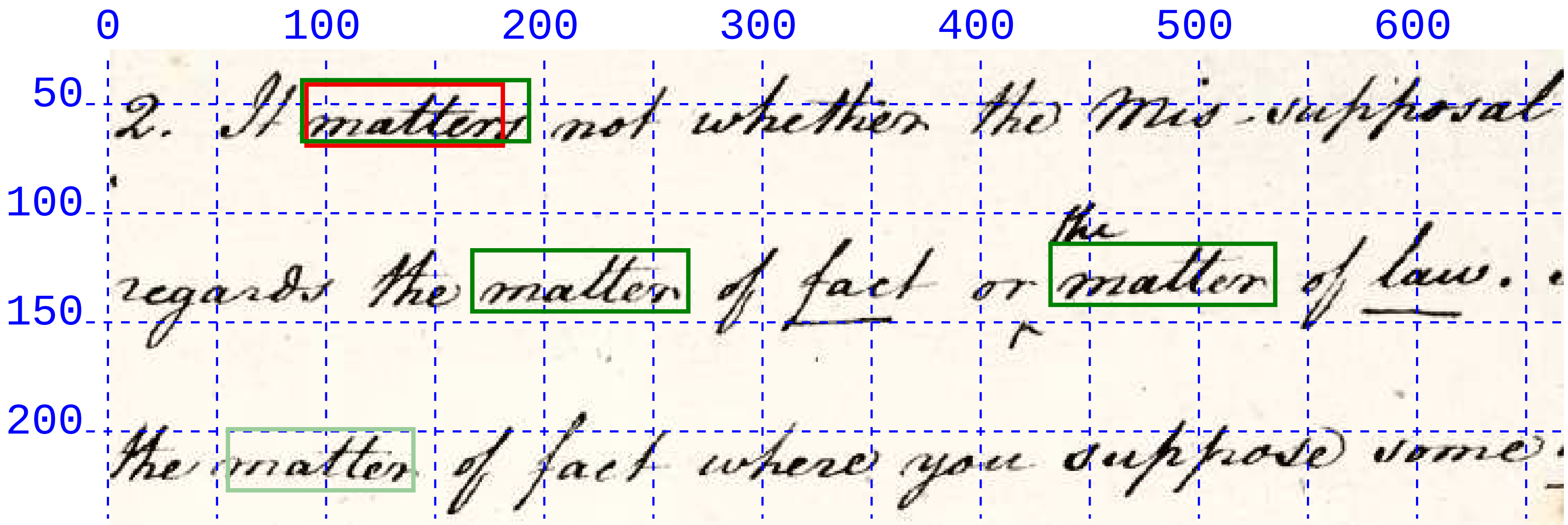
Pixel-level Posteriorgram: Probabilistic Word Indexing (PrIx)



Directly computing and using a full pixel-level posteriorgram would entail a formidable computational load and would require prohibitive amounts of indexing storage.

But, for each word, image region *relevance probabilities* and *locations* are easily derived from the Posteriorgram – and used to probabilistically index the word in an efficient way.

Probabilistic Index: Example



\# pageID="Bentham-071-021-002-part"
\# keyword relPrb bounding box
\#

keyword	relPrb		bounding box		
2	0.929	1	36	20	31
21	0.064	1	36	24	31
IT	0.982	33	36	27	31
IF	0.012	33	36	26	31
MATTERS	**0.998**	**76**	**35**	**104**	**31**
MATTER	**0.011**	**77**	**36**	**93**	**31**
NOT	0.999	216	36	47	31
WHETHER	1.000	256	36	99	31
THE	0.997	389	36	33	31
MIS-SUPPOSAL	1.000	455	36	193	31
THE	0.927	430	88	30	31
LHE	0.056	434	88	25	31
...	...	...		...	

REGARDS	0.857	5	115	84	31
UGARDS	0.138	5	115	80	31
THE	0.993	110	115	43	31
MATTER	**0.998**	**160**	**115**	**93**	**31**
OF	0.996	271	115	23	31
FACT	0.999	306	115	49	31
OR	0.973	377	115	37	31
ON	0.021	377	115	42	31
MATTER	**0.990**	**425**	**116**	**100**	**31**
OF	0.995	542	115	25	31
LAM	0.407	575	115	30	31
BIMR	0.175	575	115	55	31
...	...	...		...	
LAW	0.032	575	115	36	31
TAUE	0.031	575	115	55	31
...	...	...		...	
LANE	0.012	575	115	59	31

THE	0.990	1	198	28	31
MATTER	**0.934**	**61**	**198**	**64**	**31**
OF	0.988	141	198	28	31
FAST	0.367	182	198	62	31
FAR	0.186	182	198	36	31
...	...	...		...	
FACT	0.017	182	198	46	31
AS	0.142	200	198	29	31
HAE	0.022	200	198	29	31
WHERE	0.992	255	198	90	31
YOU	0.761	365	198	45	31
YOW	0.030	365	198	45	31
GOUS	0.064	372	198	47	31
SUPPOSE	0.975	429	198	120	31
SUPFROSE	0.024	429	198	125	31
SOME	0.834	570	198	78	31
SONER	0.016	576	198	83	31
OME	0.109	580	198	65	31
ME	0.022	620	198	22	31

Spots for **MATTER** and **MATTERS** marked in colors according to their Relevance Probabilities: low=red, high=green.

Probabilistic Indices are *NOT* Transcripts

AUTOMATIC TRANSCRIPTION (HTR)	PROBABILISTIC INDEXING (PRIX)
Generally comes after Layout Analysis	Is generally Layout-agnostic
Strictly needs carefully detected lines	Line detection helps, but only if accurate
The output is a best, unique (delicate!) text interpretation of the given image according to the models used	For the same models, the output is a robust probability distribution of words with their possitions in the images
The output is aimed to be in reading order (but this is seldom achieved)	In general, Probabilistic Indexing is reading-order agnostic
Provides plaintext output. If accuracy is high, it can be directly used in many applications	In its basic form, does not provide any text output; only images marked with word-sized bounding boxes
Usually yields only fixed and comparatively low precission-recall performance for the given trained models	Allows flexible, user-controled precision-recall tradeoffs and search preformance is generally much better for the same trained models

Probabilistic Indices are *NOT* Transcripts

AUTOMATIC TRANSCRIPTION (HTR)	PROBABILISTIC INDEXING (PRIX)
Generally comes after Layout Analysis	Is generally Layout-agnostic
Strictly needs carefully detected lines	Line detection helps, but only if accurate
The output is a best, unique (delicate!) text interpretation of the given image according to the models used	For the same models, the output is a robust probability distribution of words with their possitions in the images
The output is aimed to be in reading order (but this is seldom achieved)	In general, Probabilistic Indexing is reading-order agnostic
Provides plaintext output. If accuracy is high, it can be directly used in many applications	In its basic form, does not provide any text output; only images marked with word-sized bounding boxes
Usually yields only fixed and comparatively low precission-recall performance for the given trained models	Allows flexible, user-controled precision-recall tradeoffs and search preformance is generally much better for the same trained models

Probabilistic Indexing provides very effective search solutions where Automatic Transcription fails!

Beyond Using PrIx for Basic Information Searching

- Wild cards, approximate ("fuzzy") spelling and abbreviation expansion
- Boolean, proximity-AND and word-sequence queries

Available by by default in all the PrIx search demonstrators. See:
`http://prhlt-carabela.prhlt.upv.es/PrIxDemos`

Moreover:

- Find Hyphenated Words using just entiere-word queries
  `http://prhlt-kws.prhlt.upv.es/fcr-hyp`
- Search for Melodic Patterns in handwritten music notation
  `http://prhlt-carabela.prhlt.upv.es/music`
- Data-Base-like Information Retrieval from handwriten tables
  `http://prhlt-carabela.prhlt.upv.es/passauTab`
- Handle huge manuscript collections, *over one million pages*
  `http://prhlt-kws.prhlt.upv.es/fcr`

New:

- Text Anaytics and Big-Data Statistical Information Extraction
- Content-based Image Document Classification

Beyond PrIx Searching: Basic Text Analytics

- PrIx's are just text files containing lists of *spots*. Each spot provides:

Folder	Page	Word	Probability	Bounding-Box
5	14	MADRID	0.998414	936 585 1273 658
5	43	MARIA	0.717130	3746 910 3948 978
...	...	...	...	...

- The probability of a spot can be interpreted as the expectation that the corresponding word is actually written in the image BBx $\Rightarrow$ probabilities can be just added up to compute statistical estimates of frequencies of ocurrence.

- PrIx's can be loaded into a spreadsheet or any other DB tool and simple calculations can be made to extract intersting (statistical) information.

  Examples:
  - Estimate the frequency of occurrence of words
  - Word occurrences in specified page images or folders
  - Number of images or documents which may contain a given word
  - Word occurrences in the context of other words
  - Zipf curves and vocabulary sizes
  - Etc. . .

Estimating Word and Document Frequencies

Reminder: For an image region x and character string v, PrIx provides the (relevance) probability that v is written in x, $P(R \mid x, v)$.

The expected number of words written in x is computed as:

$$E[n(x)] \;\approx\; \sum_{v} P(R \mid x, v)$$

Thus, the number of running words of a document X, or in a full collection $\mathcal{X}$, is estimated as the sum of Relevance Probabilities of all the indexed spots for X, or $\mathcal{X}$.

Similarly, the frequencies of use of a specific word v in X, or in $\mathcal{X}$, are estimated as:

$$E[n(v,X)] \;\approx\; \sum_{x \sqsubseteq X} P(R \mid X, v)\,; \qquad E[n(v,\mathcal{X})] \;\approx\; \sum_{X \in \mathcal{X}} E[n(v,X)]$$

Fianally, the expected number of documents in $\mathcal{X}$ that contain the word v is:

$$E[m(v,\mathcal{X})] \;\approx\; \sum_{X \in \mathcal{X}} \max_{x \sqsubseteq X} P(R \mid x, v)$$

Beyond PrIx Searching: Statistical Information Extraction

- Taking advantage of word contexts, PrIx models can be directly trained to produce pseudo-word hypotheses with *Named Entity or "Semantic" tags*

- This allows distinguishing words depending on their semantic roles; for instance: **`SMITH!surname`**, **`SMITH!job`**

- This way, statistical estimates not only for plain words, but also for semantic categories, can be easily computed

Applied this idea to a collection of visa records of Spanish citizens issued between 1936 and 1939 by a Spanish consulate in Buenos Aires.

Hybrid printed/handwritten forms: Printed text represent *"attributes"* or "concepts" which convey "semantic" context for the handwritten text (*"values"*)

Work carried out by ***tS*** for the EDT (European Digital Treasures) project:

Statistical Information Extraction from PrIx of EDT-Spain

Reasson to travel

```
---------------------------------   0%
1670.0 familia Familia
 191.5 Turismo turismo Turista
  62.7 profesión
---------------------------------  75%
  57.2 repatriado Repatriado
  34.0 esposa esposo Esposo
  28.8 [Deberes] militares militar
---------------------------------  80%
  16.0 negocios negocio Negocios
   8.0 servicio
   4.9 años
   4.8 comercio comercial
   3.0 Deportado
---------------------------------  81%
   3.0 compras comprar
   2.7 padres padre
   2.4 Sur
   2.2 caso
   2.2 casa
   2.4 navegar Navegar
   2.0 hijo hijos
   1.8 trabajo
   1.0 salud
   1.0 reunirse
   1.0 posesión
   1.0 patrones
---------------------------------  82%
 467.0 [OTHER and ERRORS]
--------------------------------- 100%
```

Jobs

```
-------------------------------------   0%
1344.9 empleado
 774.7 [sus Sus] labores
 165.4 jornalero
-------------------------------------  76%
  75.4 comerciante comercio Comerciante
  22.5 artista Artista
  16.3 marino Marino marinero
  11.0 mozo
-------------------------------------  80%
  10.5 sacerdote
  10.2 camarero
   5.7 casado
   5.7 abogado Abogado
-------------------------------------  81%
   5.6 estudiante
   3.0 minero
   3.0 labrador
   3.0 escritor
   3.0 engrasador
   2.0 fogonero
   2.0 Lingüista
   1.8 mecánico
   1.5 estado
   1.3 autor
   1.2 parado
   1.2 oro
   1.1 rentista
   1.1 Vigo
   1.0 viajante
   1.0 timonel
   1.0 pintor
-------------------------------------  82%
 543.6 [OTHER and ERRORS]
------------------------------------- 100%
```

Statistical Information Extraction from EDT-Spain PrIx’s

Civil State

```
----------------------------     0%
1096.8 soltero           36%
1015.7 casado            34%
----------------------------    70%
 437.0 soltera           14%
 209.0 casada             7%
----------------------------    91%
 120.5 viuda              4%
  70.3 viudo              2%
  21.7 célibe célibero    1%
----------------------------    98%
  62.2 [OTHER and ERRORS] 2%
----------------------------   100%
```

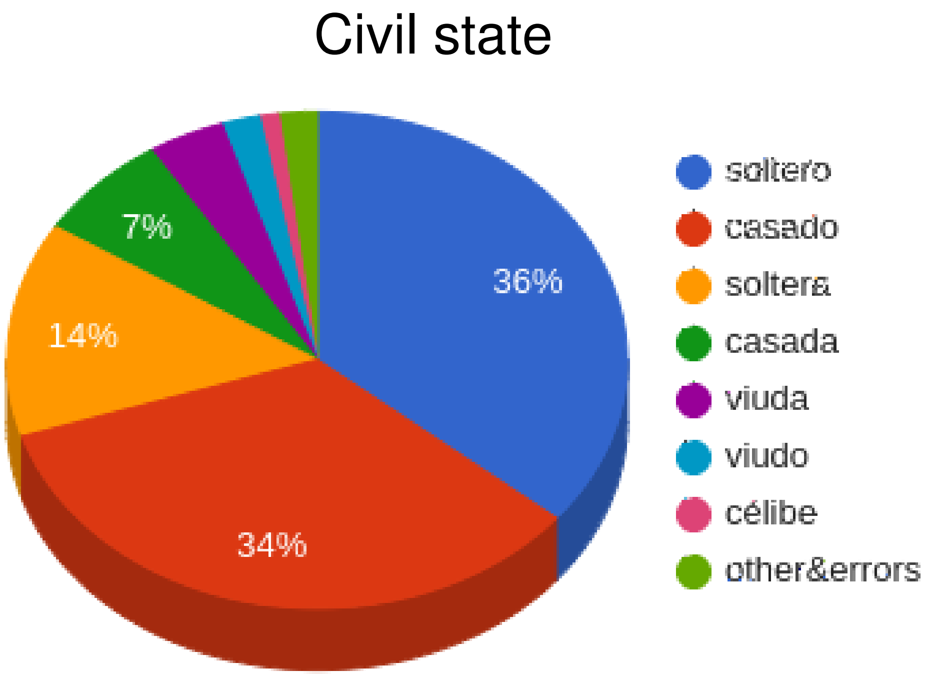
Genre

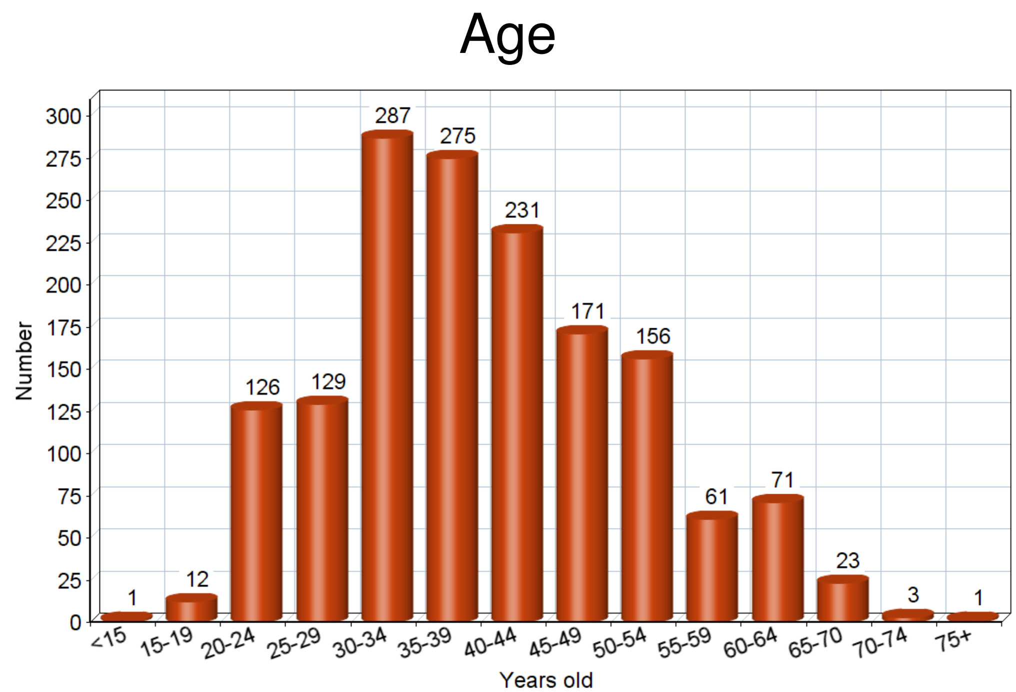
```
----------------------------
2182.8 Men:              72%
 766.5 Women:            25%
  73.9 Unknown:           3%
----------------------------
```

Age (years old)

```
-------------     0%
   1.4 <15
  11.9  15-19
-------------    10%
 126.3  20-24
 129.0  25-29
 286.6  30-34
-------------    50%
 275.0  35-39
 230.5  40-44
 170.7  45-49
 156.2  50-54
-------------    90%
  61.2  55-59
  71.4  60-64
-------------    99%
  23.1  65-70
   3.2  70-74
   1.1     75+
------------- 100%
```

Statistical Information Extraction from PrIx of EDT-Spain





Work carried out by ***tS*** for the EDT (European Digital Treasures) project: