Herramientas de transcripción asistida colaborativa de censos históricos

Alicia Fornés[1], Joana Maria Pujades-Mora[2], Oriol Ramos[1], Josep Lladós[1], Anna Cabré[2]

1. Centro de Visión por Computador, Dept. de Ciencias de la Computación, Ed.O, Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain. afornes@cvc.uab.es

2. Centro de Estudios Demográficos, Ed.E-2, Dept. de Geografía, Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain

El gran volumen de documentos almacenados en archivos históricos son un patrimonio de gran relevancia para el estudio y evolución de las sociedades que contribuyen a la preservación de la memoria histórica. En la era digital, las bibliotecas y archivos han dedicado un gran esfuerzo a digitalizar de forma masiva sus fondos, y en especial aquella documentación de carácter histórico. De este modo se asegura su preservación, pero a su vez, se abren nuevos retos sobre el acceso y valorización de los documentos digitales a través de la extracción, indexación y vinculación de sus contenidos mediante herramientas informáticas.

Las humanidades digitales son un área emergente e interdisciplinar en la que convergen las humanidades y la informática. A través del proyecto EINES, financiado por *La Obra Social La Caixa*, investigadores de las áreas de demografía y ciencias de la computación se unen para desarrollar instrumentos y procedimientos que faciliten la informatización masiva de las fuentes demográficas como los padrones. Los objetivos del proyecto son: construir bases de datos de uso público, mejorar el acceso y consulta de los documentos de archivos y construir herramientas de análisis de los datos.

El interés de los padrones reside en que son una fuente que a lo largo del siglo XIX se generalizó a la inmensa mayoría de municipios españoles, y que de forma similar, existen en muchos otros países, siendo por tanto, un proyecto escalable a nivel europeo. En concreto, se han seleccionado los padrones y censos del municipio de *Sant Feliu de Llobregat*, por ser un municipio importante de la época, con una estructura social y ocupacional diversificada, y con un número importante de padrones y censos (19 entre 1828 y 1955).

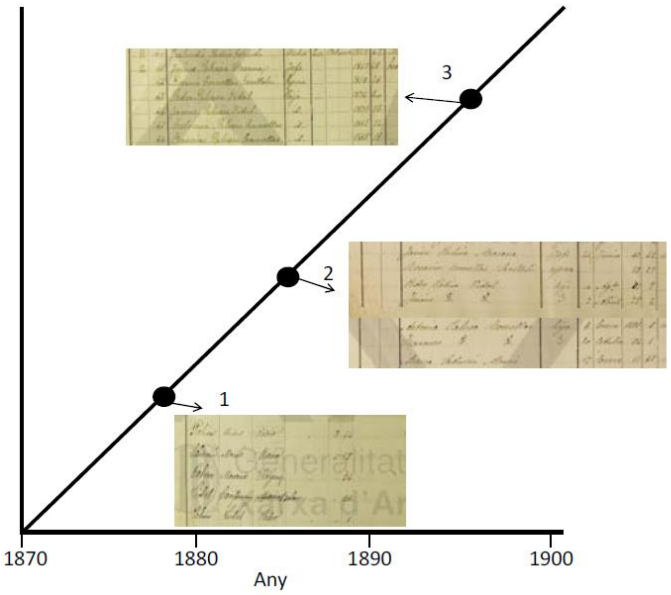
Plataforma de transcripción asistida colaborativa

Se ha desarrollado una plataforma de acceso a través de internet que permite a los usuarios transcribir de forma simultánea los contenidos de las imágenes del archivo, así como validar las transcripciones por parte de los expertos para asegurar la consistencia de los datos. La plataforma de transcripción se basa en el paradigma de "crowdsourcing", permitiendo que la tarea de transcripción se divida en múltiples tareas pequeñas (por ejemplo páginas) que puedan llevar a cabo un grupo numeroso de transcriptores. Además, dado que la información padronal se registra usando formularios, se pueden aplicar métodos de reconocimiento de manuscrito que permitan reconocer parte de la información con bastante fiabilidad (ej. nombres y apellidos frecuentes). Por esa razón, la plataforma incorporará progresivamente herramientas automáticas que permitirán la transcripción semi-automática, de manera que el tiempo dedicado a la transcripción vaya reduciéndose progresivamente.

La construcción de trayectorias vitales

A través del "*record linkage*" (vinculación nominal de registros) se pretende automatizar la generación de trayectorias individuales (ver Figura 1) y familiares a lo largo del tiempo, así como la localización espacial de redes familiares. El resultado final de este proceso será una gran red social histórica que a su vez gran parte serán multitud de genealogías entrelazadas. Este "*record linkage*" consiste en enlazar por ejemplo padres e hijos o detectar las apariciones del mismo individuo a lo largo de padrones de años sucesivos, etc.

Dado que muchos nombres, apellidos y lugares pueden aparecer con variantes ortográficas, en la plataforma de transcripción se están incorporando varias técnicas de comparación de cadenas de caracteres, como la distancia de *Levenshtein*. De este modo, para cada registro, el sistema propone enlaces entre individuos que tienen nombres y apellidos muy similares que vivan en el mismo lugar (ej. misma calle). Finalmente, el experto valida tanto las transcripciones como las relaciones propuestas entre individuos.



1	Padró 1878	Edat
	Pedro Pahisa Ribas	64
	Maria Sadurni Moner	59
	Jaime Pahisa Massana	34
	Dolores Vidal Campmany	26
	Pedro Pahisa Vidal	1

2	Padró 1886	Edat
	Jaime Pahisa Massana	42
	Rosario Comellas Sanllehy	27
	Pedro Pahisa Vidal	9
	Jaime Pahisa Vidal	6
	Antonia Pahisa Comellas	5
	Francisco Pahisa Comellas	1
	Maria Sadurni Moner	68

3	Padró 1896	Edat
	Jaime Pahisa Massana	48
	Rosario Comellas Sanllehy	38
	Pedro Pahisa Vidal	20
	Jaime Pahisa Vidal	16
	Antonia Pahisa Comellas	14
	Francisco Pahisa Comellas	11

Figura 1. Ejemplo de trayectoria individual de Jaime Pahisa Massana.

La transcripción de los padrones de *Sant Feliu de Llobregat* y la generación de trayectorias de vida no es un fin en sí mismo, sino que se busca que sea el prototipo que permita experimentar para después adaptar las herramientas generadas a otras fuentes documentales o municipios distintos. El objetivo final es poner la información demográfica histórica al servicio de la ciudadanía, creando nuevos productos y servicios a disposición de la sociedad para facilitar el acceso universal a los archivos históricos y que a partir de ésta se pueda generar conocimiento.