20 de Junio de 2016

Visualización de datos en Humanidades Digitales

Alejandro Benito Santos
abenito@usal.es

VNiVERSiDAD
D SALAMANCA

Departamento de Informática y Automática
Universidad de Salamanca

Agradezco al Prof. Dr. Roberto Therón su apoyo en la realización de este proyecto, así como a mi compañero Antonio Losada, que ha sido de inestimable ayuda a la hora de llevar a cabo esta investigación.

[1]Se empleó como formato para este documento una versión mejorada y actualizada por el autor de la plantilla de Informes Técnicos del Departamento de Informática y Automática, obtenida en la URL `http://diaweb.usal.es/diaweb/asignaturas/avisos/verAvisoTablon.jsp?cod_aviso=346`.

El yo subliminal no es de ninguna manera inferior al yo consciente; éste no es puramente automático; puede discernir, tiene tacto, delicadeza; sabe cómo escoger, adivinar. ¿Qué digo? Sabe mejor cómo adivinar que el yo consciente, ya que aquél tiene éxito donde éste ha fallado.

El valor de la ciencia
HENRI POINCARÉ

Índice

Índice de figuras

1. Introducción

1.1. Las Humanidades Digitales

La tarea de definir el campo de las Humanidades Digitales (HD de ahora en adelante por sus siglas en inglés), a pesar de tener su origen en la década de los años 40 del siglo pasado, es sumamente complicada aún a fecha de la concepción de este trabajo. Debido a la propia naturaleza del mismo, esta definición se encuentra en continua revisión y evolución por estudiosos y académicos de todo el mundo de acuerdo con los diferentes estudios y trabajos que se van publicando. Como ejemplo para el lector de la complejidad del asunto, investigadores en HD recopilaron en el año 2012 veintiuna definiciones diferentes [2], resultantes de una anterior criba de una lista online mucho mayor [3]. Ya que el objetivo de este trabajo no es dar una definición precisa del término, estableceremos como base para el resto del texto la siguiente aseveración, que es común hoy en día para la mayoría de académicos de las HD, de manera que, de forma general podemos decir que:

«HD es un campo de estudio resultante de la **intersección** entre las disciplinas de las ciencias de la **computación** y las **humanidades**. En ellas se comprenden una serie de ramas o especializaciones, que varían desde la catalogación de colecciones online hasta por ejemplo, la minería de datos de grandes conjuntos de datos culturales de todo tipo. Las HD incorporan datos digitalizados y/o digitales y **combinan metodologías de disciplinas tradicionales de humanidades** (como la historia, filosofía, lingüística, literatura, arte, arqueología, música y estudios culturales) con **herramientas proporcionadas por las ciencias de la computación** (como el hipertexto, la visualización de datos, la recuperación de la información, la minería de datos y textos o la estadística).»

A partir de esta definición, el lector quizás pueda ya suponer que las HD pivotan alrededor de una colaboración que se nutre de dos entes, lo tradicional, representado por el área de "humanidades" o "artes", y lo nuevo, las ciencias de la computación, emergiendo de dicha interacción el concepto de Humanidades Digitales, que pone a disposición de los investigadores nuevas metodologías y formas de trabajo, como podemos ver en la Figura 1.

Es en este área de confluencia donde surgen algunos de los retos a solucionar más interesantes. Como ocurre en otras disciplinas de convergencia (que no pertecen al ámbito puramente computacional), esta interacción requiere de un esfuerzo añadido para crear herramientas suficientemente potentes que investigadores de otras áreas (en el caso que nos atañe, las humanidades) puedan utilizar y comprender de manera natural. En la línea de este enunciado, el referente en visualización e investigador de la Universidad de Stanford Elijah Meeks completa la definición ya dada con un enfoque más práctico: «[Las HD] suponen la aplicación e integración de palabras clave y acrónimos en la investigación humanística [...] así como la demistificación de los métodos computacionales y su aplicación en nuevas formas y maneras no tradicionales [...] aplicados con el debido escepticismo hacia su validez final»[4].

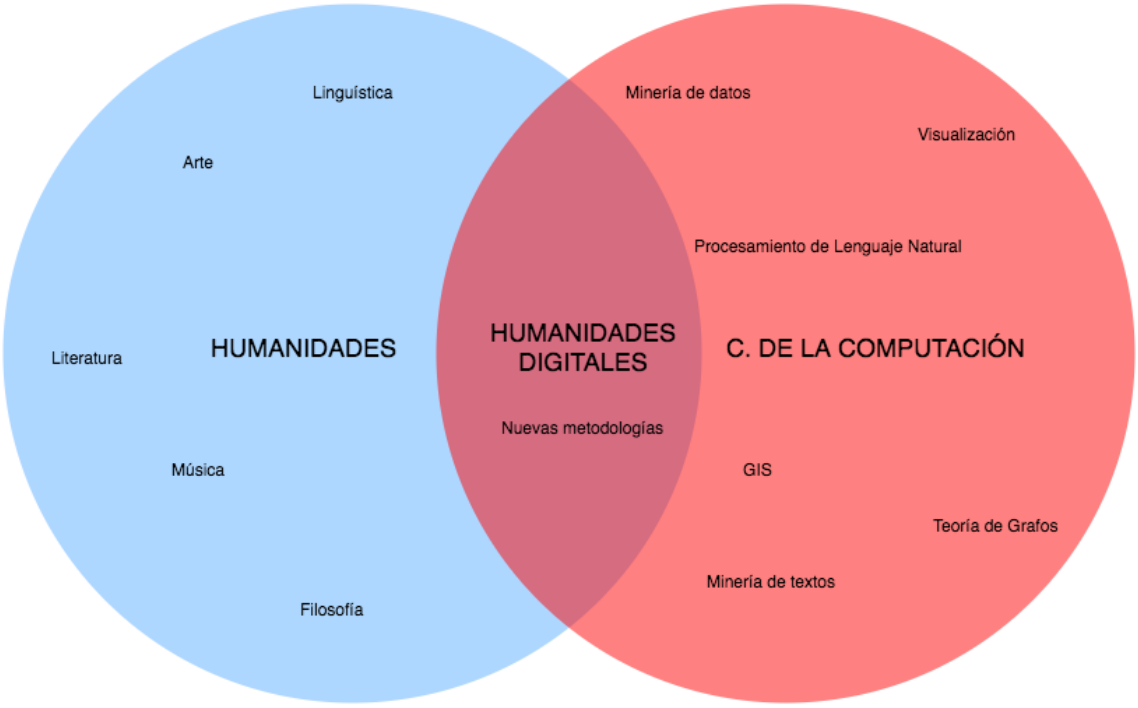


Figura 1: Esquema de la colaboración entre Humanidades y Ciencias de la Computación, resultante en la disciplina de las Humanidades Digitales

Este último punto resulta de especial interés en nuestro enfoque, ya que como veremos a lo largo de este estudio, en muchos casos los métodos computacionales que serían sin duda útiles en otras áreas de conocimiento más establecidas, pueden no tener validez en el ámbito de las HD o en el de un proyecto de investigación en concreto. Es por ello que se necesita una realimentación y validación continuas y una colaboración más estrecha entre los expertos de ambos ámbitos que en otras áreas de conocimiento para poder producir herramientas útiles que ayuden a responder correctamente a las preguntas humanísticas de los investigadores. Es decir, en el ámbito de las ciencias de la computación en general, y la visualización en particular en relación a las HD, se ha de poner especial esfuerzo en **no crear herramientas que dirijan el desarrollo de nuevos métodos estadísticos, matemáticos o de ingeniería**, sino que por el contrario, **apliquen métodos de probada eficiencia en maneras nuevas e inesperadas, que permitan el descubrimiento de conocimiento oculto hasta entonces.**

Es por estas razones que las HD requieren una **gran capacidad de innovación**, una **alta dosis de creatividad** y una **amplia base cultural y científica** por parte de los investigadores en ellas envueltas para que éstas sean capaces de conseguir avances importantes en dicho campo.

1.2. Motivaciones de las Humanidades Digitales

Recopilamos en esta sección una serie de razones por las que el autor piensa que las Humanidades Digitales son necesarias en el mundo moderno, y por las que

han de existir profesionales de la Ingeniería, en todos los ámbitos, que participen de estos proyectos. Sobre estas ideas volveremos continuamente durante la elaboración de esta tesis, ya que han servido en parte de inspiración y motor para la consecución de los objetivos de esta investigación:

1. **Las Humanidades Digitales son divertidas**: Al aportar métodos visuales e interactivos con los que el ciudadano de a pie, estudiante o investigador pueda establecer una relación con las diferentes ramas del conocimiento humanístico, se incentiva el interés por las mismas, y se consigue una mejor comprensión del conocimiento. Este es el llamado precepto de "gamificación", que consiste en convertir tareas que podrían resultar en un principio desmotivadoras para el profano en algo divertido, parecido a un juego. Las HD pueden servirse de este concepto para llegar a más personas y atraer a más personas que crean sinergias especiales necesarias para la resolución de problemas específicos, en un enfoque de Ciencia Ciudadana.

2. **Las Humanidades Digitales son inherentemente colaborativas**: Debido a su naturaleza humana, la creación y estudio de las HD hacen que el propósito de colaborar con otros surja de manera natural en los procesos de investigación y enseñanza. Las HD son más autoconscientes de estos procesos de colaboración que otras disciplinas más establecidas que la dan por sentado, y eso es algo que se ha de aprovechar para hacer que la partipicación de las partes implicadas sea lo más útil posible.

3. **Las Humanidades Digitales tienen un transfondo de Ingeniería:** Como hemos explicado con anterioridad, uno de los retos de las HD es exponer toda la potencia de los métodos de Ingeniería, aplicando éstos en maneras nuevas e inesperadas que consigan extraer la mayor cantidad de conocimiento posible de los datos. Ésto hace que esta disciplina resulte atractiva y desafiante para alguien con un perfil en Ingeniería.

4. **Las HD se aprovechan de la creciente accesibilidad a métodos computacionales**: El acceso a programas de código abierto y plataformas más potentes se ha incrementado drásticamente en los últimos años, de acuerdo a la Ley de Moore. Esto hace que métodos y prototipos que hace tan sólo unos años requerían de importantes (y caras) estaciones de trabajo para llevarse a cabo, puedan ser desarrollados en ordenadores personales o de más bajo presupuesto. Esto ha llevado por supuesto, a un incremento sustancial de creaciones en la materia, como veremos en apartados posteriores.

5. **Las Humanidades Digitales democratizan el conocimiento**: Las HD mejoran la accesibilidad al conocimiento por parte de no expertos, permitiéndoles adquirir una base cultural en dominios específicos donde el saber popular o el nivel de cultura general garantizado por los sistemas educativos no es suficiente para poder comprenderlo en un primer momento. Las HD desempeñan por tanto la importante tarea de instruir al ciudadano interesado en las estructuras abstractas sobre las que se asienta el mundo moderno.

6. **Las Humanidades Digitales dan un sentido nuevo a las técnicas computacionales**: Las HD marcan especialmente bien la utilidad de los objetivos conseguidos a través del estudio de las mismas. Por ejemplo, a un estudiante de Ingeniería que es instruido en el uso de técnicas de análisis de grafos o redes, no se le transmiten sin embargo los enfoques sociales o éticos que puede tener el uso de las mismas en el mundo real. Existe a veces por tanto una separación entre el mundo de las ideas y el de los hechos demasiado grande que hace que los conceptos abstractos no siempre sean interiorizados correctamente desde una perspectiva humana o social. Las HD, al ser más autoconscientes que otras disciplinas, y al referirse al fin y al cabo a problemas muy antiguos de la especie humana, pueden incrementar el interés y la captación de nuevo conocimiento esencial en muchos y variados aspectos por parte del ciudadano medio, lo cual, de acuerdo a algunas tesis, puede resultar beneficioso para el conjunto de la sociedad en general.

1.3. Los (3+1) pilares de las Humanidades Digitales

Como terminábamos diciendo en la sección inicial de este trabajo, uno de los objetivos principales del investigador en las HD es aplicar métodos computacionales en maneras novedosas, y es por ello que las HD toman ventaja de la creciente accesibilidad a los mismos que venimos experimentando en los últimos años. Esto significa que hoy en día es mucho más fácil realizar análisis espacial, textual o de redes mediante métodos visuales de lo que era hace diez o veinte años. Esto es especialmente importante si nos fijamos en los requisitos económicos, coyunturales y de infraestructura que envuelven estas tareas en comparación al pasado no tan lejano. Las herramientas y recursos de código libre disponibles en la red reducen considerablemente el coste de crear un prototipo funcional. Las tecnologías web y el abaratamiento de los procesadores permiten prescindir hoy en día de servidores y *workstations*, reduciendo abrumadoramente el coste estructural asociado a un proyecto de este tipo. El acceso abierto a los datos gracias a la multitud de iniciativas *Open Data* de las que gozamos los investigadores hoy en día hace que conseguir información no sea un problema. Y por último pero no menos importante: El nivel de pericia informática de la usuaria final puede ser considerablemente menor que en el pasado: Los principios de UI/UX están mucho mejor integrados en todos los *frameworks* y herramientas de software de hoy en día, reduciendo notablemente también el esfuerzo necesario por parte de investigadores y desarrolladores para crear prototipos y productos que transmitan correctamente la utilidad subyacente del software. Ya no se necesita ser un experto en grafos para realizar un análisis de redes sociales, de igual manera que no se necesita ser un geógrafo para hacer uno espacial.

En principio, no hay restricciones para aplicar cualquier método computacional en el ámbito de las HD. Por supuesto, lo útil que sea este enfoque dependerá de la pericia o buen criterio del investigador y de la situación en particular sin embargo, hay métodos que a lo largo de los años han demostrador ser más eficaces a la hora de representar el conocimiento humanístico. Hablaremos de tres pilares fundamentales en los que se apoyan la mayoría de proyectos e investigaciones en las HD (Sistemas

de Información Geográfica, Procesamiento del Lenguaje Natural y Análisis de Redes Sociales). De ellos se obtienen las fuentes de datos, métodos y procesos con los que trabajará el investigador, que finalmente unirá por medio de la Visualización de Datos, dando sentido y proyección a los mismos. Se reseñan a continuación algunas de las razones de uso y principales aplicaciones de los citados pilares. Véase la Figura 2 para una representación de esta idea.

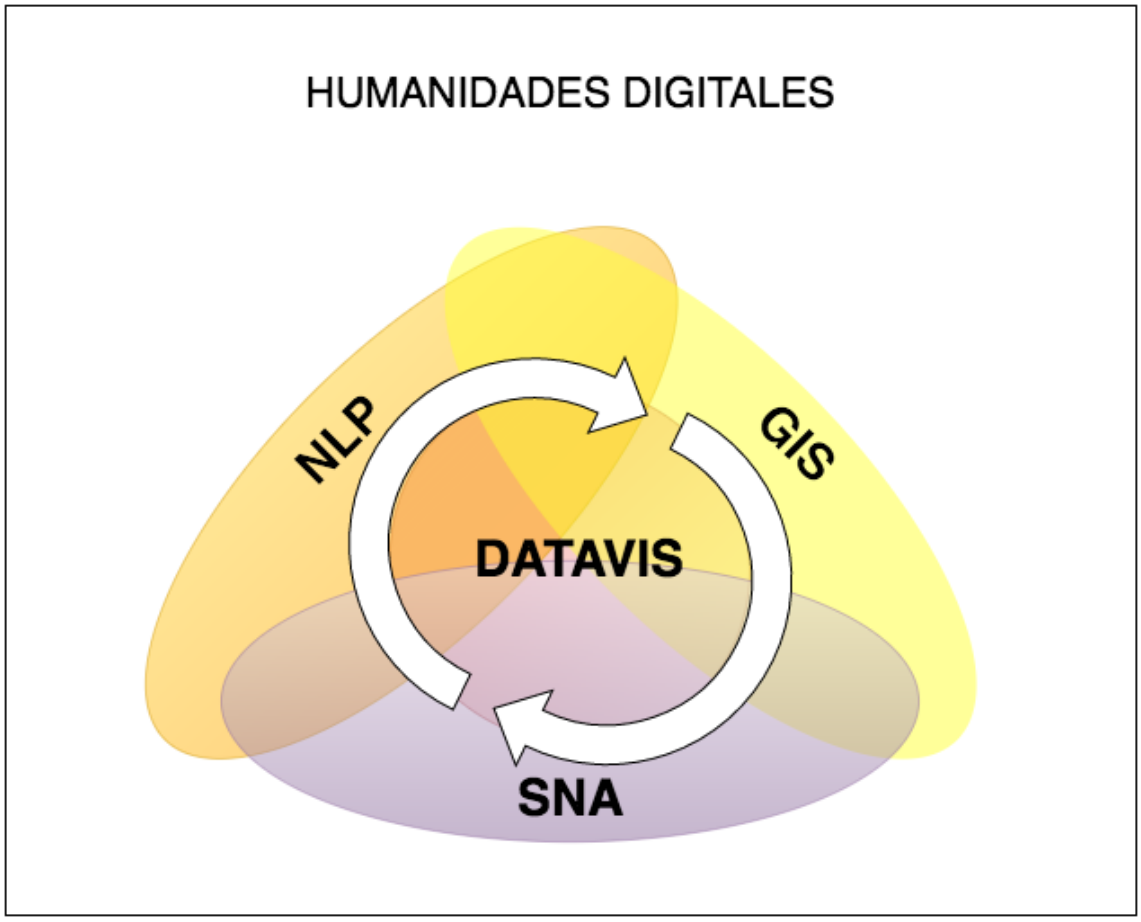


Figura 2: Los 3 + 1 pilares computacionales de las Humanidades Digitales. En el centro, la Visualización de Datos sirve para conectar las otras 3 partes y exponer cada una de sus características a la usuaria final

1.3.1. GIS: Sistemas de Información Geográfica

Tradicionalmente, el ser humano tiene una larga y rica historia relacionada con la representación de conceptos abstractos en mapas. Desde tiempos inmemoriales ha tenido la necesidad de almacenar, ordenar, cuantificar y valorar la información adquirida a través de la experiencia en estos artefactos. Hoy en día, las representaciones en papel o pergamino han sido sustituidas por el uso de estos sistemas que ha venido de la mano con un incremento en calidad y cantidad de la información que puede ser codificada en los mapas. Consecuencia directa de la capacidad ya innata con la que el ser humano es capaz de manejar un mapa es el gran auge de

estos sistemas en la actualidad. No necesitan de maestría ni de conocimiento experto añadido previo, simplemente tienen la capacidad de poder usarse por casi cualquier persona en el planeta. Hoy en día, mapas generados por un grupo de personas son empleados por cientos de miles de personas alrededor del mundo de manera casi instantánea, fácil y gratuita. El mejor ejemplo de este hecho es OpenStreetMap, un proyecto colaborativo en red para crear mapas editables con licencia libre. Una de las últimas mediciones realizada en 2014, expuso que el proyecto contaba con un número en torno a 1.840.000 usuarias registrados, de los cuales alrededor de 22.600 realizaron alguna edición en el último mes. Además, el número de usuarias crece a un ritmo de un 10 % mensual.[5]. Hoy en día se estima que este número ha superado ya los dos millones y medio de usuarias.

En el ámbito de las HD, diversos tipos de datos son visualizados por medio de mapas, en algunos casos de forma brillante, como veremos en la sección 2.

1.3.2. NLP: Procesamiento del Lenguaje Natural

El procesamiento del Lenguaje Natural (NLP por sus siglas en inglés) es un método computacional más difícil de demistificar que el GIS. De forma que en el caso de los mapas esta interacción surge de manera espontánea, no ocurre lo mismo con el NLP. En esta disciplina incluiremos técnicas de minería de textos y recuperación de la información, que tendrán el objetivo de acceder, condensar y abstraer el conocimiento subyacente en diferentes tipos de flujos de textos, normalmente a gran escala. El gran auge que han experimentado los motores de búsqueda textual en los últimos años ha potenciado que estas tareas se hayan suavizado en dificultad considerablemente. Como veremos en nuestro ejemplo, existen alternativas de código abierto que permiten indexar o buscar cadenas de texto en conjuntos de datos masivos en cuestión de milisegundos, y que son de altísimo valor a la hora de mejorar y acelerar todas las etapas del proceso de concepción de un prototipo en el área de las HD. Este hecho, unido a una correcta presentación de los resultados siguiendo los preceptos de la Visualización de la Información, supone una mejora muy importante en la calidad de los resultados obtenidos en todas las etapas del proceso investigador en HD.

1.3.3. SNA: Análisis de Redes Sociales

En HD hablamos de Análisis de Redes Sociales, y no nos referimos al concepto de Análisis de Redes más general porque (no se nos olvide) estamos hablando de conocimiento humano, que en la gran mayoría de los casos se explica con el primero. Es mucho más sencillo partir de redes sociales que sirvan como base para modelar redes de transportes, genealogías, redes administrativas o corrientes (de pensamiento, culturales o religiosas). Al fin y al cabo el término "social" se refiere a sociedades de seres humanos, que es precisamente el enfoque de los estudios humanísticos. En términos de demistificar esta rama de la computación, diremos que ocurre también de forma natural, al igual que pasaba en el caso del GIS. Casi cualquier persona con

un perfil en una red social de Internet, o que haya buscado direcciones utilizando un servicio en línea, tiene ya cierta familiaridad inconsciente con el análisis de redes.

Las teorías sobre grafos, y los algoritmos de división, *pathfinding*, recorrido y balanceo van a ser particularmente útiles en las HD. También en este aspecto, existen gran cantidad de herramientas para el análisis visual de grafos (la manera habitual de representar una red, aunque no la única), que presentaremos también en la sección 2.

1.3.4. Visualización de la Información

Se dedica un apartado especial en esta introducción a la materia a la Visualización de Datos y a las técnicas y herramientas de la misma más usadas en el ámbito de las HD:

El emplazamiento de la Visualización de Datos o de la Información como cuarto y último pilar es útil desde una perspectiva ontológica, ya que su propósito se solapa y entra en conflicto con los tres pilares anteriormente presentados. Sin embargo, le dedicamos una mención aparte (de ahí el 3+1) debido a su especial importancia como núcleo central y nexo entre los otros además de ocupar un espacio completamente distinto en el mapa mental de cualquier investigador de HD. La visualización es de vital importancia en el proceso debido a que va a proporcionar una entrada accesible y menos intimidatoria a la usuaria final a cualquiera de las otras tres ramas mencionadas arriba y por tanto, de la buena praxis del investigador al aplicar adecuadamente las técnicas de visualización va a depender que se libere todo el poder de las anteriores.

Existen muchas y diversas visualizaciones para multitud de naturalezas de conjuntos de datos y propósitos, la mayoría accesibles en la galería de ejemplos de D3.js[6], la biblioteca web de visualización de datos por excelencia. La visualización de datos es mucho menos esotérica de lo que era en un pasado, e Internet rebosa en ejemplos, información y *how-to's*, gracias en gran medida al trabajo de Mike Bostock, el creador de D3.

Una pregunta que podría asaltar al lector al leer estas líneas es: Si existen tantos tipos de visualizaciones, ¿Cuáles de ellas son adecuadas para las HD? De igual manera que existen ciertas áreas de la computación que, debido a la naturaleza del problema a resolver, son más adecuadas para las HD, ¿existe por tanto algún tipo de tendencia hacia un tipo concreto de visualización de los datos? Esta misma pregunta se hizo Katrien Verbert, profesora e investigadora del grupo de investigación sobre HCI (Interacción Humano Máquina) de la Universidad de Leuven en Bélgica muy activa en el area de las HD. En una de sus ponencias en la conferencia anual internacional de la Alliance of Digital Humanities Organizations (ADHO), que reúne a expertos en HD de todo el globo cada año, en su edición de 2015, analizó el tipo de visualización más usada en los envíos de publicaciones para la conferencia del año 2014.[7], arrojando datos a la luz de los cuales se derivan 3 preguntas principales de la investigación llevada a cabo por la autora:

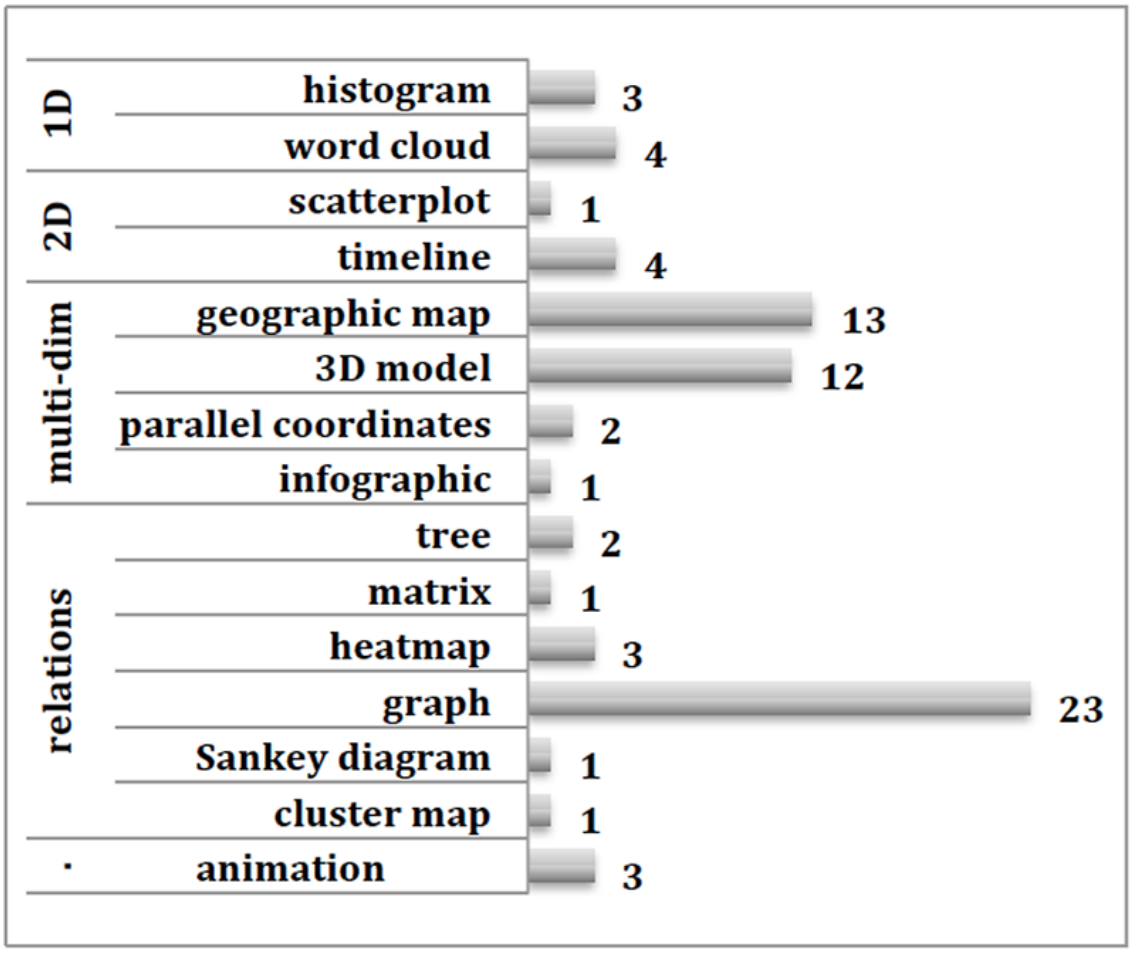


Figura 3: Los tipos de visualización más empleados en los envíos de DH2015. El mapa geográfico y el modelo 3D destacan en la representación multidimensional, mientras que el grafo es el recurso preferido para mostrar relaciones (redes).

1. ¿Qué técnicas de visualización son usadas por los prototipos y sistemas actuales? En relación con la respuesta a la primera pregunta, la autora hace explícito el hecho de que de todas las contribuciones, 58 presentan trabajo significativo en el uso de visualizaciones (Figura 3), dividiéndose a su vez por el tipo de problema que se desea atajar: Para la presentación de datos **unidimensionales**, se emplea el **histograma** y la **nube de palabras**[8]. Su uso principal es dar a entender tópicos generales sobre el conjunto de datos. En el caso de presentación de elementos **bidimensionales**, los ***scatterplots*** y las **líneas temporales** son los elementos más usados. Por último, en el análisis multidimensional, se destaca el uso de modelos 3D para representar edificios históricos que permiten una navegación virtual de los mismos. Más interesante para nuestro estudio es el empleo de mapas geográficos, que también destaca sobre el resto, empleado para representar 3 dimensiones: latitud, longitud y frecuencia de otra característica de los datos que varía dependiendo del caso concreto, que normalmente es codificada en el tamaño del punto en el mapa y/o en el color. Esta limitación en el número de coordenadas (3) hace habitual el hecho de que se complemente y enlace la vista de mapa con otras representaciones como

las coordenadas paralelas, con capacidad para representar un número teóricamente infinito de dimensiones[9].

Para finalizar con el comentario del trabajo de K. Verbert, sorprende al autor el bajo uso de animaciones (en sólo 3 de 58 prototipos) que ayuden a mejorar la comprensión de los datos, siendo ésta una técnica cuya efectividad, cuando es bien aplicada, a la hora de captar la atención de la usuaria y fijar su estado mental e interés en las partes relevantes de cada etapa del flujo de trabajo de una aplicación ha sido probada en múltiples ocasiones.[10] [11]. El impacto de las animaciones es sin duda un tema clave en el desarrollo del prototipo de esta investigación, y que es comentado en amplitud en apartados posteriores de este escrito.

2. ¿Qué técnicas de interacción son soportadas por los prototipos y sistemas actuales? En este apartado, la autora continúa con su análisis, esta vez fijándose en los tipos de interacción soportados por los sistemas pertenecientes al conjunto estudiado, que son clasificados de acuerdo al modelo propuesto por Yi et al.[12]. En su estudio se concluye que el tipo de interacción más común es **abstraer/elaborar**, que permite a las usuarias mostrar más o menos detalle dentro de la visualización, como por ejemplo recorrer ramas de un dendrograma, o el uso de *zoom* en un mapa geográfico, también bastante común. Otro tipo de interacción común es el uso de *tooltips*, que muestran información relacionada cuando se deja el cursor sobre una entidad del gráfico, también llamadas de "detalle bajo demanda". Otro tipo de interacción bastante común en este tipo de sistemas es la operación **exploratoria**. Tales operaciones permiten a la usuaria fijarse en un subconjunto de los datos en la representación visual. Una operación típica exploratoria es el llamado "panning" o desplazamiento, que posibilita a la usuaria ver partes diferentes de la representación de un grafo, o de un mapa geográfico. Casi la totalidad de los prototipos que soportan interacción incluyeron este tipo de funcionalidad. Por último, las técnicas de interacción de **filtrado**, conceden a la usuaria la posibilidad de elegir el conjunto de datos que se muestra en la visualización de acuerdo a ciertas condiciones específicas. En este ámbito son comúnmente usados los *sliders* para seleccionar rangos, y las *checkboxes*, en el caso de datos categóricos. Un porcentaje menor de prototipos incluyeron filtrado textual a través del teclado. Anticipamos al lector la vuelta a todos estos tópicos en secciones posteriores, que constituyen la justificación y base de muchas de las características que se incluyeron en el prototipo propuesto, habiendo sido algunas de ellas incluso mejoradas por combinación con otras técnicas de UX.

3. ¿Cómo son evaluados los enfoques visuales en la investigación de HD?
El último punto de la investigación plantea qué métodos son los más usados para medir la validez de un sistema o prototipo en el área de las DH, y a qué tipo de pruebas son sometidos para asegurar su congruencia o utilidad. En el caso del conjunto de sistemas analizados, sólo cuatro de 58 han sido evaluados con casos de estudio reales que midiesen su verdadero valor. En todos ellos, estas aplicaciones han servido para que las usuarias finales desempeñaran tareas de investigación en sus diferentes áreas. Sin duda estos casos de estudio son la prueba de fuego para los prototipos, y son los que más información aportan al respecto de la idealidad

y usabilidad de las diferentes técnicas de visualización aplicadas al dominio de las HD. Entre las conclusiones a las que se llegó al intentar dar respuesta a la pregunta que se formulaba en este apartado, se comprobó por ejemplo que la técnica de visualización de clusters (*cluster map*), resultó ser demasiado compleja para usuarias no experimentadas. Por otra parte, un diagrama de Venn resultó ser mucho más efectivo con ese mismo grupo de personas.

Como colofón, se comprueba que en el ámbito de las visualización en las HD, la prueba del prototipo a través de casos de estudio adecuados para la usuaria final resulta crucial a la hora de refinar y validad las técnicas empleadas. Como decíamos al principio de este documento, es necesario probar diversas técnicas de formas novedosas y atrevidas, que sean capaces de extraer conocimiento de los datos. Sin embargo, esto supone una razón de peso para prestar aún más atención si cabe a la etapa de validación del prototipo y nunca lo contrario.

2. Trabajo relacionado

Como se ha comentado en la sección 1, los avances tecnológicos de los últimos tiempos, y el incremento en la accesibilidad a los mismos, han generado una gran producción académica y no académica en las HD en los últimos tiempos. Numerosas son las instituciones públicas y privadas que han puesto sus ojos en esta antigua disciplina renovada, tratando un gran número de temáticas diferentes, que varían desde el procesamiento y visualización de textos y lenguajes hasta la representación de características, análisis de sentimientos y cultura tradicional en entornos GIS.

Esta sección comienza por un breve repaso a la historia de las Humanidades Digitales, reseñando los hechos más importantes que han tenido lugar desde su concepción a finales de los años 40. Se pasa a continuación a reseñar una selección de trabajos de toda clase que bien por su calidad o parecido en temática al trabajo aquí presentado han sido incluídos en esta lista. Debido a la naturaleza del estudio que se expone en este escrito, esta sección describirá en mayor profundidad, aquellos trabajos dentro de las HD cuya temática este más relacionado con el procesamiento del lenguaje natural y muy especialmente con el campo de la lexicografía. Además, se da una especial importancia a aquellos trabajos de visualización de características cartográficas y redes sociales, ya que todos ellos, en mayor o menor medida, han servido de inspiración y referencia para la realización de nuestro análisis.

2.1. Historia y Estado Actual

A pesar de haberse incrementado notablemente el número de proyectos relacionados en los últimos años, ya se ha comentado que éstas surgen formalmente hace más de 70 años, lo que implica la necesidad de aventurarse en este trabajo, aunque sea brevemente, en los hechos claves que han modelado el devenir de este disciplina desde los años 40 del siglo pasado. Es así que en esta sección hacemos un repaso cronológico de los mismos, para terminar analizando su estado actual y el impacto que tienen en la sociedad de hoy en día.

2.1.1. Los inicios, de 1949 a principios de los 70

Las HD descienden directamente del hecho de aplicar métodos computacionales a las humanidades, consiguiendo "representaciones formales del registro humano adaptadas al ordenador"[13]. Sus orígenes se encuentran en los finales de los años 40. El iniciador y pionero en el área fue el Padre Roberto Busa[14], un sacerdote jesuita italiano, que comenzó a realizar una tarea que aún hoy consideraríamos monumental: Realizar un *index verborum* de todas las palabras de los trabajos del filósofo y sacerdote medieval Sto. Tomás de Aquino y autores relacionados, dando como resultado un corpus de 11 millones de palabras en latín medieval. El padre Busa viajó a los Estados Unidos, donde conoció a Thomas J. Watson, un trabajador de la compañía IBM y ambos comenzaron la tarea. Progresivamente, los textos fueron

registrándose en tarjetas perforadas, que era el almacenamiento de datos existente en la época.

El padre Busa fue muy exigente en su trabajo y no se conformó con crear una red de concordancia de acuerdo a las formas gráficas de las palabras, sino que quiso crear una red de concordancia en base a los lemas de cada una de las apariciones. Este proyecto se completó no sin pocos esfuerzos, y fue el primer trabajo de impacto en el área de las HD. Hasta los años 70, otros autores imitaron el trabajo de Busa empleando corpus de otros idiomas (en su mayoría antiguos) para crear redes de concordancia. Es el caso del alemán, por Wisbey en 1963[15], concordancia de poemas en inglés, Parrish 1962[16], el francés [17] o el trabajo de De Tollenaere en el Instituto de Lexicología Holandés en Leiden, 1973 [18].

Cabe reseñar que en esta época los esfuerzos se centran en crear enfoques cuantitativos a los datos, más que puramente computacionales, y se llevan a cabo tareas que habían sido imaginadas décadas o incluso siglos antes. Por ejemplo, Augustus de Morgan propone en 1851 un estudio cualitativo del vocabulario como medio para investigar la autoría de las epístolas de S. Pablo. T.C. Mendenhall, recoge el testigo a finales del s.XIX e imagina la máquina ideal que realizaría esta tarea, que en aquel momento tuvo que hacerse manualmente. A principios de los años 60, un clérigo escocés, Andrew Morton, publica en un periódico inglés de la época que, de acuerdo a los métodos aplicados en su computador, S. Pablo sólo escribió una de las cuatro epístolas[19].

En aquel momento, la limitación en la tecnología era aún un gran problema para los investigadores. Los datos a analizar eran textuales o numéricos. Éstos seguían introduciéndose laboriosamente a través del uso de tarjetas perforadas, que cada una podía contener hasta ochenta caracteres o una línea de texto. La representación de los diferentes símbolos era también un problema sustancial, que sólo llego a ser solucionado parcialmente con el advenimiento de Unicode, en la década de los 80. Hasta entonces, este problema se resolvió marcando los caracteres con otros símbolos como asteriscos, para que fuesen reconocibles por una analista en la investigación. También se encontró la problemática de que los ordenadores en aquel entonces tenían dificultades para conseguir la tarea de identificar referencias a textos poéticos o literarios, ya que aquéllos asumían una estructura de texto de artículo.

En estas fechas tuvo lugar lo que se considera hoy en día la primera serie de conferencias de académicos interesados en la computación de las humanidades, precursora de la "Association for Literary and Linguistic Computing/Association for Computers and the Humanities" (ALLC/ACH)[20], que perdura hasta nuestros días. Las actas de dichas conferencias sentaron estándar para las publicaciones que se sucederían en el futuro. Éstas ponían el énfasis en el interés por la programación, métodos de entrada/salida, lexicografía, enseñanza del lenguaje y la estilística. Incluso en estos tiempos se puso de relieve la necesidad de una metodología para archivar y mantener textos electrónicos.

Para finalizar con el repaso de esta etapa, cabe destacar tres hechos importantes: 1) Se crea una nueva revista científica en 1966, "Computers and Humanities" que pasó a convertirse en el vehículo de transmisión de ideas en la materia por excelencia.

2) Thomas Wisbey funda el "Centre for Literary and Linguistic Computing" en Cambridge en 1963, como medio para continuar sus investigaciones sobre textos en alemán antiguo citadas anteriormente. 3) Wilhem Ott establece un grupo en la Universidad de Tübingen (Alemania), cuya producción se basa en la creación de una *suite* de programas para el análisis de textos, especialmente para la producción de ediciones críticas. Estos módulos, llamados TUSTEP[21], son de especial interés en el contexto de nuestro problema, ya que como veremos más adelante, gracias a su calidad y utilidad para los investigadores siguen en uso y desarrollo continuo hasta nuestros días. Adelantamos al lector que los datos de entrada del prototipo planteado en esta tesis se encuentran precisamente en este formato.

2.1.2. Consolidación, de los años 70 a mediados de los 80

En este período más investigadores comienzan a utilizar y mejorar en su trabajo diario métodos concebidos en la etapa anterior. Se digitalizan más textos y surgen nuevos proyectos. Esto hizo posible que más gente, no sólo del ámbito académico, empezase a preguntarse cómo las computadoras pueden ayudar en sus investigaciones y en la enseñanza de las materias.

Es en estos años cuando comienzan una serie de conferencias bianuales en el Reino Unido con el simposio de Cambridge en 1970, que se vería sucecido por el de Edimburgo (1972), Cardiff (1974), Oxford (1976), Birmingham (1978), y Cambridge (1980), todos ellos dando como resultado una prolífica producción científica en el área. En 1973 se funda la ALLC en el King's College de Londres. Por el mismo tiempo, una serie de conferencias comienzan su andadura en este caso en Norteamérica, llamadas "International Conference on Computing in the Humanities" (ICCH). Paulatinamente las conferencias en Norteamérica y Europa empiezan a converger. La ACH ve la luz en 1978 gracias a una de ellas.

Los requisitos de la computación de las humanidades empiezan a ser reconocidos dentro de los centros académicos de computación e informática, y los *mainframes* empiezan a recibir peticiones de uso por parte de investigadores de dicho campo. Surge otra suite de programas para la computación de textos en la Universidad de Oxford, llamada Oxford Concordance Program (OCP), que abarató significativamente los costes relacionados con la producción de proyectos de computación de las humanidades.

Con el advenimiento del software empaquetado y de suites de software como la mencionada en el párrafo anterior, se redujo el tiempo necesario en la programación de los *mainframes* que desempeñaran tareas comunes del dominio del problema, y progresivamente se empezó a invertir más tiempo en preparar los textos específicos que servirían de entrada para los sistemas concebidos en la época. El Oxford Text Archive (OTA) comenzó a recopilar textos ya procesados de investigadores de todo el mundo, una vez éstos hubieran terminado de trabajar con ellos, que a su vez eran puestos a disposición de otros interesados en ellos (habiendo cumplimentado los distintos derechos de copyright asociados a los mismos). Éstos fueron los comienzos de la primera biblioteca digital de la historia. Se comienzan a definir los primeros

estándares primigenios de descripción de textos.

Mientras tanto, surgió en la Universidad de California en Irvine, el primer repositorio de textos específico en una materia, el *Thesaurus Linguae Graecae* (TLG), como un intento de crear un archivo accesible de textos en Griego Antiguo que cubría más de 800 años de historia. En esta iniciativa, se llegaron a digitalizar 70 millones de palabras. Un repositorio parecido, pero esta vez con textos en latín surge años más tarde en el *Packard Humanities Institute*, que junto con el TLG sirvió de fuente para investigaciones de académicos durante muchos años. En otro aspecto, surgen más centros para la computación de las humanidades en diferentes partes del mundo, como el *Norwegian Center for the Humanities*, conocido hoy en día como HIT, o el *Center for Computer Analysis of Texts* (CCAT), en la Universidad de Pennsylvania. También surgen los primeros grados y cursos especializados en la computación de las humanidades, que concentraban su enseñanza en el aprendizaje y uso correcto de algunas de las suites de software mencionadas con anterioridad. La llegada del almacenamiento en disco y las BBDD supone un gran avance en el área, permitiendo búsquedas no secuenciales de textos por primera vez. Se presta especial atención en este período a la aplicación de cálculos matemáticos para manejar los recuentos de palabras, desplazando en protagonismo a investigadores más orientados hacia la temática pura de humanidades.

Atendiendo a la temática de los artículos producidos en esta época, se observa cierta preponderancia de estudios de vocabulario generados inicialmente por los citados programas de concordancia (OCP, TUSTEP). Existía entonces un cierto grado de escepticismo por parte de un sector de académicos de Humanidades hacia la validez o utilidad de los resultados que emitían las publicaciones de trabajos en el área asistidos por ordenador, desplazando las publicaciones de este tipo a revistas no especializadas en humanidades puras.

2.1.3. Nuevos desarrollos: Mediados de los 80 a principios de los 90

Este período se caracterizó por experimentar un *sorpasso* a las metodologías y desarrollos creados en la etapa inicial, hecho debido en gran medida al auge del ordenador personal y el correo electrónico, que permitieron reducir la duplicación de esfuerzos considerablemente y espolearon la innovación y la asunción de riesgos. Al final de la década de los 80, existían tres programas basados en MSDOS para el manejo y tratamiento de textos académicos en formato electrónico: Word-Cruncher, TACT y MicroOCP, todos ellos ofreciendo muy buena funcionalidad. Dos de ellos usaban búsqueda interactiva, mientras que que MicroOCP empleó la técnica de concordancia por lotes.

En este período, el Apple Macintosh resultó de especial interés para los investigadores en humanidades por dos razones: Primero, contaba con un interfaz gráfica mucho antes que los sistemas Windows o PC, lo que significaba que era mucho mejor que sus competidores a la hora de mostrar caracteres especiales. Era ahora posible ver caracteres en inglés antiguo, griego, cirílico o casi cualquier otro alfabeto en la pantalla, y manipular textos que contenían estos juegos de caracteres de manera

sencilla. Segundo, el Apple Macintosh contaba con un herramienta que suscitó gran atractivo a los estudiosos del momento: Hypercard (Figura 4). Esta aplicación ofrecía la posibilidad de crear tarjetas con vínculos entre ellas, así como también disponía de un lenguaje sencillo de programación que promocionó la creación de aplicaciones basadas en HyperCard entre los investigadores, destacando entre ellas el Proyecto Perseus, que presentaba textos antiguos en inglés a la usuaria con enlaces a versiones modernas de los mismos y anotaciones contextuales de varios tipos como imágenes o gráficos.

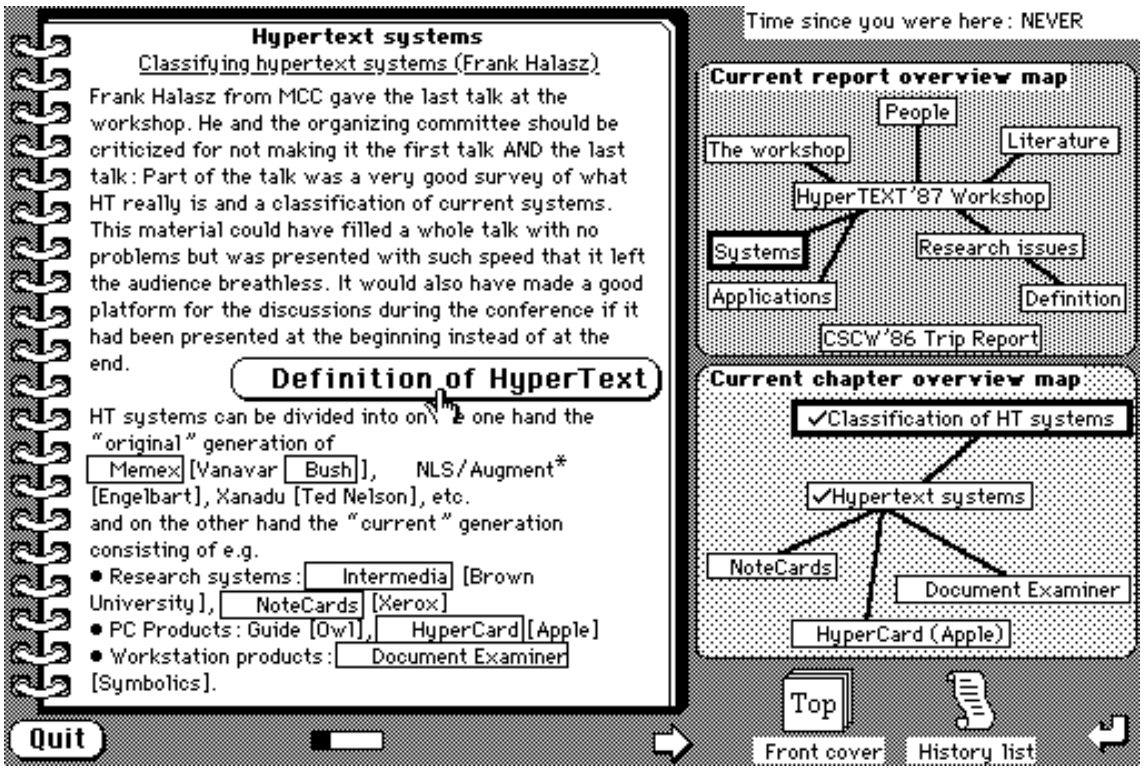


Figura 4: Detalle de la primera versión de Hypercard de Apple, mostrando en su interfaz un hipertexto primigenio.

Si nos referimos al aspecto de publicaciones, este período fue también altamente prolífico: Se crea el primer intento a gran escala de producir una bibliografía editada de proyectos, software y publicaciones, que se denominó *Humanities Computing Yearbook* (HCY). El primero apareció en 1988 y contaba con 400 páginas. El segundo una revisión aumentada y mejorada del primero, surgio en 1990 y se publicó en 700 páginas. Este segundo volumen, hasta que empezó a quedarse atrás de los tiempos, fue un recurso excepcionalmente codiciado y pasó a ocupar el lugar del *Computers and the Humanities* de épocas pasadas como referencia básica de la disciplina.

En términos de desarrollo intelectual, una actividad supera en notoriedad e importancia al resto de las cosechadas en esta etapa: Se crea finalmente el primer esquema estándar de codificación de textos humanísticos [22], poniendo fin al caos existente en este aspecto. El *Standard Generalized Markup Language* (SGML) se convirtió en estándar ISO y ofreció un mecanismo para definir un esquema de marcado que podía manejar muchos tipos diferentes de textos, metadatos y era capaz

de representar acertadamente representaciones y anotaciones de académicos, manteniendo la estructura básica de los mismos. Se crean los *Guidelines for Electronic Text Encoding and Interchange* de *Text Encoding Initiative*, que sirve de comité para discutir el estándar.

Este comité también fue el primer intento sistemático de categorizar y definir todas las características que poseen los textos humanísticos de interés para los investigadores. Se especifican 400 etiquetas diferentes, con capacidad de ser extendidas a subdominios concretos y nuevas áreas de aplicación. Esta discusión llevo a enormes esfuerzos intelectuales desempeñados en la tarea de la teoría de los lenguajes de marcado en particular y la representación del conocimiento humanístico en general. Mucho de este esfuerzo se realizó de manera colaborativa por diferentes entes alejados geográficamente por medio del e-mail y de listas públicas de discusión, hecho inimaginable tan sólo pocos años atrás. Este proyecto se convirtió en referente para un modelo de trabajo distribuido y en red, asistido por reuniones y conferencias en las que los participantes podían concretar los temas discutidos en la Internet.

En esta época, debido al creciente tamaño y grado de especialización de los diferentes proyectos de investigación, el tópico de la lingüística computacional comienza a divergir del resto al que había permanecido unido hasta entonces, dando lugar a conferencias y publicaciones especializadas en el mismo. Esta separación, a pesar de los esfuerzos de algunos académicos del momento, continuó durante largo tiempo y tan sólo hasta recientemente estas ramas (la lingüística computacional y las computación de las humanidades) han vuelto a convergir en ciertas acometidas investigadoras.

2.1.4. La era de Internet: De principios de los 90 hasta el presente

Si hubiese que elegir un hecho notorio que reseñar en la década de los 90, no sólo en el ámbito académico, ése sería el nacimiento y popularización del uso de Internet, y más en concreto de la *World Wide Web*, que cambió para siempre la manera en la que los seres humanos interactuamos con las fuentes de información y el conocimiento en general.

Al principio, este auge fue tratado con impasividad por la comunidad científica de las HD. Se comprobó desde estos entornos que el estándar de marcado HTML no era suficientemente potente para llevar a cabo investigaciones como las soportadas por SGML, y que aquél adolecía de muchos de los problemas ya observados en procesadores de textos anteriores. Es así que la WWW se empleó principalmente para buscar fuentes de datos y establecer contactos, pero nunca como una herramienta seria de investigación a tener en cuenta. Por el contrario, este surgimiento sí se presentó como una muy buena oportunidad para instituciones que típicamente no habían estado involucradas en la investigación humanística y computacional para introducirse en este mundo. La Internet era un medio perfecto para hacer llegar a un número mayor de personas no sólo los resultados de sus últimas investigaciones, sino también para promover sus actividades entre un número mayor de potenciales interesados: El formato de distribución ya no estaba limitado al papel escrito, no existía

teóricamente limitación en el tamaño y los hipervínculos suponían una útil manera de añadir anotaciones y otros comentarios, permitiendo la creación incremental de las publicaciones.

A principios y mediados de los años 90, se anunciaron muchos proyectos nuevos en el área de las HD, algunos de los cuales tuvieron éxito consiguiendo importantes inversiones de dinero que posibilitaran su viabilidad. El el área de las ediciones electrónicas de textos, desatacan los trabajos de Finneran[23], en 1996 y Bornstein y Tinkle[24], en 1998. Sin embargo, muchos de estos trabajos quedaron relegados al aspecto teórico y nunca llegaron a llevarse a la práctica.

Comenzaron los primeros debates sobre cómo llamar a las colecciones electrónicas de recursos online. Se consolida la preferencia de SGML para manejar este tipo de bibliotecas, en su mayoría impulsadas por el TEI. Se pone especial énfasis en mejorar la calidad de la navegabilidad, ya que por aquel entonces, el volumen de datos almacenado era ingente, incluso los órdenes de escala que se manejan en la actualidad. Surge una preocupación por las interfaces de usuario, usabilidad y experiencia de usuario en el manejo de estas aplicaciones. Mucha más gente, de entornos tradicionalmente reacios al uso de ordenadores, se familiariza con el uso de nuevas tecnologías en su investigación diaria.

Comienza el tiempo de los motores de búsqueda e indexación de textos, gracias a tecnologías como Lucene y Solr, que derivarían en un futuro en motores más modernes y potentes como ElasticSearch, al que se le dedica un importante apartado en este documento debido a su vital importancia en el prototipo resultante de nuestra investigación. Antes del surgimiento de esta familia Java de buscadores, destacan las iniciativas de ciertas instituciones de los Estados Unidos, que desarrollaron sus propios buscadores como OpenText[25], un motor de búsqueda textual SGML, o DynaText. Destaca también el Proyecto Orlando, que crea una base de datos documental de la historia de mujeres escritoras en lengua inglesa, almacenando biografías de las autoras, pasajes históricos y otros eventos también empleando el formato SGML. Eso permitía, empleando un enfoque "pick & match", generar nuevo material a partir de retazos de los originales, la creación por ejemplo de cronologías temáticas con mínimo esfuerzo. Este proyecto sentó un importante precedente en los métodos de trabajo de los investigadores, ya que era algo radicalmente diferente a lo empleado hasta el momento.

La llegada de la Internet también hizo posible desempeñar ya no sólo conversaciones entre grupos de investigación, sino llevar a cabo proyectos de investigación completos de forma paralela y colaborativa. La idea de varias personas contribuyendo con anotaciones sobre un mismo texto básica se había hecho realidad. El Peirce Project[26] y el Codex Leningradensis[27] fueron dos de los primeros proyectos que ofrecieron esta posibilidad.

Otra dimensión importante sobrevino en los años 90 en el ámbito de las HD, que hasta entonces habían reducido su foco de acción a fuentes textuales: **El uso del multimedia**, en forma de imágenes, audio y video. Muchos de los académicos del momento se preguntaban qué tareas de apoyo a la investigación se podrían realizar con este nuevo tipo de fuentes. Empiezan entonces los primeros intentos de

enlazar texto a imágenes, hasta la resolución de la palabra, La comunidad científica, típicamente acostumbrada a trabajar con textos, fue en un principio escéptica en cuanto al uso de este nuevo tipo de recursos sin embargo hoy en día nadie duda de su utilidad.

En el aspecto de la enseñanza, muchas universidades ofrecen hoy en día cursos y grados especialmente enfocados a la disciplina de las HD, como el King's College de Londres, la McMaster University de Canadá o la Universidad de Virginia en EEUU. En 2012, el *UCL Centre for Digital Humanities*[28] crea una infografía en el que se cuantificaron 114 centros educativos en 24 países ofertando este tipo de enseñanzas. Pasamos a comentar los resultados mostrados en dicha infografía en la siguiente sección.

2.1.5. Algunos datos y conclusiones

Es un hecho probado que las HD pueden contribuir sustancialmente al creciente interés de compartir recursos sobre herencia cultural en la Internet, no sólo para usuarias académicas, sino también para el público general con interés por la materia. Las progresiva incorporación de técnicas de lingüística computacional puede reducir enormemente los costes de la adquisición de nuevos datos sin comprometer su valor cultural o funcionalidad. El creciente uso (Figura 5) de las nuevas tecnologías hace que más y más gente se muestre interesada en el área, y por tanto el acceso a los recursos ha de ser más democrático e igualitario que nunca si cabe. Las HD tienen sin duda un compromiso con el ciudadano a la hora de poner a a su disposición estos recursos de manera accesible y manejable, en un intento de crear una sociedad más informada y autocrítica, que sea consciente de su pasado y pueda participar de manera activa en los avances que marquen el progreso de la misma.

2.2. Artículos académicos

La producción de artículos académicos en las áreas de las humanidades digitales y la visualización ha ido en crescendo en los últimos tiempos y cada vez más centros académicos van viendo aumentar su número de publicaciones en el área. Sin embargo, existe un selecto grupo de grupos de investigación que típicamente han sido origen de la mayoría de estos artículos, y que se podrían considerar referente en el área de las HD y la visualización.

Un buen ejemplo de este hecho ocurre con el Departamento de Análisis de Datos y Visualización de la Universidad de Constanza (Alemania), liderado por el Prof. Dr. Daniel A. Keim. Entre su extensísima producción académica, relacionada con áreas tan diversas como las Ciencias del Deporte, la medicina o la genómica, podemos encontrar ejemplos en el área de las HD. Uno de sus primeros trabajos en dicha materia data del año 2013-2014, y que es de especial interés en nuestro trabajo, es el titulado "Visual Analytics of Change in Natural Language"[29]. En esta publicación se propone una novedosa solución al problema de la detección y comprensión de cambios observables en textos de lenguaje natural o basados en él, entendiéndose

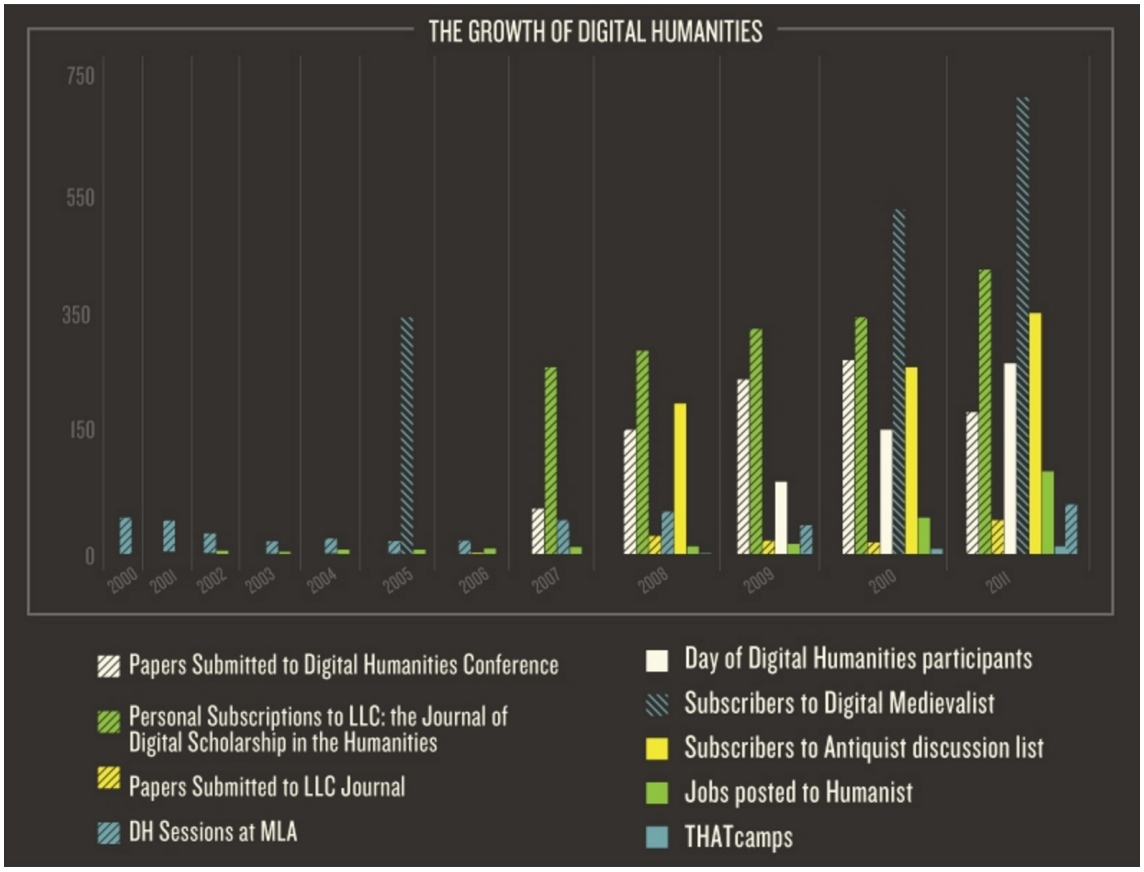


Figura 5: Histograma que refleja el creciente interés en el área de las HD en base a diferentes parámetros (académicos y no académicos) en la primera década del S. XXI.

cambio como la variación observable de características y patrones en los mismos. Este trabajo a su vez se divide en dos partes bien diferenciadas: En la primera, se incluye trabajo en aquel momento pionero en la intersección entre las áreas de la comparativa tipológica histórica de lenguajes y la analítica visual y en las sinergias surgidas en esta nueva colaboración. En la segunda, se ahonda en los métodos de analítica visual para la detección y exploración interactiva de cambios repentinos en datos textuales de gran tamaño. Además, complementa trabajos anteriores sobre analítica visual de textos basada en el tiempo de los cuales, como veremos en secciones posteriores, este trabajo intenta recoger el testigo.

Mención especial merecen los esfuerzos realizados en dicho trabajo en la comprensión de mutaciones semánticas diacrónicas en las palabras mediante recursos visuales como el uso de líneas temporales y glifos que representan características semánticas en los diferentes contextos en los que aparecen las mismas en cada una de las épocas tratadas en el estudio. Otra parte interesante del estudio es la utilización del análisis en tiempo real para el análisis de textos. Como se verá en la sección 5, este tipo de anális es una faceta altamente recomendable para el análisis visual de características dependientes del tiempo encontradas en grandes conjuntos de datos, ya que permite reducir considerablemente la carga cognitiva involucrada

en el proceso, hecho crucial para conseguir un análisis satisfactorio por otra parte.

En la misma línea de la tesis doctoral recién mencionada, destaca el trabajo y artículo de Roberto Therón et al., "Diachronic-information visualization in historical dictionaries"[30], que también ahonda en el análisis visual de información diacrónica albergada en un corpus textual proveniente de diferentes diccionarios históricos del castellano. En esta investigación se propone una solución que permite al investigador visualizar la evolución de la posición de los diferentes significados de un mismo lema a lo largo del tiempo, empleando para ello una serie temporal animada. Por añadidura se suma al prototipo desarrollado la capacidad de representar características espaciales y geográficas junto a las variables puramente semánticas, siendo éste por tanto un trabajo novel en el ámbito de aunar tiempo y espacio en el análisis visual de un corpus. Además, el artículo hace especial hincapié en la toma de decisiones a la hora de implementar herramientas visuales guiadas que permitan una correcta validación de los datos en ellas presentadas, lo cual ha servido de modelo para la realización del prototipo resultante de este trabajo.

Un proyecto que merece una sección aparte en nuestra reseña de trabajos relacionados, es el de Thomas Mayer et al., titulado "An Interactive Visualization of CrossLinguistic Colexification Patters"[31]. En este trabajo, el equipo de Mayer propone una solución visual en web al problema de visualización que presenta la base de datos CLICS, un recurso online que alberga asociaciones léxicas sincrónicas (también llamadados patrones de colexificación) de 200 lenguajes diferentes. El planteamiento es sencillo pero interesante: Se trata de representar tendencias en el uso de ciertas palabras para referirse siempre a los mismos conceptos en los mismos idiomas, y ver su evolución en el tiempo, así como comparar dichas tendencias en un lenguaje con las de otros lenguajes. Por ejemplo, afirman los autores, es un hecho que la gente tiende a asociar los conceptos "bueno" y "bonito" con cierta asiduidad. Sin embargo, no todo lo que es bueno es necesariamente bonito ni tampoco al revés pero a pesar de ello, estos dos conceptos son expresados con palabras idénticas en 27 lenguajes pertenecientes a 8 familias diferentes. A partir de este hecho, el lingüista puede plantear una serie de preguntas: ¿Dónde están estos lenguajes localizados en el globo terráqueo? ¿Cómo se distribuyen geográficamente las familias a las que pertenecen? ¿Qué otros conceptos son *verbalizados* empleando las mismas palabras? La respuesta a estas preguntas sin duda revelaría relaciones culturales entre las diferentes familias de lenguaje y otros tipos de conocimiento no aparente a simple vista.

En este aspecto, los autores tratan de ayudar a las investigadoras usuarias del prototipo a descubrir ciertas repeticiones de tendencias en el tiempo en un análisis visual exploratorio. Para ello, emplean un grafo dirigido de fuerzas[32], un recurso ampliamente usado para habilitar un análisis visual de redes. A través del uso de esta estructura, permiten obtener una buena perspectiva de la estructura general y posicionamiento de las palabras en torno a uno u otro concepto, permitiendo a su vez recuperar información bajo demanda cuando la usuaria lo exige. Una parte integral de esta visualización es un listado interactivo de todos los lenguajes que contribuyen a conformar las fuerzas de un patrón de colexificación. Para ello emplean también una escala de colores especial que permite una rápida identificación de los mismos.

Un problema que encuentran en el transcurso de su estudio es la alta densidad de nodos resultante en los grafos que representan relaciones de colexificación, ya que la muestra de la BD CLICS contiene una gran cantidad de palabras (301.498 para ser exactos), que cubren más de 1280 conceptos diferentes. En un primer procesamiento de los datos, se encontraron 45.667 casos de colexificación, que generaban un grafo con 16,239 enlaces entre el total de 1280 nodos, haciendo imposible la creación de una representación visual efectiva a partir de tal estructura.

Esta problemática es hábilmente solucionada mediante el uso de **comunidades** (para más información sobre las comunidades remitimos al lector a la sección 5, donde se desarrolla esta idea y se adapta al dominio del problema concreto de esta investigación). Mediante la aplicación de esta técnica, el equipo de investigadores consiguió crear 271 clusters o comunidades en las que se agrupan los 1280 conceptos inciales, permitiendo así la creación de grafos más pequeños y por tanto que puedan ser manejados por la usuaria del prototipo.

Una vez resuelto el problema, se muestran en el grafo dirigido de fuerzas que, debido a su funcionamiento, va a asegurar qué conceptos que están altamente relacionados aparezcan más cerca en la visualización. Esta vista del grafo se enlaza con otras dos que conforman la interfaz gráfica del prototipo web: Una vista sencilla de mapa con las proyecciones geográficas de cada una de los conceptos que aparecen en el grafo y una lista textual coloreada que relaciona los lenguajes y la familia a la que pertenecen con los elementos del mismo. En el aspecto de las técnicas de animación, emplean también el recurso visual del "highlighting", usado en los sistemas de vistas enlazadas para resaltar el conjunto de los datos que una usuaria selecciona en una vista en las otras y que posibilita la reducción efectiva de la carga cognitiva asociada al manejo de la aplicación (ver detalle en la Figura 6, para una demo online consultar [33]).

Los investigadores apuntan también en su publicación ciertos problemas que presenta el sistema propuesto: Primero, que la generación de comunidades, a pesar de ser de gran utilidad a la hora de mejorar la navegabilidad de los datos, condiciona completamente el resto del proceso, ya que todo el análisis se basa en esta construcción previa de los clústers. El algoritmo de creación de comunidades empleado, *Infomap*, no es obviamente a prueba de errores, y presentan limitaciones: Un nodo es asignado a una y sólo una comunidad, creando una estructura rígida inamovible, ofreciendo una visión parcial de los datos. En nuestro enfoque solucionamos este problema ofreciendo una solución más dinámica, que permite crear comunidades de manera automática en demanda por el investigador.

Pasamos a comentar ahora otro importante trabajo de investigación muy reciente (año 2016), perteneciente al grupo de Daniel A. Keim y centrado también en el análisis visual de patrones esta vez en series temporales de datos textuales. En el mismo se propone un *workflow* de dos pasos adaptado al entorno de aplicación del análisis de datos financieros [34], aunque los autores aclaran que podría adaptarse fácilmente a otros dominios de aplicación. A diferencia del trabajo anterior, en la producción de Wanner et al. sí se aplican métodos heurísticos de minería de datos, en una etapa de preprocesado de los datos mucho más compleja y elaborada, que

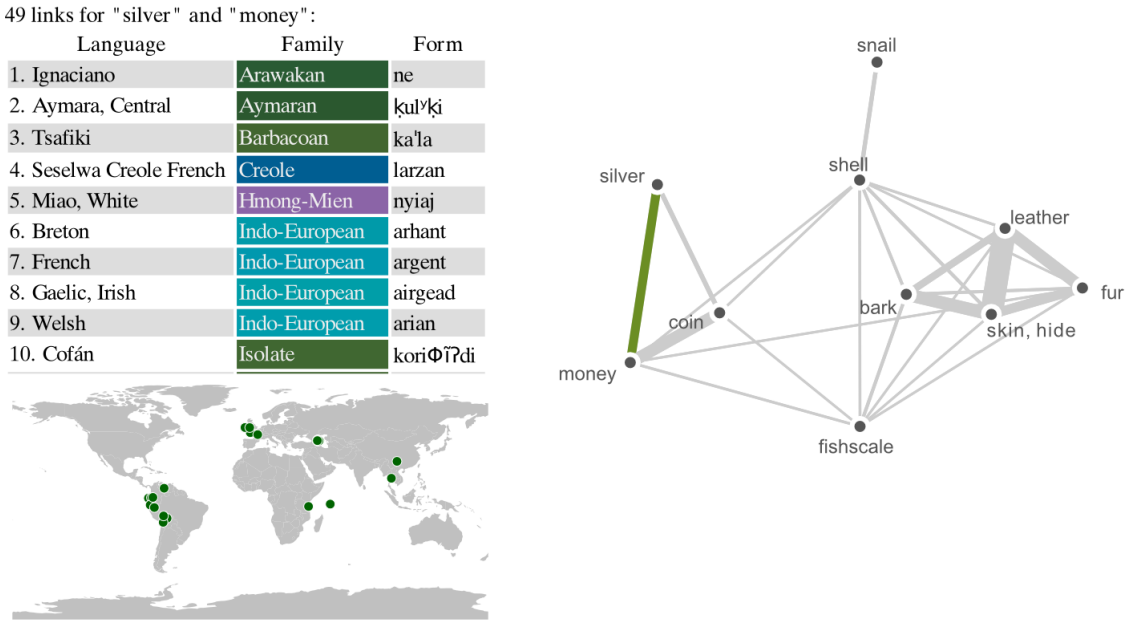


Figura 6: captura del prototipo propuesto por Mayer et al. que muestra las 3 vistas enlazas: el grafo dirigido de fuerzas (derecha), el mapa geográfico y la lista de términos (izquierda)

pasamos a explicar en el siguiente párrafo:

Como hemos dicho, el prototipo propuesto propone un flujo de trabajo compuesto de dos subflujos independientes ejecutados secuencialmente, en el que la salida del primero sirve de entrada para el segundo. En el primer flujo, se buscan patrones cuantitativos de intervalos temporales por medio de detección de puntos de interés (como subidas o bajadas abruptas en la cotización de una compañía) a los que se le aplican técnicas de clustering. Estos puntos de interés son detectados automáticamente por el algoritmo aplicando métodos estadísticos y son puestos en relieve a la usuaria empleando una combinación de colores.

En el segundo flujo, se emplean los patrones encontrados en el primero y se establece una correlación por el método a-priori, relacionando las co-ocurrencias de patrones de intervalos y las noticias. Esto es posible ya que se extraen características dependientes del tiempo y del texto (complejidad de las frases, vocabulario) de las propias noticias relacionadas con cada compañía de la que se analiza su trayectoria financiera. Por último, se visualiza la distribución de estas características textuales en un **gráfico de densidad** y una **matriz de adyacencia**. Empleando diferentes técnicas visuales, como la inspección bajo demanda de series temporales y propiedades textuales, se ayuda a generar un mapa mental en la analista sobre las dependencias y correlaciones entre los patrones encontrados y el contenido textual de las noticias. Se usa la técnica visual del **word cloud** para ofrecer resúmenes interactivos para los patrones y meta-patrones seleccionados por la usuaria (Ver Figura 7).

A pesar de que la temática y objetivos de la investigación de Wanner difieren

bastante de los planteados en nuestro estudio, éste supone un claro referente en términos visuales y de interacción que ha tratado de ser imitado hasta límites razonables. Especialmente del gusto del autor de este documento es la representación visual las palabras que conforman las noticias, y el análisis morfológico que se hace de ellas. Con todo y con eso, existen también diferencias importantes en el volumen de los datos a manejar por nuestro prototipo, que sigue siendo considerablemente más grande que el empleado en el citado trabajo. Ésto imposibilita la aplicación correcta de algunas de las citadas técnicas, habiendo sido necesario un desarrollo a mayores de las mismas.

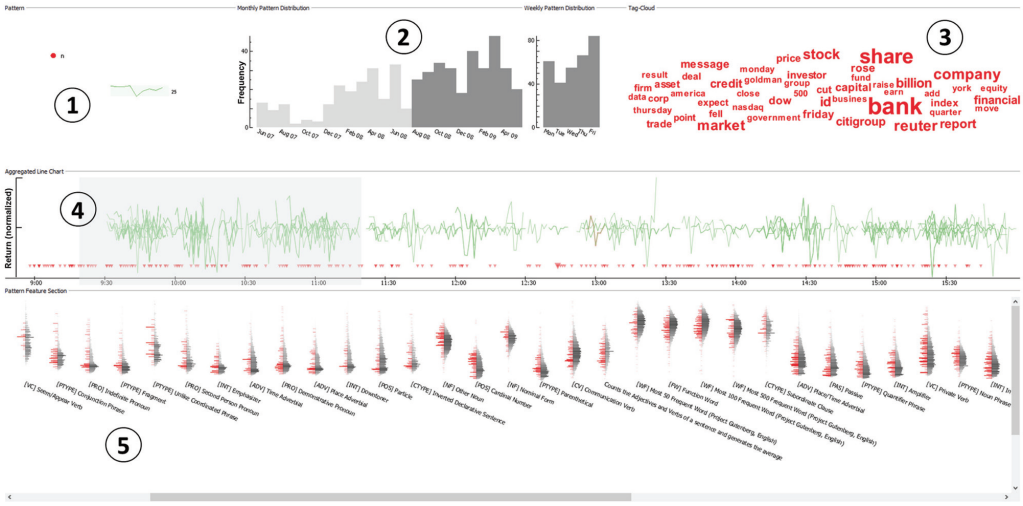


Figura 7: Las interfaz de 5 vistas enlazadas propuesta por Wanner. (1) Detalle del patrón temporal analizado, (2) Distribución en el tiempo, (3) Nube de palabras, (4) gráfico de líneas agregado con filtro activado, (5) Gráficos de densidad de las características textuales que muestran su distribución a lo largo del intervalo analizado.

Proveniente del mismo grupo de investigación de la Universidad de Salamanca, existe otro trabajo [1] que está también íntimamente ligado con el presentado en este escrito, ya que, entre otras características, emplea un conjunto de datos proveniente de diccionarios históricos del idioma alemán, con una alta similaridad al que se escogió para realizar esta investigación. Esta producción es un intento de romper con la metáfora de libro en el análisis visual de diccionarios, proporcionando recursos visuales que transcienden a la idea de lista alfabéticamente ordenada. Este giro en la interpretación de los datos y su posterior exposición posibilita en gran medida el descubrimiento de conocimiento otrora imposible de alcanzar. Remitimos al lector en este punto de vuelta a la ideas introducidas por Elijah Meeks en la sección 1.1, en las que se mencionaba la necesidad de aplicar métodos probados de maneras nuevas e inesperadas en las HD, ya que éste es un buen ejemplo de la consecución de tales objetivos.

El conjunto de datos empleado proviene de un proyecto de largo recorrido emprendido en 1911 por la Academia Austríaca de las Ciencias, el denominado Diccionario de los Dialectos Bávaros de Austria (WBÖ). En él se recopilan los léxicos empleados en diferentes partes del Imperio Austro-Húngaro y el Reino de Bavaria, con el objetivo de dar una visión detallada de los dialectos del idioma alemán,

así como crear un precedente en el estudio de la lexicografía de los mismos. En estas recopilaciones se anotaron in-situ definiciones detalladas de las palabras, así como codificación gramatical, etimología, pronunciación y demás conocimiento experto. La mayoría de estas adquisiciones se hicieron por medio del uso de cuestionarios, que se entregaban a las gentes del lugar para ser rellenados. Además, los expertos realizaron en muchas ocasiones entrevistas de alto valor histórico que sirven hoy en día para contextualizar correctamente los diferentes usos del lenguaje que se daban en aquellos tiempos. Se remite al lector a la sección 3.3 para más información sobre este conjunto de datos, que supone una parte del usado en la realización del estudio presentado en estas líneas. Estos datos, que vienen siendo progresivamente digitalizados y clasificados desde el año 2010[35], suponen una valiosa fuente de información para historiadores y lexicógrafos en nuestros días, y es por tanto que se presenta el reto de ofrecer herramientas visuales que permitan el correcto acceso a los mismos.

Como se decía, en esta producción investigadora se acomete el reto de representar cambios en el corpus alemán proveniente de diccionarios históricos a lo largo de un período de 100 años en una herramienta de análisis visual exploratorio. Para ello, el prototipo emplea una proyección geográfica de los diferentes puntos donde se podían encontrar las llamadas fuentes de cada lema del que se dispone de información digitalizada. En el mapa se agregan dichos resultados mediante la técnica del "bubble map", o mapa de puntos graduados y un algoritmo computa una serie de características para cada agregación, que a su vez son representadas en un conjunto de coordenadas paralelas[9].

Como se puede observar en la Figura 8, cada burbuja del mapa esta representada en una polilínea en el gráfico de coordenadas paralelas, que muestra a su vez el número de subregiones, lemas, personas (autores de los cuestionarios y/o registradores de los términos lingüísticos) y documentos de cada una. A la derecha, se presentan los datos originales extraídos de la BD, que permiten a la usuaria del prototipo disponer de toda la información disponible. Es así que mediante el proceso de análisis visual planteado en la herramienta, la investigadora va a poder descubrir patrones relacionados con la distribución de los datos en base a su posición geográfica (la herramienta guía el proceso investigador). Una vez descubierto un hecho llamativo, el foco de atención de la usuaria permanece en una sección o subconjunto de los datos gracias al filtrado interactivo que se hace de los mismos, y finalmente los detalles se proporcionan bajo demanda a la usuaria, que son ocultados en una primera fase en la que serían innecesarios y abrumadores para la usuaria.

Esta técnica, conocida coloquialmente como el "mantra de la visualización", que reza: **Vista general primero, zoom y filtrado, finalmente detalles bajo demanda**[36], es de probada y reconocida efectividad, y éste es un buen ejemplo de cómo aplicarlo en el ámbito de las HD en general, y al de un problema histórico-lingüístico en particular. Volveremos en secciones subsiguientes a este tópico ya que ha sido tenido en cuenta de manera importante en todas las etapas de concepción del prototipo propuesto.

A pesar de la calidad de los resultados, debemos decir que la funcionalidad de este prototipo es limitada y además carece de una etapa de procesamiento de los

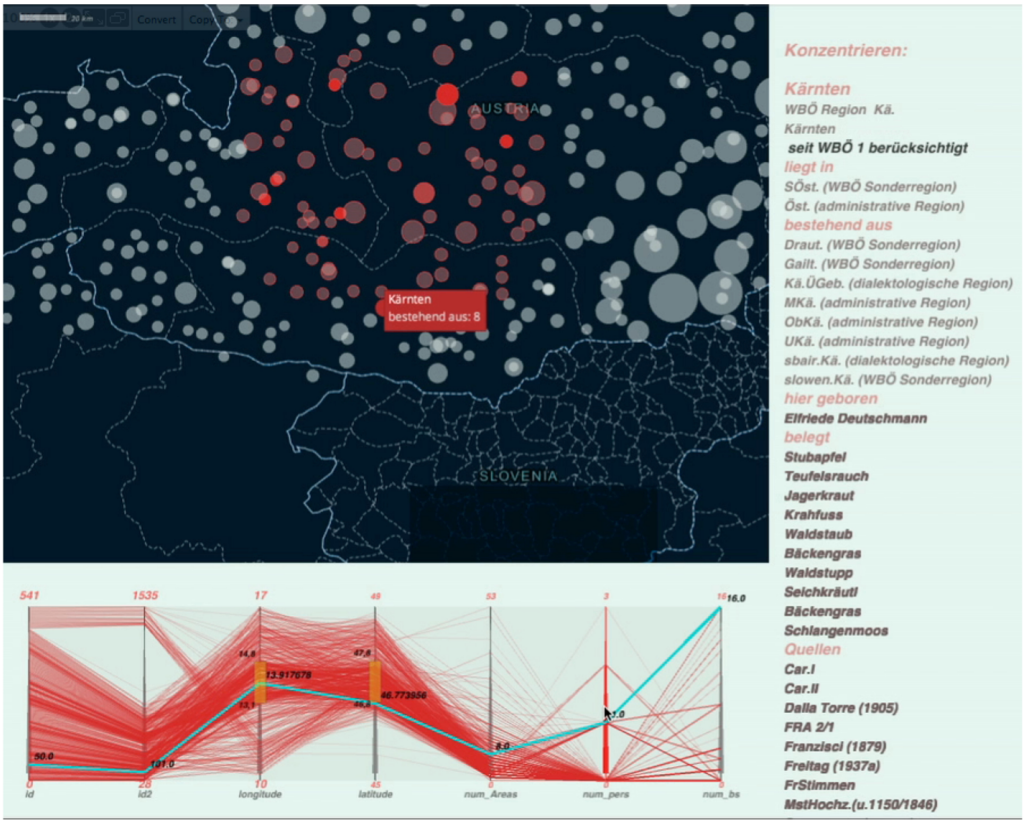


Figura 8: Captura de la interfaz del prototipo, que muestra las tres vistas enlazadas: "bubble map" (arriba), coordenadas paralelas (abajo) y datos originales (derecha)

datos que podría haber mejorado la utilidad del prototipo en gran medida, ya que el algoritmo propuesto se limita a cuantificar ciertas variables en base a las diferentes zonas geográficas, pero no expone conocimiento extraído a través de ninguna técnica de minería de textos. Además, el volumen de datos que maneja es excesivamente pequeño, y no ofrece ninguno tipo de filtrado textual ni espacial, que aumentaría considerablemente las prestaciones del software y su valor como herramienta de apoyo a la investigación. En la sección 5 se comentan ampliamente los métodos y mejoras implementados en el sistema propuesto que tratan de suplir estas carencias.

2.3. Proyectos de referencia

En esta sección compilamos una serie de proyectos y recursos online en el ámbito de las HD, que por diversos motivos han sentado referente en la aplicación de los principios discutidos en las primeras secciones de este trabajo pero que no han generado necesariamente publicaciones en revistas científicas del ramo. Por este motivo, y a diferencia de lo que ocurría con los trabajos de la sección anterior, nos centraremos más en los aspectos arquitectónicos de los mismos, así como en aplicaciones de las técnicas de visualización e interacción que resuelvan problemas muy concretos en dominios del problema con alto nivel de semejanza al de nuestro caso. Es importante también remarcar en este punto la acusada tendencia que existe

en el ámbito de las HD a crear proyectos atractivos artísticamente hablando, siendo este un hecho que no podía ser ignorado en esta investigación. Por estas razones se incluyen también trabajos en este apartado que, a pesar de no destacar en su aspecto puramente académico, científico o computacional, son incluidos como modelo de belleza estética e imaginación para conseguir crear un impacto emocional en la usuaria.

Muchos de estos proyectos, como veremos, emplean conjuntos de vistas enlazadas en planteamiento de herramienta visual, y casi siempre la proyección geográfica de los datos supone el núcleo del sistema, dando lugar a exploraciones visuales dirigidas espacialmente, de alto poder comunicativo gracias a su potencia y facilidad de uso por parte de perfiles de usuaria muy variados, por las razones ya citadas en los comienzos de este documento.

2.3.1. ORBIS

Comenzamos la serie de reseñas citando el que seguramente ha sido el proyecto que más influencia ha ejercido sobre el sistema de software propuesto en este estudio. Se trata de la obra de el ya citado humanista digital y experto en visualización de la Universidad de Stanford, Elijah Meeks. No sólo es referencia importante este proyecto por su enfoque visual de la solución a un problema de humanidades, sino también por la gran producción en forma de artículos de blog y diversa literatura online que dicho autor generó durante la producción de este trabajo [37]. La maestría de Meeks en el ámbito de la Visualización de Datos queda patente en todos los aspectos relacionados con el proyecto, así como su gran capacidad como comunicador y humanista.

El proyecto ORBIS[38] aborda el problema de la exploración del Imperio Romano, centrándose en las rutas comerciales que se utilizaban para viajar de una a otra parte del mismo. Mediante una acertada aplicación de diferentes algoritmos de navegación de grafos, en combinación con un sistema de proyecciones por capas, posibilita a la usuaria realizar una exploración visual del tiempo y el coste estimados del viaje en la época, dando una representación moderna de ideas muy antiguas. A su vez, existe la posibilidad, de como en los sistemas de direcciones comerciales de hoy en día, preferir rutas más rápidas, más baratas o más cortas en las diferentes estaciones del año, para distintos medios de transporte y cargas, de acuerdo a fuentes de datos históricas fidedignas.

Aplicando algoritmos de pathfinding el sistema es capaz de calcular dinámicamente en el grafo distancias entre ciudades por distintos parámetros, como la duración del viaje o el coste. Empleando técnicas de inundación del grafo analiza las rutas más usadas en la época y otras características. Además, Meeks implementa acertadamente como se ha dicho, diversas técnicas visuales. Una de las más notorias es la **deformación de las proyecciones geográficas** en base a la forma que adopta el grafo de acuerdo a los criterios de búsqueda introducidos. Otros recursos visuales utilizados en la aplicación son la generación de celdas de Voronoi, que consigue dividir el mapa visualmente en base a una búsqueda, o el uso del "convex

hull" para representar conjuntos de puntos relacionados. En la Figura 9 podemos ver algunas de estos recursos visuales.

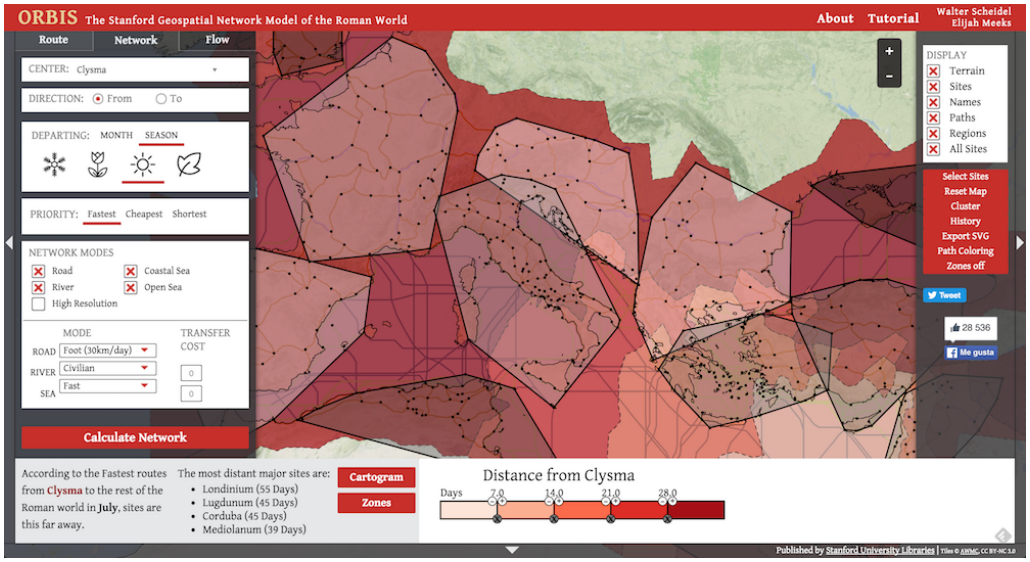


Figura 9: Análisis de la red de transportes generada por un punto del mapa. Vemos como el algoritmo que maneja el grafo agrupa puntos con costes de viaje semejantes, agrupamiento que es representado visualmente por medio del "convex hull" de los puntos.

2.3.2. TransVis

El proyecto TransVis[39] proporciona una visualización de los lugares en los que se redactaron libros y escritos de Shakespeare. Cada localización detectada incluye el siglo y el año concreto en que se compusieron los textos. En caso de conocerse, se incluye también el nombre exacto del lugar. Ofrece un filtrado temporal que actualiza los resultados del mapa, y genera pequeños grafos bajo demanda que vinculan localizaciones relacionadas (por ejemplo el lugar donde se escribió con el lugar en el que finalmente se publicó cierta obra). Además de habilitar la exploración espacial, el sistema ofrece la posibilidad de hacer una búsqueda dirigida seleccionando una obra concreta, que es a su vez remarcada en el mapa.

2.3.3. The Contested Corners of Asia: A Visual Companion

Este trabajo surge de la iniciativa de la *Asia Foundation* por crear un repositorio *online* sobre conflictos de guerra modernos (1974 a 2010) en dicho continente. Dentro de los diferentes recursos ofrecidos, se creó una visualización espacial que permite ver los puntos de conflicto en el mapa. Cuenta con una línea temporal que se actualiza junto con la animación principal, que recorre cronológicamente los diversos lugares afectados por combates. Esta sencilla pero acertada animación merece una mención aparte ya que transmite perfectamente la idea de movimiento y cambio en las diferentes localizaciones a la usuaria, pero sin resultar molesta o imposibilitar la

exploración a la misma. Permite un filtrado muy básico seleccionando zonas interactivas o "puntos calientes" del mapa, que dan información detallada en demanda acerca de los conflictos.

2.3.4. Manifest Destiny

En esta visualización interactiva[40] se representa la evolución de los distintos territorios y regiones de los Estados Unidos de América. Emplea múltiples mapas para destacar la evolución de las regiones en base a los distintos propietarios. Se da una vista global de todos los mapas o individualizada, incluyendo en cada caso textos y pasajes relacionados con los cambios experimentados por las fronteras entre los territorios.

2.3.5. Geography of the Post

En esta iniciativa por parte de investigadores del Centro de Humanidades Digitales de la Universidad de Stanford se analiza el periódo histórico llamado como la "Conquista del Oeste", que tuvo lugar entre los años 1846 y 1902. En la visualización ofrecida, se representan las aperturas y cierres de oficinas postales dentro del período analizado, como medio para comprender la migración de colonos hacia el Oeste de los EEUU. Se ofrecen dos vistas enlazadas, un mapa que muestra las localizaciones de las oficinas y un *timeline* que habilita un filtrado temporal de los resultados que se dibujan en el mapa.

3. Descripción del problema

3.1. Introducción

En el siguiente apartado de esta memoria se definen los objetivos y motivaciones iniciales de los que parte la investigación. Se explican las razones por las cuales en un comienzo se pensó en crear un sistema como el resultante de este estudio, cuyas particularidades son introducidas en la sección 5. Comentamos también los dos conjuntos de datos que se han manejado en el transcurso de la concepción de este trabajo, así como su historia, orígenes y otras características relevantes para el lector.

3.2. Objetivos

La interacción de los investigadores con diccionarios históricos y corpus lingüísticos a través de computadores ha supuesto históricamente un reto importante para muchas clases de académicos y entusiastas de las HD. Como se vió en la sección de *Trabajos relacionados*, los métodos de los que disponemos en nuestros días para su indexación, clasificación, almacenamiento, estandarización y estudio son resultantes del esfuerzo de muchas personas a lo largo de más de 50 años. Sin embargo, aún hoy en día existe una carencia de soluciones que enfoquen el análisis de los mismos de una manera visual adecuada y muy especialmente en el ámbito de los diccionarios históricos.

Muchas de las aplicaciones que ofrecen acceso *online* a estos repositorios históricos, sufren importantes carencias en relación a su usabilidad, experiencia de usuario, validez y utilidad de las técnicas de analisis visual. En el caso que nos concierne, parte de los datos han sido adaptados y estandarizados previamente por otros investigadores, e incluso se han llevado como decimos a la web, en un intento de crear sistemas accesibles para investigadores y la ciudadanía en general. Aún hoy en día la mayor parte de los datos están almacenados en estándares como SGML o XML, que hacen muy difícil su puesta en línea sin llevar a cabo antes un proceso de migración masivo altamente complejo. Existen también en la actualidad intentos de migrar esta información a formatos más actuales, como OpenLink/OpenData[41], que permitirían exponer el contenido en la novedosa y actual Web Semántica, lo que resolvería muchos de los problemas de formato. No obstante, este proceso es laborioso y se encuentra en una etapa muy primigenia aún.

Como ya se ha explicado, estos formatos estándar fueron específicamente diseñados por académicos del ámbito de la lingüística computacional y la lexicografía, y es precisamente en esas áreas donde reside su potencia y es el único uso viable que se le ha podido dar a los datos en ellos codificados. Es un hecho constatable que a pesar de los grandes avances en muchas de las disciplinas de las ciencias de la computación, lexicógrafos de todo el mundo siguen empleando los mismos métodos de trabajo y continúan teniendo la misma perspectiva de los datos que se tenía

hace décadas. Este anquilosamiento de las metodologías es debido en gran medida al fuerte acoplamiento existente entre los datos y sus representaciones formales lo cual hace que estas disciplinas no puedan beneficiarse plenamente de nuevos avances tecnológicos que serían sin duda altamente útiles para la adquisición de nuevo conocimiento resultando, el autor se aventura a opinar, en una traba importante a la investigación.

Ya hemos introducido estudios que han probado la utilidad del análisis visual en diferentes campos de la lingüística y la lexicografía (recuerde el lector por ejemplo los trabajos del grupo de investigación del Prof. Daniel A. Keim). Es por tanto razonable pensar a priori que las mismas metodologías son aplicables al contexto de los diccionarios históricos, la dialectología y la lexicografía dialectal, más aún si cabe teniendo en cuenta el hecho de que ya se han realizado intentos concretos de crear Sistemas de Información Geográfica a partir de las ya citadas migraciones parciales de los datos.

3.2.1. Una crítica al actual modelo

Si observamos algunas de las iniciativas llevadas a cabo para interactuar con textos relativos a diccionarios históricos, se repite en ellas el hecho negativo de que las representaciones conceptuales de los mismos siguen interpretando el concepto de diccionario bajo el paradigma clásico de una lista de palabras ordenada alfabéticamente. Obviamente esta es una generalización demasiado pobre para definir un diccionario, pero sin duda afirmar que un diccionario es una lista ordenada de palabras es tautología. Sin embargo, y este hecho es importante ya que pone de manifiesto el cambio de paradigma en la concepción del concepto de diccionario, esta lista no ha de estar ordenada necesariamente por orden alfabético. La preasunción de este concepto a la hora de crear sistemas de análisis visual de diccionarios supone, según las tesis defendidas por varios autores[1][30], un error fatal a la hora de diseñar este tipo de aplicaciones del que es necesario alejarse.

En el enfoque propuesto en esta investigación, el orden de los elementos en el diccionario pasa a ser completamente multidimensional, y pasa a tener en cuenta características temporales, espaciales o morfológicas, (o una combinación de ellas), habilitando una nueva forma de mirar a los datos con el objetivo de permitir el descubrimiento de conocimiento y el advenimiento de nuevas metodologías útiles en el dominio del problema.

Se hace por tanto necesario el desarrollo de aplicaciones que posibiliten un análisis visual y multidimensional de los datos de diccionarios históricos, convirtiéndose ésta en la principal motivación de la investigación aquí desarrollada.

En las secciones siguientes presentaremos en más detalle el Diccionario de los Dialectos Bávaros de Austria (Wörterbuch der bairischen Mundarten in Österreich [WBO]), su origen y estado actual, que ha servido de conjunto de datos en la primera parte del desarrollo de esta investigación. También nos detendremos en TUSTEP[42], una suite de herramientas de procesamiento de textos desarrollada en la Universidad de Tübingen (Alemania), que ya fue mencionada en la sección 2.1 y más especial-

mente en la sintaxis XML de TUSTEP, TXSTEP.

3.3. Conjunto de datos

3.3.1. TUSTEP

En las secciones iniciales decíamos que TUSTEP fue una suite de procesamiento de textos académicos concebida en los años 60 gracias a los esfuerzos de investigadores en la Universidad de Tübingen. Este sistema (evidentemente revisado y actualizado) continúa siendo empleado en la actualidad por investigadores de la lengua alemana. Con posterioridad, se concibe una interfaz en XML para dar soporte a la suite, denominada TXSTEP. Los registros, anotaciones y fuentes de los diferentes lemas son por tanto almacenados en este formato, a los que los investigadores acceden por medio de *queries* en formato XQUERY. Este formato de búsquedas, altamente versátil y parte del estándar XML, permite recuperar registros de texto albergados dentro de colecciones de documentos XML usando una sintaxis parecida al lenguaje SQL de BBDD. Empleando este lenguaje de búsquedas, los investigadores realizan habitualmente en sus investigaciones búsquedas textuales dentro del repositorio de datos, que es un conjunto bastante extenso de documentos XML. Cabe reseñar que en la actualidad no toda la información perteneciente a diccionarios históricos en Austra ha sido migrada a XML, poniendo de relieve un problema de gran importancia para la Academia de las Ciencias Austríaca en la actualidad: La **fragmentación de los datos**.

En la actualidad una gran parte de los documentos de texto de los que dispone la Academia de las Ciencias Austríaca se encuentra dividida en diversos subconjuntos: Primeramente existen datos que aún no han sido digitalizados en absoluto o no en su totalidad (El proyecto DBÖ sigue encargándose de esa tarea para el diccionario WBÖ). Segundo, entre los diccionarios que sí han sido digitalizados existen diferentes conjuntos de datos resultantes de los esfuerzos de diversos proyectos con objetivos similares a DBÖ pero empleando otros diccionarios como fuente.

En nuestro caso manejaremos sólo datos provenientes del proyecto DBÖ que, debido a esta fragmentación de la que hablábamos, son dados en dos formatos diferentes: MySQL y XML, que usaremos para distintos propósitos y que pasamos a comentar a continuación:

3.3.2. WBÖ y dbo@ema

No nos detendremos demasiado en esta sección en la historia del WBÖ, ya que fue ya reseñado en la sección de Trabajos Relacionados 2 de este documento. No obstante, repararemos con más énfasis en el análisis cuantitativo y en los formatos digitales de los datos de dicho diccionario. La información relativa a este diccionario (normalmente presente en formato papel) comenzó a ser progresivamente digitalizada en 1993 a través de otra iniciativa del país austríaco: El proyecto “Datenbank

der bairischen Mundarten in Österreich" (DBÖ). La suite empleada para el procesamiento de estos textos una vez digitalizados fue TUSTEP.

Como explicamos en la sección 2, el método de trabajo tradicionalmente empleado por la academia para registrar los usos del lenguaje se basaba en el uso de cuestionarios. En ellos se pedía a la población que respondiese a preguntas específicas sobre el léxico que empleaban para referirse a conceptos relacionados con cierta temática. La información recopilada por la persona encargada de esta tarea era también complementada con trabajo de campo, normalmente en forma de entrevistas personales. En ellas se concretaban los conceptos que hubiesen aparecido en los cuestionarios y que fuesen de especial interés lingüístico por diversas razones. Al final de este proceso la lexicógrafa trataba de dar una definición lo más exacta posible de cada término que iba a parar a una tarjeta. Estas tarjetas son también especialmente importantes por el hecho de que muchas veces (dependiendo de la habilidad de la lingüista), contenían dibujos descriptivos que ayudaban a contextualizar el concepto. Por ello eran recopiladas, ordenadas y almacenadas en registros junto a los cuestionarios y servían de base para futuras investigaciones.

Refiriéndonos de nuevo al proceso de digitalización de fuentes llevado a cabo por iniciativas como DBÖ, y como ha sido costumbre en las HD, éste se centró en la transcripción de los textos originales de estas tarjetas que se mencionaban en el párrafo anterior. No obstante, gracias al reciente abaratamiento de las tecnologías de almacenamiento de datos, existe una nueva línea de trabajo en la que se producen imágenes digitales de alta calidad de las tarjetas y cuestionarios originales, así como de otros artefactos resultantes de la producción investigadora de la academia (mapas, cartas, volúmenes de diccionarios o fotografías), en un intento de perpetuar la información que contienen. Hoy en día se sigue trabajando en esta tarea, pero a pesar de los grandes esfuerzos invertidos, sólo una mínima porción de los originales cuenta con una versión digital completa (transcripción textual e imagen digital). En la Figura 10 podemos ver las versiones digitales de un cuestionario realizado en Alta Austria a principios del S.XX. y una tarjeta creada en los mismos años y zona.

Es debido al comienzo del proyecto dbo@ema (DBÖ Electronically Mapped) y en especial a la necesidad de dotarlo de una interfaz web que nuestro primer subconjunto de datos tiene su origen. Esta interfaz web implementa un nuevo sistema de búsquedas compatible con la web basado en MySQL que permite a las usuarias realizar consultas por lema, bibliografía, persona o lugar. Es así que para poder llevar a cabo este proyecto, se realiza una primera migración de registros en formato XML a MySQL. Como además uno de los objetivos del proyecto era ofrecer una visión espacial de los lemas del diccionario WBÖ, se genera esta información GIS a partir de un proceso de minería de textos, en el que se asocia una cadena (el nombre de un lugar) a su posición geográfica en el mapa (proceso denominado como *geocoding*. Esta posición se almacena en la BD empleando las extensiones geo-espaciales de MySQL[43], que posibilitan el almacenamiento de secuencias binarias que codifican dicha información, así como la realización de consultas SQL espaciales (Ver Figura 11). Es importante señalar que, debido al carácter histórico de los datos que se tratan, el proceso de geocoding llevado a cabo no es tan sencillo como en el caso de tratar con datos del presente. Las fronteras geográficas de Austria han cambiado

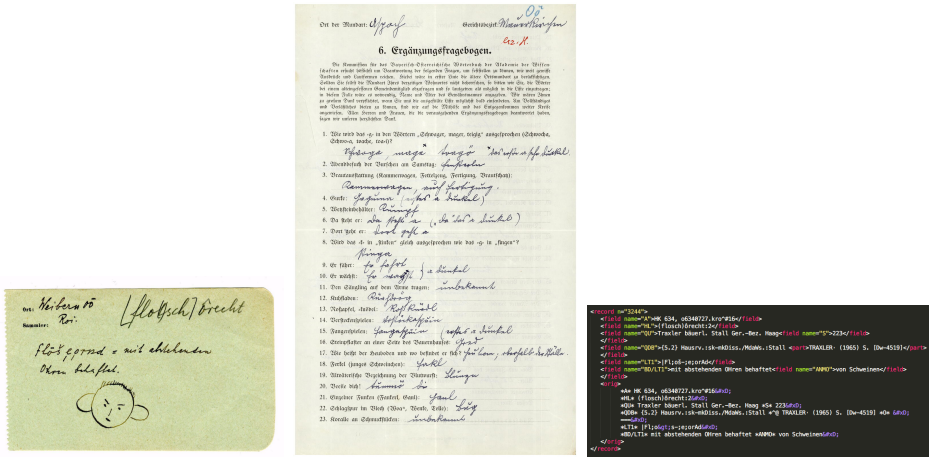


Figura 10: Izquierda: Una tarjeta con notas y un dibujo que contextualizan la palabra definida: “floschorecht” (orejudo). Izquierda: Copia escaneada de un cuestionario de principios de siglo realizado en la región de la Alta Austria. Derecha: Detalle del registro TUSTEP que hace referencia a la tarjeta. En el campo “BD/LT1” que hace referencia al significado, se lee la transcripción de la nota original: “mit wegstehenden Ohren behaftet” (afectado por orejas grandes).

significativamente a lo largo de los años y en un análisis inicial de los datos hay fuentes que datan de los tiempos medievales (El año más antiguo encontrado es el 1040). Es por esta razón que existe la posibilidad de que algunos de los sitios hayan cambiado de nombre, que no pertenezcan a la Austria actual (este hecho es habitual) o que ni siquiera existan. Este trabajo de geocodificación de los datos ha de estar supervisado por expertos en humanidades como historiadores o geógrafos, que puedan dictaminar la corrección de los resultados obtenidos y es por esta razón que esta parte de la información contenida en dbo@ema es tan altamente valiosa para el estudio.

A pesar de considerar dbo@ema como un prototipo avanzado que pretendía exponer diccionarios históricos, este enfoque, como se ha indicado previamente en la introducción presenta varios problemas: A los ya citados de interacción, experiencia de usuario y aplicación de los patrones de visualización, se suma ahora el de la estructura de datos. Al efectuar la migración de los mismos desde un sistema abierto y con formato extensible y abierto como es XML a una estructura relacional mucho más rígida, se pierden muchas de las ventajas de búsqueda e indexación con las que se contaba con el otro enfoque, además de hacer el sistema mucho más difícil de actualizar y mantener. La realidad es que MySQL no fue concebido como una base de datos documental y por lo tanto el enfoque no es óptimo para solucionar la problemática planteada. Es por eso que es difícil encontrar soluciones de almacenamiento y búsqueda de textos que se apoyen exclusivamente en estructuras relacionales, ya que las razones para tal elección serían difíciles de justificar.

Al no tener los documentos (por el hecho de ser documentos) una estructura fija preestablecida (Esto quiere decir que pueden tener un número variable de campos), hace que al ser forzados a seguir una estructura relacional se produzcan de manera

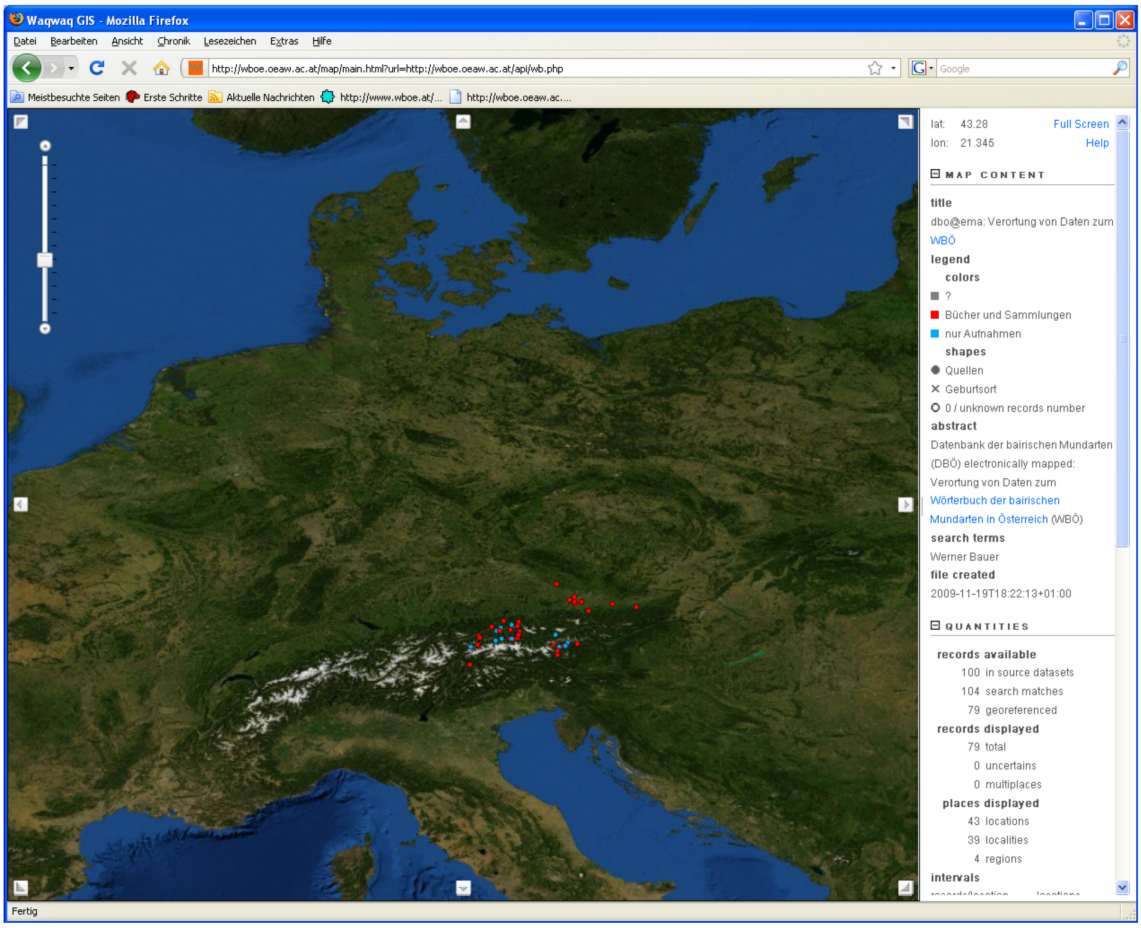


Figura 11: Detalle de la interfaz de dbo@ema mostrando la localización de una entrada de la base de datos en el mapa e información asociada.

natural, graves deficiencias que afecten a la estabilidad, seguridad, rendimiento y escalabilidad de todo el SGBD, y que quizás por el pequeño volumen de los datos manejados (alrededor de 100.000 entradas) no se pusieron de manifiesto de forma aparente desde un principio. Un ejemplo de esta problemática es la aparición de elevados porcentajes de entidades con campos vacíos, o la generación de relaciones tan dispares que harían imposible conseguir un sistema normalizado y óptimo.

En el proceso de la investigación, se generó un diagrama relacional por ingeniería inversa de la BD, que arrojó a la luz los problemas de los que adolecía el enfoque seguido en dbo@ema. (Ver Figura 12)

Por tanto una de las primeras conclusiones a las que se llegó después de entrar en contacto con el citado conjunto de datos fue que, partiendo de la premisa de un sistema que destacase por su facilidad de uso y buena experiencia de usuario, ésta iba a ser muy difícil o imposible de conseguir mediante el empleo de modelos relacionales. Fue por entonces cuando se comenzaron a buscar alternativas más acordes con la naturaleza del problema. Durante la investigación nos percatamos de que no todos los problemas se arreglarían con un cambio en la elección del soporte de los datos: Si se prescindía del SGBD y se volvía a usar XML se perdería el ingente trabajo de

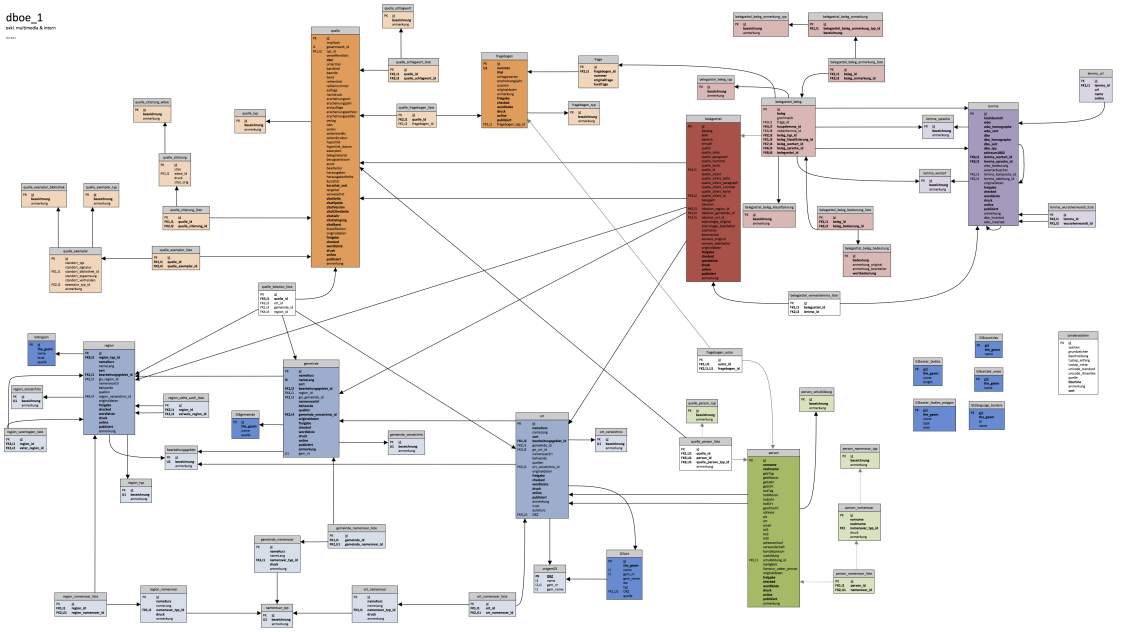


Figura 12: Diagrama de la BD MySQL empleada en dbo@ema. El número de entidades es exageradamente grande para la complejidad del dominio del problema. Se observa también una caótica distribución de las relaciones entre entidades.

geocoding almacenado en MySQL, llevándonos esta tesitura a hacernos la pregunta: ¿Cómo podríamos añadir funcionalidad sin ir en detrimento de la ya existente?

Adelantamos ya que la respuesta fue seguir un enfoque híbrido, que nos permitiría no tener que volver a realizar todo el proceso de geocoding y a la vez no forzar una estructura relacional en los datos. Las conclusiones a las que llevó el desarrollo de este enfoque son discutidas en la siguiente sección.

3.3.3. TUSTEP-XML

Por las razones comentadas en el anterior apartado, la idea dotar de una estructura relacional a los datos de los que se alimentaría el prototipo surgido de esta investigación fue descartada desde un primer momento. Por suerte, se tuvo la posibilidad de trabajar con otro subconjunto de los datos en formato original TUSTEP-XML de tamaño mucho mayor. Éste consistía en cerca de 9 GB de datos repartidos en un total de casi 3000 ficheros XML ordenados alfabéticamente. Cada uno de estos ficheros contiene una serie de registros TUSTEP codificados como entidades XML. Esta fuente de información fue generada gracias al proyecto DBÖ, que había sido el encargado de transformar los textos anotados que originalmente empleaba la suite TUSTEP en TXSTEP.

Cada uno de los registros a su vez representaba una fuente registrada (cuestionario, cita bibliográfica) de uso de un lema. A través de un conteo inicial se estimó el número total de referencias en más de 2 millones, un número considerablemente

alto en comparación con los conjuntos de datos que se manejaban en dbo@ema y en otros proyectos de referencia en HD. Este número, lejos de resultar abrumador, supuso otro de los retos principales de la investigación: el concebir un prototipo de exploración visual interactivo que pudiese manejar esta cantidad de datos de manera transparante.

Continuando con el aspecto y formato de los datos, la estructura de los datos está basada en un estándar (XML), lo que permite efectuar tareas de minería de datos y recuperación de la información de manera mucho más eficiente que si se tratase directamente con texto plano. No obstante, el contenido que es delimitado por el lenguaje de marcado, al provenir de textos relativamente libres (después de un análisis preliminar a gran escala se advirtió el uso de ciertas convenciones de estilo en las anotaciones de los académicos, pero no tan rigurosas como para constituir un lenguaje formal) requirió de un procesamiento heurístico más complejo que permitiera la extracción de características del mismo.

En la Figura 13 se adjuntan dos registros TUSTEP-XML elegidos al azar de ficheros pertenecientes a las letras "o" y "d", respectivamente.

Como se puede observar en la figura 13, existen concordancias y disonancias en el formato de los registros en cada caso. El registro localizado en la parte superior de la imagen cuenta con un campo "O", que denota el lugar de procedencia de la fuente citada, en este caso *Joglland in St*, que no está presente en el de la parte inferior. Este campo puede hacer referencia a una población, a una comunidad o a una región, en orden creciente de extensión. De manera análoga, el campo "ETO" del registro inferior, que denota el origen etimológico falta en el de arriba. Hay un total de 38 campos diferentes que contienen diferentes tipos de informaciones referentes al registro de la fuente, desde anotaciones manuales del investigador que creó la tarjeta original hasta el número de cuestionario del que proceden (si se aplica el caso).

Ya que la explicación pormenorizada de los 38 tipos de campo del estándar TUSTEP-XML resulta innecesaria para el objeto de esta investigación, pasamos a reseñar sólo los dos campos comunes a todos los registros que han sido sometidos a tratamiento textual de algún tipo con el objetivo de generar características textuales útiles para el estudio.

- "HL": *Hauptlemma* o lema principal. Especifica el lema o palabra principal a la que hace referencia la fuente. Este tipo de contenido puede aparecer dividido en dos raíces léxicas, que harían referencia a otros dos lemas.
- "QDB": *Quelle/normiert* o fuente estandarizada. Este campo surge del esfuerzo de los lexicógrafos por estandarizar el formato textual de citación de fuentes y sustituye al campo "QU", que en nuestro estudio ha sido ignorado (aunque se encuentra presente en muchos de los registros). De este campo se extrae, mediante minería de textos, la información temporal y el nombre del lugar asociados al registro, en caso de estar presentes.

En esta sección hemos presentado los conjuntos de datos que sirven de punto de entrada para el desarrollo de la investigación. Como hemos intentado remarcar, nin-

```
<record n="1074">
   <field name="A">HK 631, 06130618.PUD^#17</field>
   <field name="HL">oben:3</field>
   <field name="QU">Eiselt, Slg. Umg. Vorau, St. Nachtrag (1984)<field name="S">566</field>
   </field>
   <field name="QDB">{3.5b,3.5c,3.5g} nöOSt. (v.1950)
      <part>WbMs.EISELT· (1984) Nr. [HA-5086:"Joglld."; wrTr] ** Exz.Bosmanszky:1985</part>
      <field name="O">Joglland in St.</field>
   </field>
   <field name="LT1">-ou·m</field>
   <field name="BD/LT1">oben</field>
   <field name="KT1">-ou·m |And int<field name="KL">Fg.</field>
   </field>
   <field name="BD/KT1">da und dort</field>
   <orig>
   *A* HK 631, 06130618.PUD^#17&#xD;
   *HL* oben:3&#xD;
   *QU* Eiselt, Slg. Umg. Vorau, St. Nachtrag (1984) *S* 566&#xD;
   *QDB* {3.5b,3.5c,3.5g} nöOSt. (v.1950) *^@ WbMs.EISELT· (1984) Nr. [HA-5086:"Joglld."; wrTr] ** Exz.Bosmanszky:1985 *O* Joglland in St.&#xD;
   ===&#xD;
   *LT1* -ou·m&#xD;
   *BD/LT1* oben&#xD;
   ===&#xD;
   *KT1* -ou·m |And int *KL* Fg.&#xD;
   *BD/KT1* da und dort&#xD;
   </orig>
</record>
```

```
<record n="4">
   <field name="A">HK 153, d1530910.pir, korr. I.G.</field>
   <field name="HL">(Glatt)dick:1</field>
   <field name="QU">Steir.Wb.<field name="S">293</field>
   </field>
   <field name="QDB">{3} ^@ SteirWb.(1903) S. HA-3600a-e [Ausg. 1968]</field>
   <field name="LT1">Glattdick [m]</field>
   <field name="BD/LT1">der zur Familie der Störe gehörige Accipenser glaber</field>
   <field name="ETO">s. Heckel=Kner Sw.F. 332</field>
   <orig>
   *A* HK 153, d1530910.pir, korr. I.G.&#xD;
   *HL* (Glatt)dick:1&#xD;
   *QU* Steir.Wb. *S* 293&#xD;
   *QDB* {3} ^@ SteirWb.(1903) S. HA-3600a-e [Ausg. 1968]&#xD;
   ===&#xD;
   *LT1* Glattdick [m]&#xD;
   *BD/LT1* der zur Familie der Störe gehörige Accipenser glaber&#xD;
   *ETO* s. Heckel=Kner Sw.F. 332&#xD;
   *DBO* x-&#xD;
   </orig>
</record>
```

Figura 13: Comparativa de dos registros TUSTEP-XML. Nótese la gran disparidad en el formato del contenido para campos iguales, como por ejemplo *QDB*, y la divergencia en el tipo y número de campos disponibles en cada uno. Por último el campo "orig" delimita el texto original empleado para generar el registro.

guno de los dos subconjuntos poseía las características estructurales o de contenido necesarias para cumplir con los objetivos marcados al principio del proceso. En la siguiente sección volveremos sobre el tópico de los conjuntos de datos, y en particular sobre los métodos que se adoptaron para conseguir un formato de entrada válido que cumpliese los requisitos idóneos demandados por el sistema de visualización que se había proyectado al comienzo.

4. Métodos y Herramientas

En este apartado introducimos las técnicas y herramientas empleadas en el desarrollo de la investigación, así como las metodologías de software que se usaron para la construcción del prototipo propuesto. Para ello, los clasificamos en las siguientes categorías, dependiendo del rol que han tenido dentro de la concepción del prototipo y la investigación en general: Estándares y Paradigmas, Evaluación y Minería de Textos, Indexación de los Datos y Motor de Búsquedas, e Interacción y Visualización de datos. En la sección 5 nos referiremos a las particularidades de implementación de cada caso.

4.1. Estándares y Paradigmas

4.1.1. Modelo de desarrollo del prototipo

El desarrollo del prototipo propuesto siguió un modelo iterativo, guiado por expertos y centrado en el usuario propuesto por Bernard et al.(2015)[44] en su trabajo de investigación de HD. En las primeras fases del modelo se crean pequeños prototipos que prueban cierto tipo de funcionalidad y sirven como medio para comprender los datos. Estos prototipos se muestran a los expertos en reuniones que ayudan a refinar y orientar el trabajo en las subsecuentes etapas. Es en éstas que se emplea el conocimiento adquirido por el investigador para crear nuevos prototipos o modificar los ya existentes. En el curso de nuestra investigación se crean 6 prototipos iniciales que resultan en el prototipo final. (Figura 14)

Aplicando este paradigma de desarrollo, se consigue poner de manifiesto en la investigación uno de los principios de las HD propuesto en las secciones iniciales: La **colaboración**, indispensable y clave en el desarrollo de nuestro estudio. En la Figura 15 podemos observar un esquema del modelo propuesto por Bernard.

4.1.2. Orientado a la web y escalable

Teniendo dbo@ema como referencia de funcionalidad básica, era necesario que nuestro prototipo propuesto tuviese un marcado carácter de orientación hacia la web y es así que este principio se tuvo en cuenta desde las primeras etapas de concepción de éste. En el proceso de elección de tecnologías y herramientas se favoreció la elección de aquéllas que respetasen los estándares marcados por la WWC y que su despliegue y mantenimiento *online* fuesen naturalmente fáciles de llevar a cabo. Dado que según avanzó la investigación se fueron manejando diferentes conjuntos de datos, se necesitaba una arquitectura web que se adaptase a todos los escenarios y que las diferentes versiones de los prototipos soportasen un conjunto de datos variable en tamaño. En esta tarea también resultó de especial ayuda ElasticSearch, ya que la carga de procesamiento se desplaza hacia el servidor, que es capaz de ejecutar búsquedas mucho más rápidas, devolviendo sólo los resultados relevantes

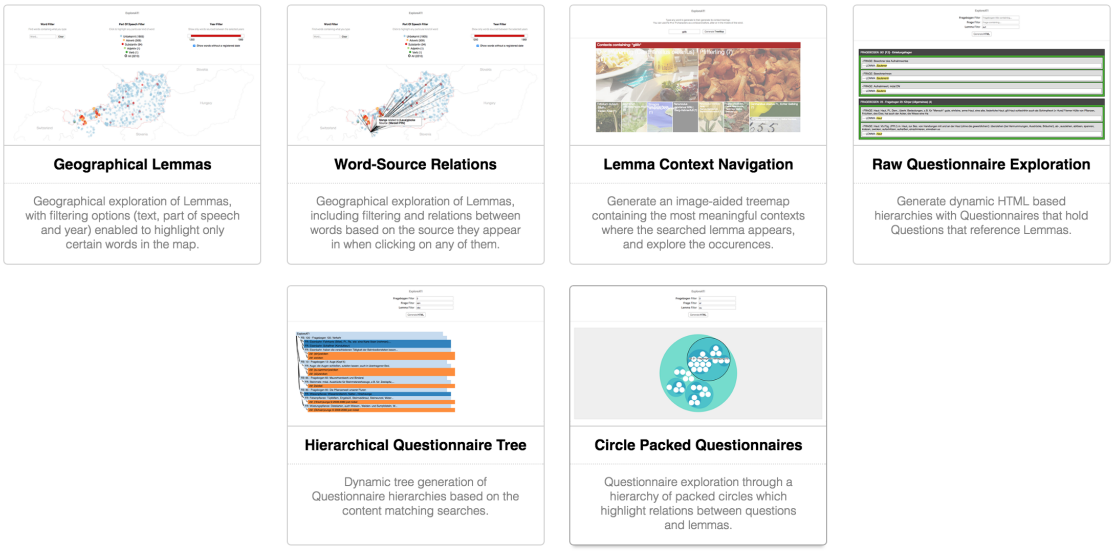


Figura 14: Vista general de los 6 prototipos creados en el curso de la investigación. 2 de ellos se centran en la distribución espacial de los datos en dbo@ema, basándose en trabajos previos[1]. Del resto, 3 hacen hincapié en la visualización de cuestionarios de preguntas a través de diferentes técnicas y 1 presenta una navegación contextualizada de los lemas.

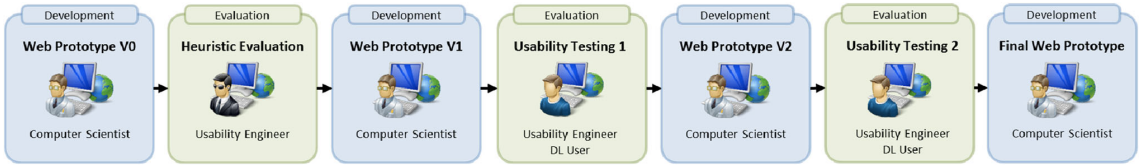


Figura 15: Proceso iterativo de desarrollo propuesto por Bernard et al. En cada fase intervienen diferentes *stakeholders* y se producen nuevos prototipos o versiones mejoradas de los existentes.

que sean necesarios para la etapa específica del análisis visual en la que se encuentre la investigadora. Como veremos, esta elección tiene un profundo impacto en términos de usabilidad de la aplicación y en la experiencia de usuario final, permitiendo adoptar soluciones de interacción ejecutables prácticamente en **tiempo real**.

4.1.3. JSON: El pegamento de Internet

El formato de intercambio de datos ligero JSON se ha impuesto como estándar de facto en muchos servicios de la Internet moderna. La claridad y sencillez de JSON han hecho que éste se impusiese a otras alternativas como XML/HTTP SOAP, que se han visto relegadas a un segundo plano recientemente. Sin embargo, a juicio del autor, la característica fundamental que ha conseguido desmarcar a JSON del resto de alternativas existentes en el pasado ha sido una: Su **gran capacidad de adaptación** a la realidad del mundo de Internet, que no es sino un reflejo del mundo real. Los naturaleza de los datos que se manejan en el día a día de la gran mayoría de

negocios, instituciones e individuos participantes de la red es cambiante y altamente proclive a cambios inesperados, en un flujo e intercambio continuo de información que se va incrementando en volumen y acelerando progresivamente con el paso del tiempo. Los problemas de estandarización en TUSTEP que reseñamos en la sección dedicada a los Conjuntos de Datos 3.3, unido al declive del uso de XML en favor de JSON en la mayoría de *frameworks* web fueron también determinantes a la hora de reemplazar XML por JSON en nuestro prototipo.

Se da sin duda la tendencia de que los sistemas menos rígidos y que imponen menos restricciones al modo en el que se vayan a manejar, como hemos visto que es el caso de los sistemas de búsqueda lexicográfica, sobreviven mejor a los tiempos, y en definitiva producen mejores resultados en la mayoría de los casos de uso a los que son sometidos. JSON no es un formato especialmente potente en términos computacionales en relación a sus competidores originales, como demuestra la misma existencia de la extensión JSON-RPC, pero es importante notar que lo que se supone que había de resolver lo hace extremadamente bien. Si bien es cierto que se podría argumentar al contrario, aduciendo que un exceso de laxitud en los formatos de datos va a dar como resultado un bajo rendimiento de los procesos de computación, o un decremento en la seguridad e integridad, **la búsqueda del equilibrio en el uso de elementos contrapuestos supone con casi toda seguridad la virtud**, y es con JSON que se ha conseguido.

4.1.4. NOSQL

Con JSON como ganador en la carrera de los formatos de datos de Internet, y de la mano de motores de búsqueda de texto como Lucene, han hecho posible la emergencia de alternativas a los SGBD relacionales SQL. Existe una gran dificultad para muchos de los profesionales del sector a la hora de elegir (o incluso distinguir) entre SGBD NOSQL, sistemas de almacenamiento de documentos y motores de búsqueda.

Sea como fuere, la verdad es que curiosamente estos sistemas experimentan su auge a finales de la década del siglo pasado, fecha que coincide recordemos con la primera versión de Lucene. Estas bases de datos, como se ha comentado anteriormente, rompen de pleno con el enfoque clásico relacional de los datos de los años 70 y 80. Además, y como consecuencia no exponen una interfaz SQL (Structured Query Language), sino que ofrecen otro tipo de alternativas para la recuperación de los datos, comúnmente en la forma de búsquedas de texto. Estos SGBD son tipificados normalmente como horizontalmente escalables, es decir, escalan mediante la adición de más nodos al entorno, en contraposición a la escalabilidad en vertical, que supone la adición de más recursos a un sólo nodo o conjunto de nodos. Es por esta razón que se adaptan mejor a las necesidades y topología de la red actual, motivada por el aumento de velocidad en las comunicaciones. Como se puede apreciar en la Figura 16, gracias a este diseño orientado a la escalabilidad, este tipo de BBDD son capaces de manejar volúmenes de datos notablemente mayores que sus parientes relacionales, sin experimentar un descenso en el rendimiento. Estas BBDD, de manera análoga a lo que ocurría con el formato JSON, han sabido adaptarse mejor

a las exigencias del denominado *Big Data* que otras alternativas. A pesar de esto, merece la pena reseñar sin embargo que a fecha de escribir estas líneas, los sistemas relacionales siguen siendo los más usados a nivel mundial.

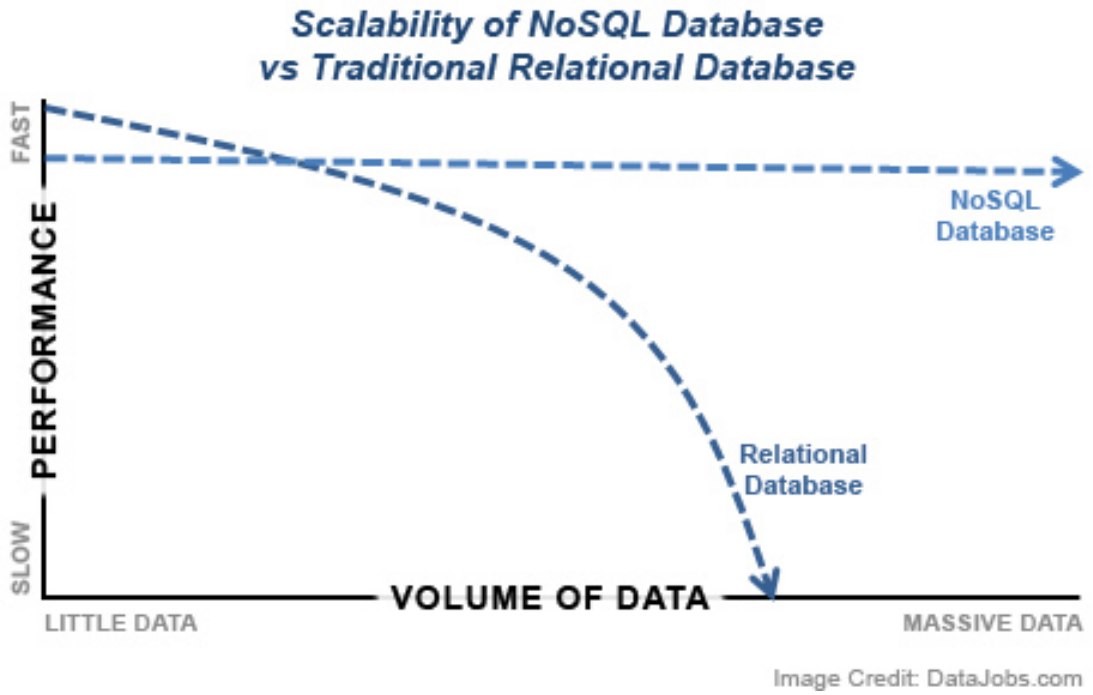


Figura 16: Comparativa de rendimiento entre BBDD NOSQL y Relacionales en base al volumen de datos manejado. En nuestro caso particular de estudio, el gran número de entidades diferentes (2 millones) y el formato libre de los mismos hacen que la opción NOSQL sea la más adecuada.

Existen multitud de proyectos de SGBD NoSQL que potencialmente podrían haber sido objeto de estudio, No obstante después de un primer proceso de filtrado, basándonos en el último estudio de popularidad, reseñamos aquí los 3 sistemas más usados a fecha de escribir este texto: MongoDB, Redis y Cassandra. [45]. Dentro del espectro, cada una va a venir a cubrir un tipo de necesidades diferentes, pero siempre manteniendo una marcada inclinación hacia la posibilidad de ser distribuidas. Mientras que MongoDB se caracteriza por mantener algunas propiedades de SQL como los *queries* o los índices, o permitir el guardado de ficheros grandes y el manejo de conjuntos de datos muy variables en tamaño, Redis se va a centrar en realizar rápidamente operaciones de actualización de un conjunto de datos limitado en tamaño. Por su parte, Cassandra, también un proyecto que en años recientes también ha pasado a formar parte de la fundación Apache, va a permitir el guardado de enormes conjuntos de datos manteniendo una interfaz muy parecida a SQL, llamada CQL3, salvo por aquellos elementos imposibles de realizar en ambientes distribuidos: Operaciones JOIN y sin funciones de agregación. A pesar que Cassandra se postuló en un principio como un primer candidato para reemplazar a TXSTEP/XQuery fue descartado rápidamente al no disponer de estas funciones de agregación, que son clave para lograr una correcta aplicación del mantra de la visualización[36].

Estas BBDD de datos son mencionadas en los sitios de Internet junto a Elastic-

Search, poniendo de manifiesto las mínimas diferencias existentes en la actualidad entre los motores de búsqueda de texto y los sistemas de BBDD, si bien es cierto existen muchos ejemplos de uso conjunto en los que, por ejemplo, Cassandra se emplea para conseguir una distribución adecuada y transparente de los datos, mientras que en la misma configuración, ElasticSearch se va a emplear para indexar y permitir la búsqueda de los datos. Muchas de estas bases de datos NoSQL presentan importantes deficiencias en la búsqueda de textos, aunque bien es cierto que todas ellas implementan algún tipo más o menos rudimentario de la misma y es por ello que todas estas opciones fueron descartadas en favor de ElasticSearch en nuestra investigación.

4.2. Evaluación y Minería de Textos

Debido al bajo grado de estandarización y alta heterogeneidad de los conjuntos de datos empleados para el estudio, fue necesario realizar una serie de evaluaciones heurísticas previas que ayudasen a construir un mapa mental previo antes de la construcción del prototipo. Para ello se emplearon una serie de herramientas y *scripts* que extrajesen variables cuantitativas globales de los datos. En un principio se emplearon los *scripts* de *bash grep*, *count*, *sort* y *order*.

4.3. Indexación de los Datos y Motor de Búsquedas

Como ya anticipamos en las secciones 3.3 y 4.1.4 cuando hablábamos de las limitaciones que imponían los SGBD relacionales en las tareas de búsqueda y almacenamiento de documentos textuales, surgió la necesidad de encontrar un motor de búsqueda textual que imitase la funcionalidad de XQuery pero que también soportase la interacción con elementos GIS y además estuviese preparado para funcionar correctamente en un entorno web. Esta elección resultó ser clave para la viabilidad de la investigación y por ello se derivaron notables esfuerzos en el proceso de documentación y búsqueda de alternativas viables cuyos resultados recogemos en esta amplia sección dedicada al motor de búsqueda documental ElasticSearch (ES en adelante). A pesar de su corto tiempo de vida, ES ha experimentado un gran éxito entre el público debido a sus altas prestaciones y facilidad de configuración. Este corto tiempo de vida no obstante, hace que la literatura sobre su uso y aplicación en el ámbito de las HD sea muy escasa. Sin embargo, en el proceso de investigación se encontró una referencia sobre un caso particular de aplicación exitosa de ES en HD (Hauswedell et al. 2015) [46], hecho que sin duda animó al autor a continuar su investigación por este camino.

Se comienza citando la historia y motivaciones que dan lugar a la concepción de este motor de búsqueda, y los indicios que se fueron encontrando sobre la idoneidad del uso del mismo en el ámbito de esta investigación. Se continua con una descripción de las funcionalidades básicas del software y se apuntan las técnicas empleadas en relación al mismo para conseguir el rendimiento y funcionalidad esperados en un principio.

4.3.1. Al principio fue Lucene

La búsqueda documental asistida por máquinas es una práctica relativamente nueva en la historia de la humanidad, y emergió en un primer momento con la aparición de los primeros ordenadores, aunque su diferenciación con respecto a otras ramas es bastante reciente. En un principio la recuperación de la información se englobaba dentro de la disciplina del almacenamiento y las bases de datos pero al aumentar la complejidad y la capacidad de procesamiento de los sistemas informáticos existentes, y tal como ha ocurrido con otras ramas de las ciencias de la computación, se hizo necesario distinguirla de otras prácticas.

Es por esto que hay que remontarse a las bases de datos relacionales y a su auge en los años 80 para encontrar los orígenes de la disciplina que nos atañe, y de los sistemas de software representativos de la misma que han sido creados a lo largo del curso de los años. Es así que los sistemas de bases de datos SQL se empleaban para almacenar documentos de forma estructurada, siendo ya en aquellos daños el volumen a manejar muy grande, del orden de los cientos de gigabytes. Como ya mencionamos en secciones anteriores, estos sistemas demostraron ser altamente ineficientes para la búsqueda documental, principalmente debido a la **rigidez** en los formatos de los datos que aquéllos imponían, al **alto grado de acoplamiento** existente entre ellos y el sistema de ficheros y por último a la **imposibilidad** de permitir **búsquedas no estructuradas** de documentos en tiempos aceptables.[47]

Es así que la biblioteca Lucene surge de la rama de los SGBD, en principio alejada de los sistemas de búsqueda textual propios de las HD, como TUSTEP. No obstante, la funcionalidad ofrecida por aquella rama de sistemas sí va a ir solapándose progresivamente con la de estos últimos hasta nuestro días.

En el último año del S.XX se produce uno de los mayores hitos en la historia de la búsqueda documental y los sistemas de información: Doug Cuttin escribe la primera versión de Lucene. Lucene fue un intento de flexibilizar las limitantes condiciones impuestas por los SGBD tradicionales, Se introdujo en este software la noción de **documento**, que es el concepto básico que se maneja en cualquier sistema de búsqueda textual en la actualidad, definiéndose como la entidad que contiene campos de texto en formato libre, la unidad de información atómica. Esta aparición abrió una puerta a un mundo completamente nuevo y diferente de las búsquedas en SGBD tradicionales, por lo que se puede considerar como la fecha de nacimiento de la búsqueda documental moderna. Nótese que este alejamiento de los SGBD tradicionales que caracteriza a Lucene (motor de ES) encaja perfectamente con el problema planteado en el dominio de nuestra investigación.

Apache Lucene fue originalmente concebido como una librería para Java, que es precisamente el lenguaje en el que se escribe ElasticSearch, lo cual dota a estos motores de capacidad de ejecutarse en diferentes sistemas operativos, siendo esta capa abstraida en la propia implementación del software, maximizando así la portabilidad de los datos, hecho de especial importancia hoy en día, y la intercomunicabilidad entre diferentes instancias de ejecución del software. El auge de Internet no hizo más que aumentar y promocionar el éxito de Lucene, que se postuló como una excelen-

te alternativa a la búsqueda basada en texto. Lucene fue rápidamente incorporado a la pila de servidores Jakarta de la Apache Software Foundation en el año 2001, conviertiéndose en un proyecto de primer nivel dentro de la misma en el año 2005, debido a su gran popularidad.

Es así que se pasó de pensar en optimizaciones de *queries*, formas normales y lectura de ficheros a otro dominio, en el que otra clase de problemas eran planteados, como la ordenación de resultados de búsqueda, o cómo lograr una indexación eficiente de los documentos de texto basándose en la mejor combinación posible, que redujese al máximo el almacenamiento ocupado por la representación de los mismos.

Lucene introdujo además el concepto de **índice invertido**, que va a asociar palabras a unidades de información (los documentos). Este simple pero inteligente cambio en el paradigma fue una de las novedades más revolucionarias e identitarias de Lucene. Además, establece precedente su sintaxis de búsqueda de cadenas basada en expresiones regulares, muy parecida a la ofrecida por Xquery.

4.3.2. Lucene es sólo una biblioteca: Solr

A pesar de todas las bondades mencionadas en la última sección acerca de Lucene, existen áreas que se escapan del ámbito y enfoque original del proyecto que han sido complementadas progresivamente a lo largo de los tiempos desde la aparición del software. Al ser un proyecto enmarcado en el contexto de la recuperación de la información, y como consecuencia de haber querido desmarcarse desde un principio de los SGBD tradicionales, iba a presentar carencias precisamente en este aspecto. El esfuerzo que se hizo para concebir Lucene fue dirigido hacia la descripción de los datos, no a la persistencia de los mismos. Esto motivó la aparición de sistemas encargados de llenar este vacío, como es el caso de Solr.

Solr se crea en los laboratorios de CNET en el año 2004, como un proyecto interno dirigido a soportar búsquedas en los sitios administrados por la popular compañía de medios americana. En 2006, su código fuente es liberado y se integra también en la familia Apache, junto con Lucene. Mientras que Lucene es una librería Java, Solr es una aplicación completa que hace uso extensivo de la misma y además añade diversa funcionalidad para dar una solución de software compacta y completa a la problemática de la búsqueda documental (Lo que se denomina en jerga técnica un *wrapper*). Solr por tanto, se persona como una **capa de abstración sobre Lucene**, fácilmente instalable en los entornos de la época y que no requería apenas de ningún conocimiento de programación para obtener un sistema de búsquedas configurable para los intereses del público general. Esta facilidad de instalación y uso, contribuyó como es natural muy notablemente a la expansión de los motores de búsqueda en texto en aquellos años, cuyo efecto es notorio aún en nuestros días.

De especial interés para nuestra investigación es una de las características aportadas por Solr, el uso de **facets** en las búsquedas, que marcó también un importante antes y después en los sistemas de búsqueda documental y que derivarían en el futuro en **las búsquedas agregadas**, usadas profusamente en la implementación de nuestro prototipo.

4.3.3. ElasticSearch: "You know, for search"

ElasticSearch ha sido sin duda alguna la gran revolución en los mótores de búsqueda de texto modernos, a raíz de un proyecto previo de su creador, llamado Compass. Compass fue creado por Shay Banon en el año 2004 con el objetivo de permitir la búsqueda lexicográfica en cualquier aplicación Java de la manera más sencilla posible.[48] Compass, que se encuentra descontinuado en estos momentos, al haber sido sustituido completamente por ES, marcó muchas de las líneas de desarrollo que definen a ES hoy en día, y ya incorporaba muchas que eran pioneras en el campo de las búsquedas textuales. La inclusión de OSEM (Object Search Engine Mapper), que permitía mapear clases del modelo a documentos del motor de búsquedas, el soporte XML y JSON, junto con su orientación hacia la escalabilidad son ejemplos de ellas, todas presentes en versiones recientes de ES. ES es un proyecto de código abierto bajo la licencia Apache 2, lo cual le colocaba como un perfecto candidato para su uso en nuestra investigación.

ElasticSearch, de manera análoga a *Solr*, va a abstraer el uso de la potente pero compleja librería Lucene. En palabras de sus creadores, es mucho más que simplemente Lucene, y mucho más que sólo búsqueda de textos, así que se describe como:

- Un **almacenamiento distribuido** en tiempo real de documentos, donde cada *campo* puede ser indexado y buscado.
- Un motor distribuido de búsquedas con analíticas en **tiempo real**.
- Capaz de **escalar** hasta **cientos de servidores y petabytes** de datos estructurados y no estructurados.

A pesar de que estas características podrían describir a un sistema altamente complejo, ElasticSearch se va a orientar hacia la facilidad de uso y va a permitir que investigadores noveles en el campo de la recuperación de la información y la búsqueda textual sean capaces de ser productivas mediante el uso de esta herramienta [49].

4.3.3.1. Distribuibilidad

Es así que la primera versión de ElasticSearch, lanzada en 2010, se creó como una gran reescritura de Compass (lo que habría sido Compass 3.0), a partir de los importantes cambios que se incorporaron en Lucene 2.9, que la hicieron necesaria. Aprovechando esta reescritura, el autor inteligentemente dotó a esta nueva versión del software de capacidad para hacerse distribuido desde su misma concepción. Podríamos decir por tanto que ElasticSearch **nació para ser distribuido**, y para ello hubo de sacrificar importantes perstaciones como el concepto de **transacción**. Esto dio como resultado la introducción de un nuevo concepto denominado *shards* (literalmente traducido, esquirla), que se define como una instancia de Lucene manejada por ES. Cada documento va a tener un **shard primario y uno de réplica**

que van a ser definidos por la usuaria. El objetivo de este *shard* de réplica va a tener dos objetivos principales [50]:

1. Fomentar confiabilidad y la disponibilidad de los documentos en caso de caída del shard primario
2. Incrementar el rendimiento de las búsquedas, que podrán ser manejadas por el shard primario o sus replicas basándose en aspectos de topología de la red de distribución, como la latencia u otras.

4.3.3.2. Interoperabilidad

Como no podría ser menos, y por las razones previamente explicadas, **JSON** fue el formato para permitir la comunicación con otros lenguajes de programación y plataformas, y se encuentra embebido en la plataforma en todas sus vertientes. **Toda representación de un documento va a venir dada en este formato**. ElasticSearch además expone una interfaz HTTPS, que va permitir realizar las consultas a través de este protocolo, permitiendo la ejecución de consultas incluso a través un navegador web o mediante el empleo de la herramienta *curl*. Además, a raíz de la utilización de JSON como formato de preferencia en el sistema. Se crea también un lenguaje específico de búsqueda llamado **Query DSL**, basado en JSON. Este lenguaje va a permitir realizar búsquedas de texto empleando una sintaxis JSON, que introducimos en la siguiente sección.

Por añadidura, ES, hereda todas las características importantes enumeradas en secciones anteriores referidas a la recuperación de información, como son

4.3.3.3. Estructura

Ya hemos reseñado en los parráfos anteriores la topología de una red ES, explicando que los documentos pueden ser distribuídos en diferentes shards. Pero existe otro tipo de clasificación dentro del shard, el índice, que va a acoger grupos de documentos listados en base a alguna de sus propiedades. El API de búsquedas va a permitir realizar búsquedas a un grupo de índices o a todos los que existan distribuidos en los diferentes shards.

4.3.3.4. Mapeado

En el proceso de mapeado se define cómo ElasticSearch va a tratar los campos de un documento, y la manera en la que van a ser almacenados e indexados. No tendría sentido por ejemplo tratar un campo que contiene coordenadas geográficas como numérico o textual (a pesar de contener éste números), ya que ese tipo de búsquedes no van a ser capaces de extraer ningún tipo de información valiosa de los mismos. Se van a determinar en este proceso por tanto, qué campos de cadenas deberían ser tratados como texto, o cuáles de ellos contienen números, fechas o coordenadas

geográficas, etc. Esta importante decisión va a generar un tipo de análisis para cada campo, que permitirá en turno efectuar un tipo u otro de búsquedas sobre ellos.

ES añade a todos sus documentos unos campos descriptivos o meta-campos internos, como el índice (_index), el identificador (_id) o el documento original (_source), que van a ser generados en la inserción de los mismos. Se definen además en este proceso los tipos de datos de cada campo (cadena, fecha, booleano...) con la imposición de que todos los campos con el mismo nombre en el mismo índice han de ser mapeados siempre de la misma manera. Estos metacampos presentan además una clara correlación con el esquema propuesto en TUSTEP-XML (Figura 13).

4.3.3.5. Búsquedas

Se permiten en ES dos tipos de búsquedas: **lite y complex**. Mientras que las búsquedas lite o reducidas se basan únicamente en búsqueda de cadenas, el verdadero poder de ES reside en su API de búsquedas complejas, que permite queries más complejos por medio del empleo del lenguaje DSL. Es así que el tipo de búsquedas permitido en ES va a poder ser: 1)Un query estructurado en campos concretos, ordenado por un campo del documento, similar a lo realizado en SQL, 2)Un query de texto completo, que va a encontrar todos los documentos que coincidan con las palabras clave, y va a producir un resultado, especificado en la clave del objeto JSON de respuesta *hits*. Dentro de este objeto encontraremos los documentos ordenados y filtrados por relevancia u otros campos de acuerdo a la naturaleza de la consulta, como se verá en los siguientes párrafos. El primer tipo de búsqueda que se permite es la búsqueda vacía ({}), la cual va a encontrar todos los documentos dentro de un conjunto de índices. Se pueden dar valores también para la *paginación*, o la cantidad de resultados que se desean obtener dentro de los que encajen con la búsqueda.

Query String: La búsqueda simple de cadenas va a ser ilustrada en profundidad en el caso de uso práctico de la siguiente sección de este documento 5, así que de momento nos limitamos a enumerar en este apartado sus características principales y operadores. Los query string son altamente potentes a la par que sencillos de utilizar. Éstos simplemente van a buscar la aparición de un aserto lógico de cadenas en una serie de campos de los documentos. Existe para tal efecto el modificador de campo *_all*, con el que la búsqueda se va a ejecutar en todos los campos del documento en cuestión, que es el que se va a lanzar por defecto cuando no se determina ningún campo específico. Query string va a comprender por tanto en sí mismo un lenguaje mínimo, basado en el lenguaje de expresiones regulares y como tal, nos va a permitir hacer búsquedas de proximidad entre palabras, rangos numéricos o lexicográficos, expansiones, expresiones booleanas (must o must not), etc. Remitimos al lector a de nuevo al caso práctico para comprender en más profundidad las capacidades de este sub-lenguaje.

Query DSL: El lenguaje DSL es flexible y expresivo, y sirve a ES para extraer todo el poder que reside en las instancias de Lucene que componen la instalación.

Se recomienda encarecidamente emplear este lenguaje en entornos de producción, ya que es más fácil de depurar y de entender. La sintaxis de DSL es de árbol, siendo las consultas tratadas primero por un procesador de consultas, de manera parecida lo que se hace en BBDD SQL. DSL soporta dos tipos de sentencias:

Cláusulas hoja - Leaf query clauses Van a buscar un valor en un campo dado en dicho el query. En este grupo se engloban los queries **match, term y range**. Pueden usarse individualmente y no necesitan de otras sentencias para producir resultados válidos.

Cláusulas compuestas - Compund query clauses Combinan múltiples cláusulas hoja de manera lógica, empleando operadores lógicos como AND, OR o NOT.

La consulta más básica que se puede hacer es el "Match all", que va a encontrar todos los documentos dentro de un índice, asígnándoles a todos una puntuación de 1. En las consultas de texto completo, se va a efectuar la búsqueda en base al análisis efectuado de los documentos, comunicándose con este con el objetivo de obtener los mejores resultados posibles. Estos analizadores se van a componer de *tokenizadores*, que van a extraer las palabras clave de los documentos analizados. Además, también van a contar con *filtros de tokens*, que van a recibir la salida de los tokenizadores y van a aplicar un primer procesamiento de los textos (por ejemplo, convertir a minúsculas). A su vez, los tokenizadores pueden recibir la entrada directamente del texto a analizar o de los filtros de caracteres, que van a eliminar caracteres extraños y signos de puntuación. Además de "Match all", contabilizamos en el siguiente listado algunos de los tipos de búsquedas DSL de texto más importantes:

- Match: Acepta búsquedas de texto, números y fechas. Existen tres tipos
  - boolean: Tipo por defecto. Va a producir valores booleanos para ser combinados con otras cláusulas. Es equivalente a una cláusula SELECT en lenguaje SQL.
  - phrase: Va a hacer que los documentos devueltos en la respuesta:
    - Contengan **todos los términos** que se indican en la cláusula
    - Además, lo hagan en el **mismo orden** que el que se indica en ella
  - match_phrase_prefix: Funciona de igual manera al tipo phrase, excepto que se va a usar la frase entera como prefijo a encontrar en el texto.
  - query string: Se va a permitir introducir query strings como los explicados en la sección anterior en consultas más complejas.
- multi-match: Se apoya en los *match* queries para construir búsquedas en base a varios campos. Permitir especificar prioridades para los diferentes resultados encontrados en cada campo.
- términos comunes: Es una alternativa moderna a las palabras de parada de cada lenguaje, sin reducir la precisión de las búsquedas: Esto es, las palabras de parada no se eliminan completamente de la búsqueda, sino que dividen a los términos de la búsqueda en más importantes (los que aparecen con poca

frecuencia y serán más relevantes para la búsqueda) y menos importantes (los que lo hacen con más frecuencia y por tanto contribuirán menos al resultado final).

Aparte de los *queries* de texto, se permiten otros muchos tipos de consultas: a nivel de **términos**, que operan sobre los términos exactos almacenados en el índice invertido. Otro tipo digno de mención son los **Join** queries, que permiten obtener una funcionalidad parecida a la ofrecida por las BBDD relacionales. Principalmente ésto se consigue almacenando objetos completos en campos especiales del documento, que van a poder ser consultados empleando el tipo de query *nested*. También se habilitan las búsquedas taxonómicas de documentos, que retornan resultados de búsqueda en relaciones padre-hijo establecidas previamente entre ellos por el desarrollador en el mapeado de los datos. Otras búsquedas permitidas en ES, heredadas de sistemas como Solr, comprenden aquellas referentes a **localizaciones geográficas**, queries span, que tienen en cuenta la **distancia entre palabras** y queries especializados, que permiten búsquedas complejas empleando scripts, búsquedas de similaritud y búsquedas de código HTML. Como veremos en el desarrollo de la solución en la siguiente sección,

4.3.3.6. Agregación de búsquedas

Una mención aparte merece esta funcionalidad ofrecida por ElasticSearch, que ayuda a proveer datos agregados en base a los términos de búsqueda proporcionados. Se basa en bloques simples denominados agregaciones, que sirven para componer resúmenes complejos de los datos. Existen muchos tipos de agregaciones (geógraficas, de términos, de muestreo...), pero todas ellas se agrupan en 3 categorías fundamentales:

1. **Buckets:** Esta familia construye *buckets* (literalmente del inglés balde o cubeta), que son estructuras de datos en las que se recogen los elementos que, tras ser evaluados, cumplen cierto criterio de búsqueda previamente fijado. Cuando esto ocurre, se considera que los elementos "caen" en el *bucket* correspondiente. Al final del proceso de agregación, se obtiene una lista de *buckets*, cada uno con un conjunto de documentos que "pertenecen" a él.

2. **Métricas:** Este tipo de agregaciones mantienen y computan una serie de métricas estadísticas sobre un conjunto de documentos como la media, varianza o valores máximos y mínimos. En el caso de la visualización de datos, son especialmente útiles para crear escalas.

3. **Pipeline:** Estas agregaciones agregan resultados provenientes de otras agregaciones. En nuestro caso, se emplean para el análisis visual de redes, ya que permiten generar rápidamente las estructuras de datos necesarias para la creación de diagramas de grafos.

4.4. Interacción y Visualización de los datos

4.4.1. d3

Para la interfaz se emplea la conocida biblioteca de visualización D3. Se elige por la cantidad de ejemplos existentes online, amplia documentación y probada solvencia en la resolución de problemas comunes que se pueden encontrar a la hora de crear navegaciones interactivas en el navegador. No nos extenderemos mucho en este apartado ya que son muchísimas las referencias existentes a esta biblioteca, y sería inútil tratar de condensar aquí todas ellas.

4.4.2. d3-carto-map

Este plugin de d3 es la base de la aplicación Orbis mencionada en la sección de trabajos relacionados. Se escoge por facilitar en gran medida al programador el uso de mapas de tiles, así como ofrecer partes de funcionalidad que guardan relación con nuestro proyecto. Se crea un *fork* del proyecto sobre el que se van añadiendo mejoras o adaptaciones de la misma a las necesidades de la investigación.

4.4.3. Crossfilter

Crossfilter es una biblioteca javascript que ayuda al programador en la tarea de crear una estructura de vistas enlazadas. En nuestro caso se emplea profusamente para relacionar las tres vistas de la interfaz que presentaremos en las secciones siguientes, en concreto la funcionalidad de *highlighting* y filtrado dinámico de los datos.

5. Desarrollo de la solución

En este apartado discutimos la implementación de las soluciones a los problemas encontrados en el proceso de crear el prototipo de análisis visual exploratorio multidimensional propuesto.

5.1. Adquisición de los datos: Enfoque híbrido

Como se apunta en anteriores secciones, fue necesario un procesamiento previo de los datos que extrajese características de los textos para su posterior análisis visual. Recordemos que en un principio se disponía de dos conjuntos de datos en diferentes formatos: MySQL y ficheros XML. Ya que la naturaleza de los mismos hacía imposible su tratamiento y aplicación directa de técnicas de análisis visual, se generaron una serie de scripts en lenguaje *Javascript* que aplicarían este primer tratamiento. Estos scripts habrían de combinar información proveniente de ambos conjuntos de datos para crear un tercero que contuviese la información esperada. El objetivo pues fue conjugar e indexar toda la información disponible en una instancia del motor de búsqueda ElasticSearch, añadiendo las dimensiones textual, espacial y temporal cuando fuese posible, que permitiesen plantear un análisis visual multidimensional de la información. En esta sección explicamos en qué consistió este enfoque híbrido y qué técnicas de minería de datos se aplicaron a los textos para lograr los resultados esperados.

En nuestro enfoque se emplea el conjunto de datos XML como fuente primaria de información, mientras que el conjunto proveniente de dbo@ema sirve de conjunto de datos de soporte al primero. El primer tipo de información que se deseaba incluir era la espacial ya que es en ésta que se centra el modelo de herramienta de análisis propuesto. Recordemos que el formato XML contenía también información espacial pero la misma no estaba geocodificada, ya que sólo se disponía del nombre del lugar al que hacía referencia la fuente. Por otro lado, la base de datos usada en dbo@ema sí que contenía tablas en las que se relaciona el nombre de un lugar tal como se puede encontrar en las fuentes XML de TUSTEP y su información GIS asociada. Por otro lado, la mayoría de la información se encontraba aún en formato XML. Era pues preciso combinar estas dos fuentes de información, extrayendo la valiosa información GIS de la base de datos MySQL e insertarla dinámicamente en los nuevos documentos creados en ElasticSearch junto con la proveniente de los registros TUSTEP. Veremos a continuación con un ejemplo este proceso de adquisición de dimensiones para un registro escogido al azar de entre los datos. En la Figura 17 se muestra una captura de pantalla ilustrando el contenido del registro XML y el registro de la BBDD que contendría la información espacial asociada.

En ella podemos observar los diferentes campos que componen este registro. De especial interés es el campo “QDB”, que como vemos tiene información temporal de una fuente (en este caso un cuestionario, FbB) e información espacial contenida en un subcampo “O” emplazado a tal efecto. Vemos también que no existe una correlación directa entre el nombre del topónimo en XML y su representación en

1)
```
<record n="447">
   <field name="A">HK 869, z8690113.kro^#8</field>
   <field name="HL">(Hals)zapfen:1</field>
   <field name="QU">Blaindorf Stmk. Fabiani</field>
   <field name="QDB">{3.5g02} uFeistritzt.:mOSt. <part>FbB.FABIANI· (u.1913) [SFb.]</part>
      <field name="O">Blaindf. St.</field>
   </field>
   <field name="NR">4L7: Gaumen</field>
   <field name="LT1">H;olsz-abfrl [D2]</field>
   <orig>
      *A* HK 869, z8690113.kro^#8&#xD;
      *HL* (Hals)zapfen:1&#xD;
      *QU* Blaindorf Stmk. Fabiani&#xD;
      *QDB* {3.5g02} uFeistritzt.:mOSt. *^@ FbB.FABIANI· (u.1913) [SFb.] *O* Blaindf. St.&#xD;
      ===&#xD;
      *NR* 4L7: Gaumen&#xD;
      *LT1* H;olsz-abfrl [D2]&#xD;
      ===&#xD;
   </orig>
</record>
```

2)
```
id,nameKurz,nameLang,sort,bearbeitungsgebiet_id,gemeinde_id,gis_ort_id,namensvarErl,behoerde,quellen,ort_verzeichnis
_id,originaldaten,freigabe,checked,wordleiste,druck,online,publiziert,anmerkung,trust,menschkurz,OKZ,autokurz
16650,Blaindf.,Blaindorf,999999,2,995,15887,,,NULL,1,NULL,0,1,0,0,1,0,NULL,3,Blaindf.,15091,Blaindf.
```

Figura 17: 1: Registro 447 de un fichero XML de TUSTEP, referente al lema "Hals-zapfen". En el campo QDB podemos encontrar las dimensiones temporal y espacial asociadas a la fuente. 2: Representación CSV del registro asociado a Blaindorf en la BD MySQL. Obsérvese que la coincidencia de los nombres no es exacta.

MySQL, haciendo imposible la automatización del proceso.

5.1.1. Extracción de la dimensión espacial

Mediante la aplicación de técnicas heurísticas adquiridas mediante el estudio de los datos, se llegó a la conclusión de que la mejor opción era implementar una serie de reglas que codificasen el conocimiento experto adquirido. Cuando estas reglas no fuesen capaces de asociar correctamente las dos entradas provenientes de ambos conjuntos de datos, se requeriría de la intervención humana para finalizar el proceso, bien creando la asociación manualmente o bien descartando cualquier asociación, quedando el registro final sin información GIS asociada.

En un principio se contabilizan 1.861.878 registros (80 %) con algún tipo de referencia espacial asociada no estandarizada (la referencia al lugar no se encuentra dentro del registro "QDB". 322.459 registros (14 %) contienen dicho campo. Por otro lado, no siempre la información espacial hace referencia a un punto en el mapa: En otros casos se hace referencia a las divisiones administrativas de Austria: una comunidad o una región, dando lugar a diferentes resoluciones espaciales. En la Figura se muestra un ejemplo de la codificación de tales características, y su representación visual:

En una primera versión del prototipo se recogen únicamente los registros con resolución espacial de comunidad o polígono (830.489 / 35 %) o localidad o punto (115.248 / 4 %), que suponen al final del proceso de importación un total de 945.943

{ "type": "Polygon", "coordinates": [[[16.63201, 48.26002], [16.63161, 48.26051], [16.63147, 48.26073], [16.63141, 48.26092], [16.63144, 48.2611], [16.6313, 48.26142], [16.63212, 48.26321], [16.63196, 48.2636], [16.63114, 48.26555], [16.62995, 48.26618], [16.62987, 48.26626], [16.62978, 48.26637], [16.62906, 48.26681], [16.62834, 48.26809], [16.62814, 48.2682], [16.62796, 48.26829], [16.62761, 48.26838], [16.62725, 48.26843], [16.62683, 48.2684], [16.62647, 48.26833], [16.6261, 48.26821], [16.62596, 48.26821], [16.62506, 48.26803], [16.62323, 48.26629], [16.62087, 48.26576], [16.62068, 48.26583], [16.62023, 48.26607], [16.61992, 48.26626], [16.61858, 48.26693], [16.61839, 48.26705], [16.61824, 48.26719], [16.61765, 48.26797], [16.61731, 48.26839], [16.6168, 48.26905], [16.6163, 48.26969], [16.61625, 48.26983], [16.6163, 48.27017], [16.61635, 48.27074], [16.61634, 48.27101], [16.61629, 48.27125], [16.61621, 48.27148], [16.61604, 48.27173], [16.6158, 48.27199], [16.61532, 48.27237], [16.61518, 48.27244], [16.61493, 48.27263], [16.61476, 48.27278], [16.614, 48.27393], [16.61383, 48.27414], [16.6135, 48.27438], [16.61281, 48.27492], [16.61165, 48.27591], [16.61125, 48.27615], [16.61034, 48.27652], [16.60979, 48.27678], [16.60951, 48.27688], [16.60915, 48.27698], [16.60827, 48.27716], [16.60777, 48.27721], [16.60801, 48.27758], [16.60813, 48.27775], [16.60818, 48.27804], [16.60829, 48.27845], [16.60848, 48.27881], [16.60854, 48.27888], [16.60866, 48.27919], [16.60901, 48.28364], [16.60726, 48.28563], [16.60532, 48.28802], [16.60197, 48.29191], [16.60689, 48.29915], [16.60863, 48.30108], [16.60934, 48.30149], [16.61018, 48.30239], [16.61086, 48.30251], [16.61099, 48.30261], [16.61201, 48.30449], [16.61234, 48.30488], [16.61237, 48.30496], [16.61222, 48.30541], [16.61222, 48.30552], [16.61227, 48.30567], [16.61318, 48.30724], [16.61361, 48.30776], [16.61439, 48.30851], [16.61452, 48.30867], [16.6187, 48.30885], [16.6216, 48.30894], [16.62389, 48.30906], [16.62596, 48.3093], [16.62642, 48.30925], [16.62947, 48.30865], [16.63038, 48.30831], [16.63113, 48.30794], [16.63287, 48.30708], [16.63306, 48.3072], [16.6341, 48.3079], [16.63417, 48.30716], [16.63426, 48.30623], [16.63424, 48.30616], [16.63407, 48.30549], [16.63365, 48.30356], [16.647, 48.29953], [16.65108, 48.2976], [16.6556, 48.29487], [16.65682, 48.29539], [16.65839, 48.29373], [16.66053, 48.29159], [16.66125, 48.29069], [16.66281, 48.28873], [16.66324, 48.28787], [16.66646, 48.28007], [16.66665, 48.2796], [16.66669, 48.2796], [16.66674, 48.27914], [16.6679, 48.27728], [16.66795, 48.27725], [16.6684, 48.27626], [16.66843, 48.27619], [16.66893, 48.27555], [16.66965, 48.27439], [16.6698, 48.27426], [16.66992, 48.27423], [16.67035, 48.27345], [16.67185, 48.26903], [16.66955, 48.26589], [16.66784, 48.26317], [16.66738, 48.26254], [16.66632, 48.26112], [16.66381, 48.25871], [16.66225, 48.25696], [16.66175, 48.2551], [16.66088, 48.2543], [16.66015, 48.25382], [16.66003, 48.2537], [16.65922, 48.25362], [16.65703, 48.25442], [16.65531, 48.25332], [16.6536, 48.25258], [16.65196, 48.25222], [16.65153, 48.25222], [16.65074, 48.25224], [16.6504, 48.25229], [16.65012, 48.25238], [16.64997, 48.2525], [16.64988, 48.25265], [16.64991, 48.25284], [16.65001, 48.25316], [16.65053, 48.25401], [16.65077, 48.2543], [16.65097, 48.25461], [16.65108, 48.25486], [16.65113, 48.25509], [16.65116, 48.25568], [16.65123, 48.25602], [16.65132, 48.25623], [16.65164, 48.25671], [16.65191, 48.25721], [16.65201, 48.25752], [16.65205, 48.25779], [16.65205, 48.25817], [16.65197, 48.25868], [16.65186, 48.25911], [16.65161, 48.25923], [16.64964, 48.25984], [16.64878, 48.2599], [16.64836, 48.25977], [16.64794, 48.25968], [16.64653, 48.25947], [16.64591, 48.25941], [16.6456, 48.25942], [16.64519, 48.25949], [16.64429, 48.25974], [16.64331, 48.26016], [16.64298, 48.26028], [16.64254, 48.26034], [16.64212, 48.26037], [16.64152, 48.26031], [16.64103, 48.26021], [16.64059, 48.26005], [16.64018, 48.25981], [16.63947, 48.25925], [16.63915, 48.25903], [16.63875, 48.25884], [16.63816, 48.2586], [16.63753, 48.2584], [16.63729, 48.25835], [16.63705, 48.25834], [16.63657, 48.25841], [16.63633, 48.25845], [16.63569, 48.25859], [16.63496, 48.25874], [16.63436, 48.25889], [16.63256, 48.25954], [16.63236, 48.25967], [16.63201, 48.26002]]] }

Figura 18: Detalle de la representación de una comunidad. Ésta se define como un polígono que se proyecta en unas coordenadas. Arriba, el polígono representado visualmente. Abajo, el polígono en formato GEOJSON, que se inserta en el nuevo índice creado.

(40 %).

5.1.2. Extracción de la dimensión temporal

En el caso de la información temporal, se codificaron también reglas heurísticas para extraer dicha dimensión. Esta información puede venir en forma discreta o de intervalo cuando la fuente es por ejemplo un volumen que se extiende a lo largo de un período de tiempo, o cuando la fecha de origen exacta es desconocida y se proporciona una estimación. Para resumir, la información temporal va a venir representada en varios formatos, cada uno con una interpretación asociada diferente. Un 71 % de los registros presentan información temporal asociada.

- 1945 (Cuatro dígitos, rodeados por un número variable de caracteres): Ofrecen el máximo nivel de resolución temporal, el año. Se recuperan 509.929 registros, que suponen un 22 % del total.
- 1945-50 (Cuatro dígitos + guión + dos dígitos): Resolución temporal menos de un año pero mayor que década. Se extraen 155.240 (6 %) registros.
- 193x (3 dígitos + "x"): La resolución temporal es de década. Suponen un 2 % (54.374) de todos los registros.
- 19xx (Dos dígitos + "xx"): La resolución temporal es de siglo. En los registros sólo se encontraron ocurrencias de este tipo que hacen referencia al siglo XX. Un 41 % (954.126) de los registros presentaban este formato.

Al terminar el proceso de extracción de características que combina los dos conjuntos de datos, se arrojan los siguientes números sobre el conjunto final creado, que se considera preparado para someterse a un análisis visual, completándose así esta primera fase del estudio:

- Se importan correctamente 2.206.227 registros (95.3 %) de los 2.314.031 originales. El resto son descartados por errores en el formato original de los datos.
- De los registros importados correctamente, un 9.8 % contienen información referente a las dimensiones **espacial y temporal**.
- Un 26.6 % contiene información **temporal** pero **no espacial**.
- Un 32.4 % contiene información **espacial** pero **no temporal**.
- Por tanto, un 31.1 % de los datos **no contiene** ningún tipo de dimensión **temporal** o **espacial**.

En la siguiente sección se detallan las particularidades del prototipo alcanzado que permite la exploración del conjunto de datos creado en esta fase.

5.2. Prototipo propuesto

El prototipo resultante de la investigación es una herramienta de análisis visual multidimensional de la información recuperada y adaptada en la fase anterior de adquisición de datos. A pesar de que el enfoque es multidimensional, éste le da una mayor importancia inicialmente a la dimensión espacial, que en un análisis exploratorio en el que no dispongamos de ningún tipo de entrada por parte de la usuaria, va a ser la que sirva de guía al proceso. En nuestro enfoque, consideramos que el porcentaje de los datos generados en la anterior etapa de adquisición de los mismos es válido para guiar el flujo de trabajo por esta dimensión.

En la Figura 19 recogemos una captura de la interfaz de la aplicación con todas sus vistas desplegadas:

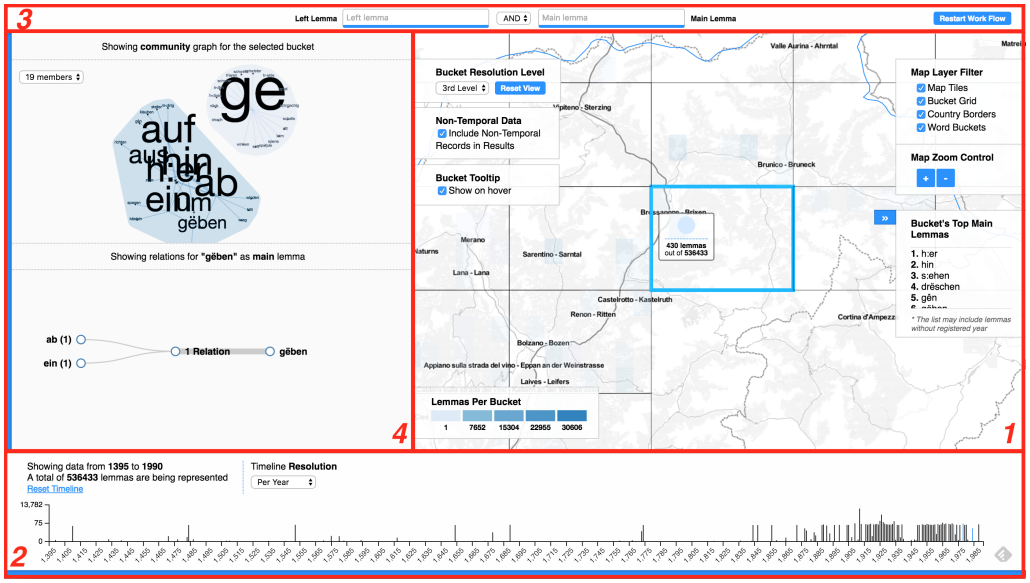


Figura 19: Interfaz del prototipo propuesto con 1) Proyección espacial o mapa, 2) Proyección temporal o *timeline*, 3) Barra de búsqueda textual, 4) Vista de análisis de redes

La interfaz muestra 4 vistas principales, de las que sólo 3 estarán disponibles en un principio (la vista 4 se muestra y oculta dinámicamente dependiendo de la fase del flujo de trabajo en la que se encuentre la analista en un momento dado). Describimos brevemente la funcionalidad de estas vistas:

1. **Vista de mapa:** En ella se presenta la dimensión espacial de los datos. En una primera versión del prototipo se agregan en ella las proyecciones geográficas de las fuentes que contienen información espacial a resolución máxima (puntos en el mapa). Comentaremos las razones de esta decisión en profundidad en la sección dedicada a esta vista. En ella realizaremos filtrado espacial mediante diversas técnicas de UX como zoom, desplazamientos y selecciones de elementos.

2. **Línea temporal:** Proyectamos en esta vista, que está enlazada a la primera,

todos los documentos que contienen información temporal asociada. En ella se realiza también el filtrado de elementos dinámico en base a dicha dimensión, que actualizará los elementos de la primera vista.

3. **Barra de búsqueda textual:** En ella se permite a la usuaria realizar una búsqueda dirigida de los elementos en base a su información textual. Se aplica con ella la técnicas de UX de búsqueda instantánea, también llamada "Search as you type".

4. **Vista de exploración de redes:** En esta parte de la interfaz se plantea el análisis de redes sociales (SNA) que va a ayudar a la analista a encontrar patrones estructurales en las relaciones entre los datos.

Como introdujimos en las secciones previas, este prototipo propone un flujo de trabajo basado en el mantra de la visualización, recordemos: "Visión general primero, zoom y filtrado, por último detalles en demanda". En las siguientes secciones, basándonos en este precepto, describiremos el funcionamiento general del prototipo y cómo se aplican las técnicas de visualización y UX a cada una de las partes del flujo de trabajo para conseguir un análisis visual del diccionario.

5.2.1. Visión general primero

Como ya se apuntó en el anterior apartado, la aplicación realiza su primera carga mostrando una vista general de los datos que sirve como punto de partida para un análisis visual exploratorio multidimensional y dirigido espacialmente. Es ya desde este punto donde nuestro enfoque trata de ser radicalmente diferente a lo ofrecido por otros trabajos de investigación y/o visualización de la información en HD: Es costumbre en el mundo de la visualización, y así se hace en los ejemplos y referencias mencionados en este trabajo, llevar a cabo la tarea de dar una visión global de los datos mediante 3 técnicas bien diferenciadas:

1. Se hace una carga inicial de todos los datos y se computan sobre ellos los algoritmos necesarios para crear las visualizaciones.

2. Los datos reciben un tratamiento previo que genera estructuras de datos que son cargadas junto a los datos en memoria. Dependiendo del nivel de detalle aplicado, se accede a dichas estructuras o los datos para generar las visualizaciones.

3. Se emplea una combinación de las dos anteriores: Dependiendo de la complejidad del algoritmo a aplicar se emplean estructuras de datos estáticas creadas a tal efecto o se generan dinámicamente.

Estas técnicas suelen ser suficiente cuando el volumen de los datos no es excesivamente grande, y se puede, aún en un entorno monohilado como el de la Web, aplicar alguna de las técnicas en tiempo de ejecución. En nuestro caso no es posible,

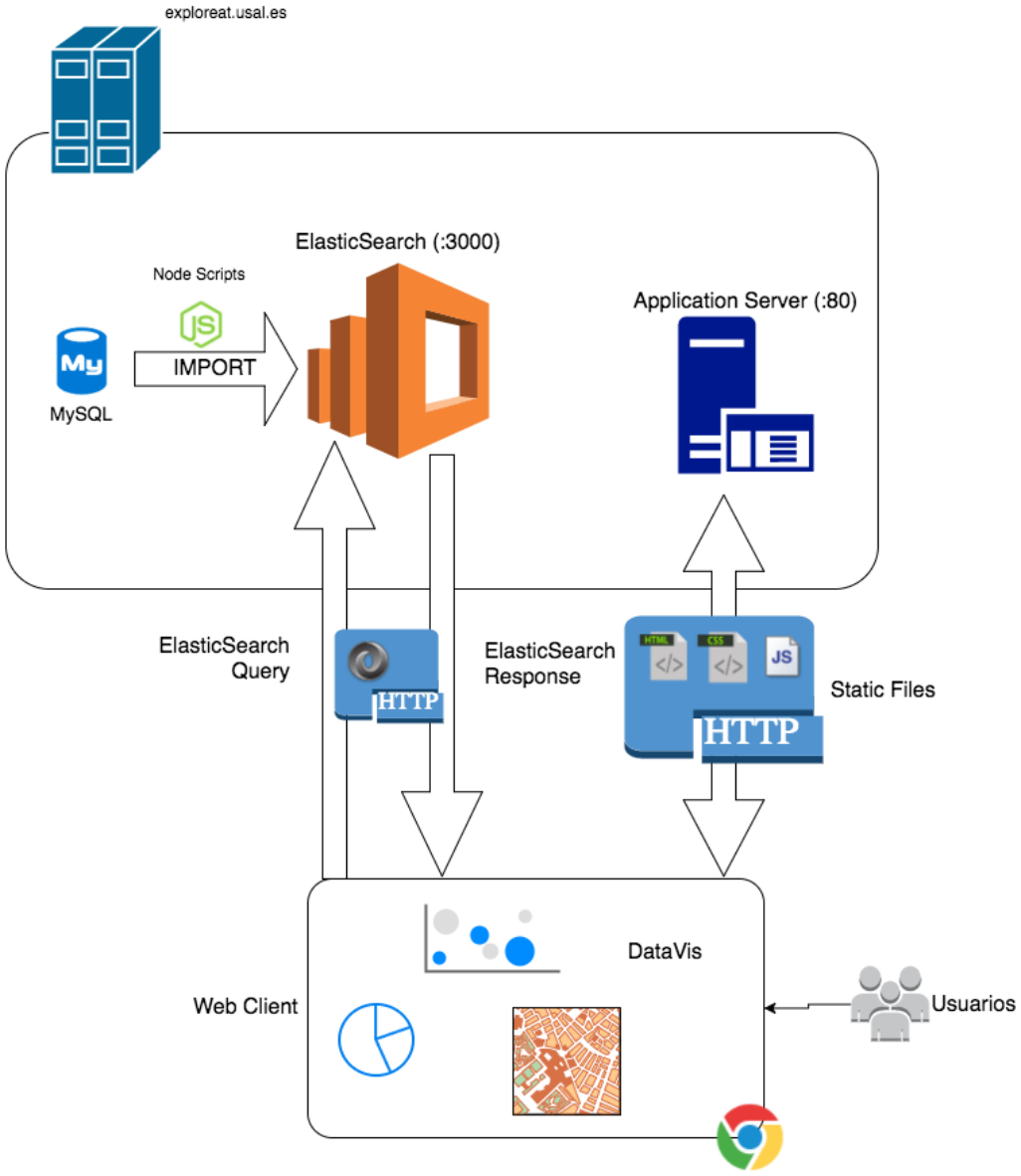


Figura 20: Arquitectura Web empleada en el sistema. El cliente recibe *assets* estáticos desde el servidor de aplicaciones. Sólo la información necesaria es transmitida desde el motor de búsquedas al motor de búsquedas al cliente en cada momento, con el consiguiente ahorro de recursos y mejora del rendimiento.

sin embargo, aplicar estas técnicas sin comprometer la experiencia de usuario. Encontramos, por tanto, los siguientes problemas: 1. Se necesitan estructuras de datos que contengan métricas generales sobre los datos analizados en base a agrupaciones. 2. Estas métricas no pueden ser construidas en un proceso previo ya que se perderían muchas de las opciones de análisis que se podrían ofrecer sobre el conjunto de datos. Para ejemplificar este hecho, imaginemos la situación siguiente: Se quiere realizar un *bubble map* que agregue diferentes localizaciones geográficas de elementos en el mapa, como ocurría en el prototipo creado por Therón et al a partir de la BD de dbo@ema[1]. En el momento que cambiemos los criterios de búsqueda o filtrado (empleando las vistas 2 o 3 en nuestro caso) necesitaremos recalcular todas

estas métricas de nuevo en base a un nuevo subconjunto de resultados que encaje con dichos criterios. Esto lleva a al problema 3. Este tipo de operación no es factible en nuestro enfoque porque llevaría a a) Disponer de todos los datos en la primera carga y aplicar las métricas en tiempo real, lo cual es inviable con el volumen de información o b) Precalcular todas las métricas para todas las situaciones de filtrado y búsqueda posibles, lo que en nuestro ejemplo tampoco es viable, ya que este número de posibilidades es, a efectos prácticos, infinito.

Esto hacía necesario otro tipo de solución que mantuviese tiempos de respuesta de la interfaz cercanos a las otras soluciones típicamente empleadas en visualización de datos sin sacrificar la funcionalidad. Para solucionar este problema empleamos el motor de búsqueda documental ElasticSearch y en concreto su funcionalidad de agregaciones por *buckets* presentada en la sección 4.3.3. Con esta técnica vamos a recuperar *buckets* o *clusters* de los datos de acuerdo a los criterios de búsqueda y filtrado introducidos por la usuaria. Este subconjunto de respuesta devuelto por el motor va a venir acompañado de una estructura de datos como la que buscábamos en un primer momento, pero con la ventaja de que este cálculo no se realiza en el cliente, cuya capacidad de procesamiento se emplea en agilizar las interacciones u otras tareas.

En la Figura 21 vemos una petición *Query-DSL* a ElasticSearch en formato JSON, dividida en dos partes fundamentales: *query* y *aggs*. En la primera se piden documentos que tengan información temporal disponible. En la segunda, se pide también que se agreguen los resultados coincidentes con el criterio de búsqueda en dos tipos de *buckets*, empleando la agregación *pipeline*: El primero agrega los resultados en función de la dimensión temporal, y estas agregaciones temporales se dividen también en función de la dimensión espacial. En el otro caso ocurre a la inversa.

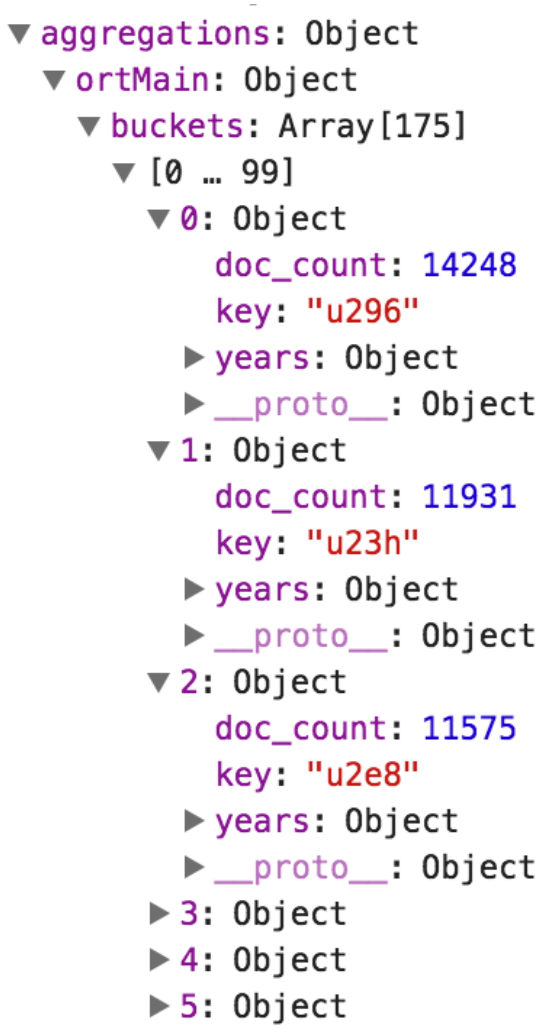
El resultado es que en todo momento van a existir dos estructuras de datos que van a ser la base de la visualización y que contienen datos a la resolución necesaria, maximizando así el rendimiento y el aprovechamiento de los recursos disponibles (Figura 20). Como veremos en los apartados siguientes, dentro de cada petición, la sección **query** controla el **conjunto de datos** a visualizar, mientras que las **agregaciones** manejan el **nivel de resolución** de los mismos.

5.2.1.1. Vista espacial

Como se ha apuntado, el mapa es la vista central sobre la que se basa el flujo de trabajo propuesto para el análisis exploratorio de los datos. Presentamos un detalle de la vista de mapa y de sus elementos en la Figura 23, de la que explicaremos a continuación sus diferentes partes:

Geohashes: Para la representación de las agregaciones empleamos una representación de *geohash*[51], un sistema de geocodificación que soporta búsquedas textuales. En este sistema, una cadena codifica una porción rectangular de terreno,

```
"{
  "aggs": {
    "ortMain": {
      "geohash_grid": {
        "buckets_path": "years",
        "field": "gisOrt",
        "precision": 3
      },
      "aggs": {
        "years": {
          "date_histogram": {
            "field": "startYear",
            "interval": "365d",
            "time_zone": "Europe/Berlin",
            "min_doc_count": 1
          }
        }
      }
    },
    "yearsMain": {
      "date_histogram": {
        "field": "startYear",
        "interval": "365d",
        "time_zone": "Europe/Berlin",
        "min_doc_count": 1
      },
      "aggs": {
        "ort": {
          "geohash_grid": {
            "buckets_path": "years",
            "field": "gisOrt",
            "precision": 3
          }
        }
      }
    }
  },
  "query": {
    "bool": {
      "must": [
        {
          "exists": {
            "field": "startYear"
          }
        }
      ]
    }
  },
  "size": 0
}"
```



```
▼ aggregations: Object
  ▼ ortMain: Object
    ▼ buckets: Array[175]
      ▼ [0 … 99]
        ▼ 0: Object
            doc_count: 14248
            key: "u296"
          ▶ years: Object
          ▶ __proto__: Object
        ▼ 1: Object
            doc_count: 11931
            key: "u23h"
          ▶ years: Object
          ▶ __proto__: Object
        ▼ 2: Object
            doc_count: 11575
            key: "u2e8"
          ▶ years: Object
          ▶ __proto__: Object
        ▶ 3: Object
        ▶ 4: Object
        ▶ 5: Object
```

Figura 21: Petición de *buckets* a Elastic-Search empleando una búsqueda abierta.

Figura 22: Respuesta a la petición. El tiempo de respuesta fue de 250 ms.

de manera que cuanto más larga es la cadena, más pequeña es la porción que define, y por tanto más resolución se obtiene. Lo bueno de este enfoque es que hay un número finito de *geohashes* inicialmente a mínima resolución y el resto de posibles *geohashes* tiene a éstas por prefijo. Siempre que un *geohash* es extensión de otro de menos resolución significa que el primero está contenido en el segundo. Este método es muy conveniente para nuestro enfoque ya que basa su potencia en la búsqueda textual, en la que ElasticSearch destaca. En la Figura 23.1 vemos la representación de un *geohash/bucket*. El mapa muestra todos aquellos donde se han encontrado resultados.

Escala: Cada *geohash*, por tanto, va a representar una agregación de todos los elementos que caen dentro de sus límites mediante una escala de colores que asocia el número de ocurrencias encontradas a tonalidades azules en orden creciente de oscuridad. (Figura 23.2)

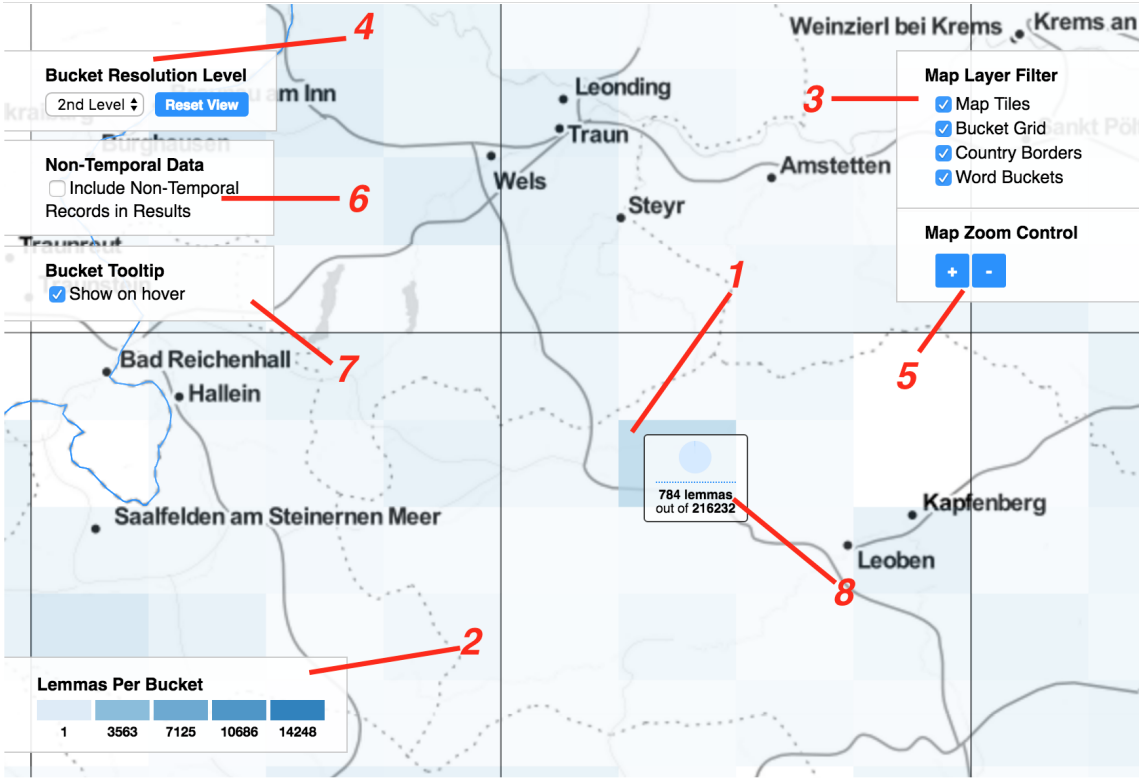


Figura 23: Detalle de la vista del mapa. 1)Geohash/Bucket espacial, 2)Escala, 3)Control de capas, 4)Control de resolución de los datos, 5)Control de zoom, 6)Control para incluir resultados sin información temporal, 7)Mostrar/Ocultar vista resumen del *bucket*, 8)Vista resumen del *bucket*.

Control de capas: El mapa muestra diferentes capas, cada una aportando un tipo de información diferente. En nuestro prototipo vamos a emplear 4 capas diferentes:

1. Capa de *tiles*: Muestra las imágenes del mapa. En ellas se muestran las distintas localidades, orografía, etc. Esta capa es fundamental y sirve para contextualizar la información mostrada en las otras capas.

2. Capa de *grid*: Muestra el nivel anterior de resolución al elegido. Sirve también para contextualizar en este caso los *buckets* que se muestran en cada momento, aportando una referencia visual sobre cuál es el *geohash* "padre" de cada uno.

3. Capa de fronteras: Remarca las fronteras de los diferentes países mostrados en el mapa. Esto es útil en el aspecto geográfico e histórico, ya que la analista tiene una referencia en todo momento de la posición de las fronteras actuales y así puede determinar si estas han variado o no desde el momento en el que se data una fuente, por ejemplo.

4. Capa de *buckets*: En esta capa se muestra la información de los términos en *buckets*, como explicamos en el anterior apartado.

Estas capas pueden ser ocultadas o mostradas a petición de la usuaria. Esta ocultación permite que la usuaria no sea distraída o molestada por elementos que no son relevantes en el momento de la exploración en el que se encuentre. (Figura 23.3)

Control de resolución: Como ya hemos comentado, las agregaciones son las responsables de controlar el nivel de resolución de los datos. Esta petición va a ser posible modificarla sin cambiar el nivel de zoom. En la Figura 23 podemos observar la misma porción del mapa mostrando datos a distinta resolución. En la marca número 4 de dicha imagen vemos el control que maneja la resolución en la vista del mapa.

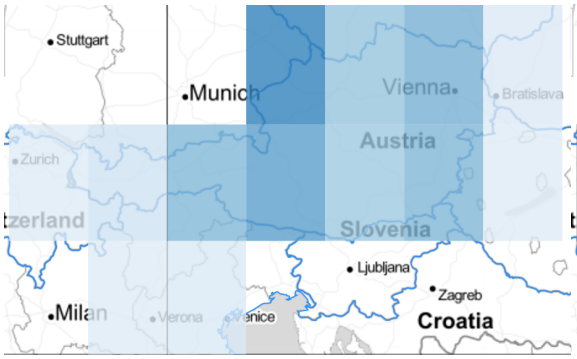


Figura 24: Menor resolución posible.

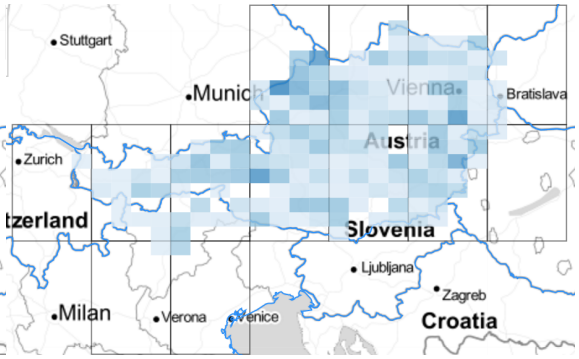


Figura 25: Un nivel más de resolución.

Control de zoom: Al igual que se dispone de control de la resolución de los datos, se ofrece también la opción de cambiar la proyección del mapa para que parezca que la usuaria se encuentra más cerca o más lejos de una parte concreta del globo terrestre. Este cambio de la proyección se llama zoom y se ofrece en todas las visualizaciones empleadas en el prototipo excepto en la línea temporal. En las Figuras 26 y 27 observamos cómo se comporta este cambio de proyección para una parte concreta del mapa. En la Figura 23.5 vemos el control para esta característica en la vista del mapa.

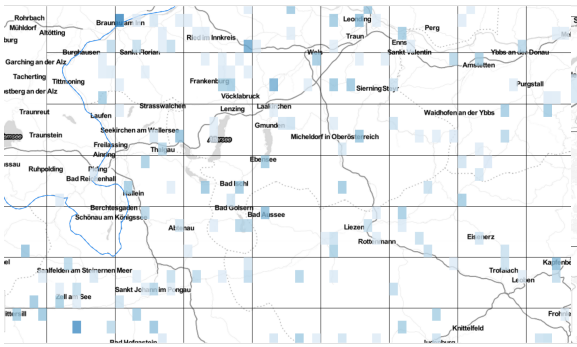


Figura 26: Una zona del mapa mostrada a un nivel de zoom.

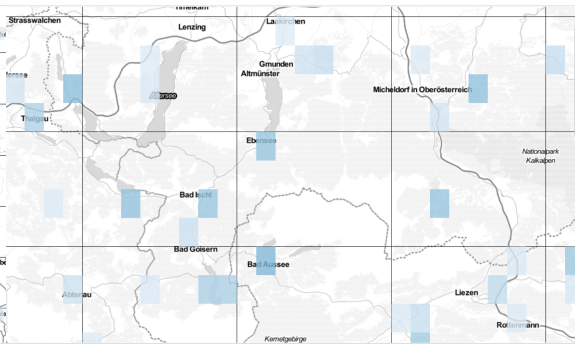


Figura 27: Misma zona mostrada a un nivel mayor de zoom.

Inclusión de resultados no temporales: Ya que existen diferentes subconjuntos de los datos, cada uno con una o dos dimensiones asociadas, se decidió por defecto trabajar con el subconjunto que contiene las dimensiones espacial y temporal asociadas en el mapa, y con el subconjunto que tiene información temporal (aunque no disponga de espacial) en el timeline. Sin embargo, no se podía ignorar el hecho de que los conjuntos que no contienen alguna de estas dimensiones pueden contener información valiosa para los investigadores y por tanto se decidió también incluir este subconjunto en los resultados por medio del control de la Figura 23.7.

Tooltip: Dentro del mapa, ésta es una de las vistas dedicadas a la última parte del *mantra de la visualización*: "[...], luego detalles bajo demanda". Cuando la usuaria desliza el puntero encima de un *bucket*, se muestra una vista resumen del mismo, que indica el número exacto de elementos recogidos en éste, así como el porcentaje que representa dentro del total de los resultados (Figura 23.8).

5.2.1.2. Línea temporal

De manera análoga a la vista espacial del mapa se implementa la funcionalidad de la línea temporal. En ella se proyectan todos los datos que cuentan con este tipo de información. Como hicimos en el apartado dedicado a la vista espacial, presentamos un detalle de la misma en la Figura 28:

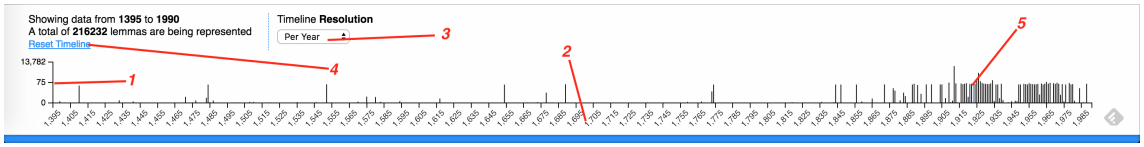


Figura 28: Detalle de la línea temporal. 1)Escala, 2)Representación de la dimensión temporal, 3)Control de resolución, 4)Texto explicativo y función de reset, 5)Barras y highlighting

Proyección: En esta línea temporal se proyectan los resultados obtenidos a través de la búsqueda en base al año en el que se registraron. En el eje X se aplica la dimensión temporal, mientras que en el Y se muestran el número de ocurrencias en cada año por medio de barras en base a la escala mostrada en la marca 1 de la figura. Esta escala relaciona la longitud de las barras con este número de ocurrencias (a más ocurrencias, la longitud de la barra es mayor, en base a un mínimo y a un máximo globales). Ambos ejes varían en base a los resultados obtenidos, cambiando la escala y la representación de los años en ambos ejes. En la Figura 29 vemos la misma barra con sus escalas modificadas en base a un nuevo conjunto de resultados.

Control de resolución: Otro aspecto importante de esta representación es el control de la resolución, que permite adaptar la escala a petición de la analista

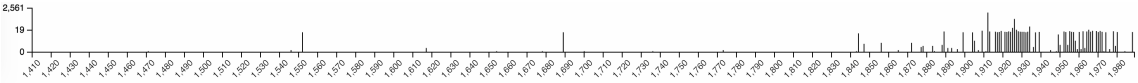


Figura 29: La línea temporal mostrando un nuevo conjunto de resultados. Nótese cómo varía la escala mostrada en el eje Y en base al mínimo y máximo encontrados, manteniendo la misma longitud. De manera análoga, el eje X muestra un nuevo conjunto de años en base a los mismos criterios.

para facilitar la visualización de ciertas características. En nuestro prototipo permitimos elegir entre 4 opciones a través del control de la Figura 28.3. En la figura a continuación mostramos una comparativa del mismo conjunto de datos mostrado a resoluciones de 1 y 25 años, respectivamente:

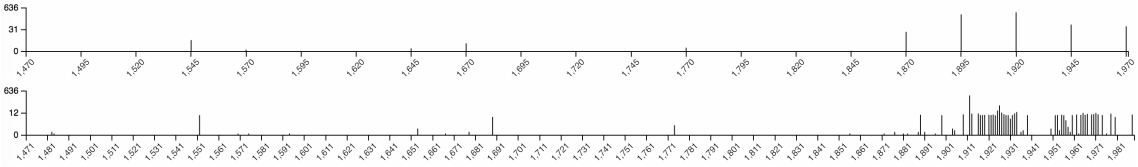


Figura 30: Conjunto de datos proyectado a resoluciones de 25 años (arriba) y 1 año (abajo)

Vista de detalle y control de reset: En este apartado se muestra una vista de detalle con unas cifras que resumen lo que se está viendo en la línea temporal: La cantidad de registros y los años mínimo y máximo encontrados en los mismos. También se incluye un control de reset, que elimina los filtros que se hayan aplicado a la línea temporal, lo cual se explica en las siguientes secciones.

Highlighting: La vista temporal soporta también una técnica de visualización ya introducida en las secciones iniciales de este documento, denominada *highlighting*. En esta práctica, se resaltan en colores diferentes aquellas porciones de datos que se seleccionan en otras vistas enlazadas, como el mapa o el grafo. Ésto aligera también la carga cognitiva asociada a la exploración de los datos y acelera la llegada a conclusiones significativas por parte de la analista.

5.2.2. Zoom y filtrado

En esta segunda parte del flujo de trabajo se requiere la interacción por parte de la analista, que interaccionará con la aplicación a través de las observaciones realizadas en la primera gracias a las ayudas visuales explicadas. Es ahora donde la usuaria aplicará el filtrado y el zoom para reflejar su estado mental, centrado en una parte de los resultados mostrados. Nuestro prototipo soporta tres tipos de filtrado de los diferentes documentos: 1. Espacial, a través de los elementos interactivos mostrados en el mapa; 2. Temporal, empleando los recursos ofrecidos por el *timeline*; y 3. Textual, que permitirá hacer búsquedas complejas de cadenas en el campo “HL”

del registro original, que fue debidamente procesado en la etapa de adquisición de los datos. El flujo de trabajo propuesto va a ir combinando estos tipos de filtrado en sucesivas etapas de refinamiento de los datos hasta que se produzca el descubrimiento de conocimiento.

5.2.2.1. Filtrado espacial

Antes de pasar al filtrado textual, volvamos momentáneamente al mapa para explicar este proceso: Cada uno de los *buckets* mostrados en el mapa es susceptible de interacción. Cuando la usuaria, bien a través de la vista general ofrecida de los datos o bien con la ayuda de las vistas auxiliares que aparecen al deslizar el ratón sobre cada *bucket*, lo selecciona, hace explícito su interés por continuar su análisis en esa parte del mapa. Cuando este sucede, se generan una serie de acciones y animaciones en la interfaz que listamos a continuación:

1. Se realiza una acción de zoom sobre la zona geográfica comprendida por el *bucket* seleccionado.
2. Se cambia automáticamente la resolución del mapa a un nivel adecuado para dicho nivel de zoom.
3. Se recuperan elementos desde el motor de búsqueda y se muestran en los diferentes *buckets*. Se actualiza también la escala de colores del mapa.
4. Se presentan los elementos de análisis visual de redes en pantalla (Ver sección correspondiente más adelante).
5. La interfaz refleja un nuevo estado mental de la usuaria. Es ahora donde se ofrece una nueva vista general de manera análoga a lo visto anteriormente, pero que emplea un subconjunto de los datos creado a partir de los de la anterior etapa.
6. El flujo se repite indefinidamente, quizás con la inclusión de otros tipos de filtrado o zoom provenientes de otros elementos de la interfaz.

En la Figura 31 ilustramos este proceso. En un primer instante, una cierta zona del mapa llama la atención a la analista, que interacciona con el *bucket* (Arriba). Al pulsar sobre él, se lanza la cadena de eventos que termina con el estado de la interfaz mostrado abajo.

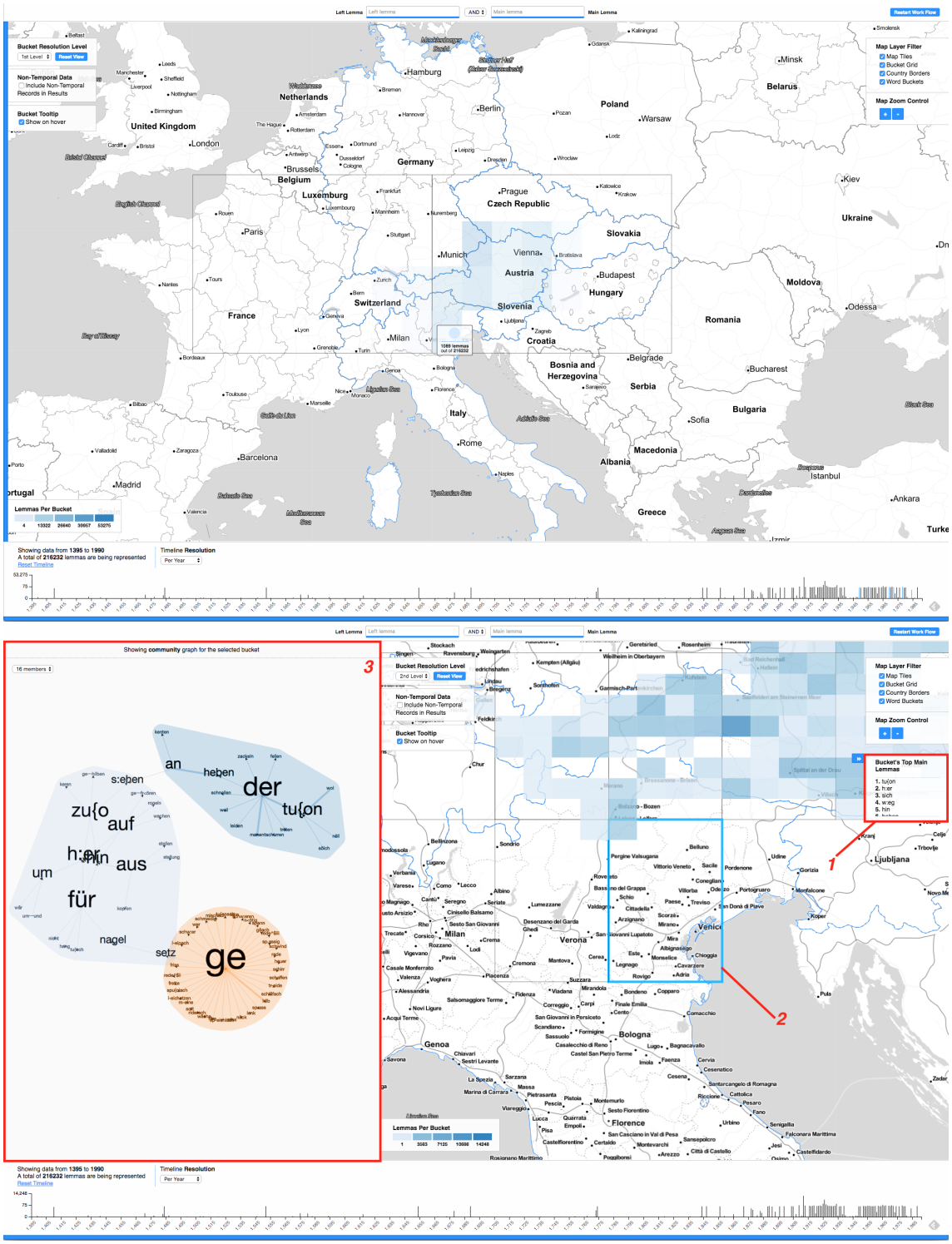


Figura 31: Los dos estados de la interfaz antes y después de realizar el filtrado espacial.

En 1 y 3 se muestran los nuevos elementos de la interfaz correspondientes al análisis de redes. En 1, los lemas (nodos) más importantes, en base al número de relaciones con otros lemas, son mostrados a la usuaria. En 3, una representación visual de las relaciones encontradas en forma de grafo dirigido de fuerzas. En 2

se añade una ayuda visual que permite a la usuaria recordar qué *bucket* de nivel inmediatamente superior fue pulsado al inicio de la interacción.

5.2.2.2. Filtrado temporal

Además del filtrado espacial, la usuaria puede decidir trabajar con un subconjunto de los datos elegido a través de la realización de un filtrado de la variable temporal. Al existir una relación entre las dimensiones en memoria, este tipo de filtrado no necesita de nuevas peticiones y es, por tanto, sensiblemente más rápido. Este filtrado se consigue mediante la acción de arrastrar, como es común en este tipo de elemento visual. La zona elegida se representa por un sombreado que, al modificarse actualiza los datos mostrados en el mapa. Una vez fijado el filtro temporal, se mantiene en etapas subsecuentes de refinamiento si no es modificado por la usuaria. En la Figura 32 se muestra el mismo conjunto de datos filtrado primeramente en el intervalo formado por los años 1404-1932 (arriba) y 1932-1987 (abajo). Obsérvese cómo la distribución espacial cambia en el mapa, así como la escala de colores, para reflejar cada uno de los dos subconjuntos formados.

5.2.2.3. Filtrado textual o búsqueda de cadenas

El filtrado textual en tiempo real supone uno de los avances más importantes de la investigación. Gracias a esta capacidad, la analista puede buscar patrones de coincidencia entre fuentes coincidentes con los criterios de búsqueda textual introducidos. Es aquí donde las capacidades del motor ElasticSearch sobresalen sobre otros tipos de implementaciones (Por ejemplo búsqueda textual en MySQL o BBDD NOSQL). Gracias al modelo elegido, la interfaz se mantiene en un estado responsivo en todo momento y permite una interacción dinámica y natural con los datos que reduce en gran medida el esfuerzo necesario inherente a la tarea de análisis desempeñada.

Sintaxis de búsquedas: Como decíamos en anteriores secciones cuando presentábamos el motor de búsquedas, la búsqueda de cadenas es la especialidad de ElasticSearch, tanto que podríamos afirmar que éste fue concebido especialmente para esta tarea. Para ello se emplea la sintaxis de búsqueda de *Lucene* basada en expresiones regulares. Esta potente sintaxis va a soportar un extenso conjunto de operaciones: lógicas, difusas o de proximidad que son de alto valor en el ámbito del estudio de la lexicografía. Con ellas, la analista va a poder buscar conjuntos de palabras que tienen la misma raíz léxica o cuya pronunciación es parecida, que sirvan de base para su investigación. Gracias a la visualización de datos, conseguimos abrir la compleja potencia del motor de búsquedas y sus técnicas de Procesamiento del Lengua Natural a la usuaria no experta en estas materias, cumpliendo así uno de los objetivos de los proyectos de Humanidades Digitales marcados al inicio de este estudio.

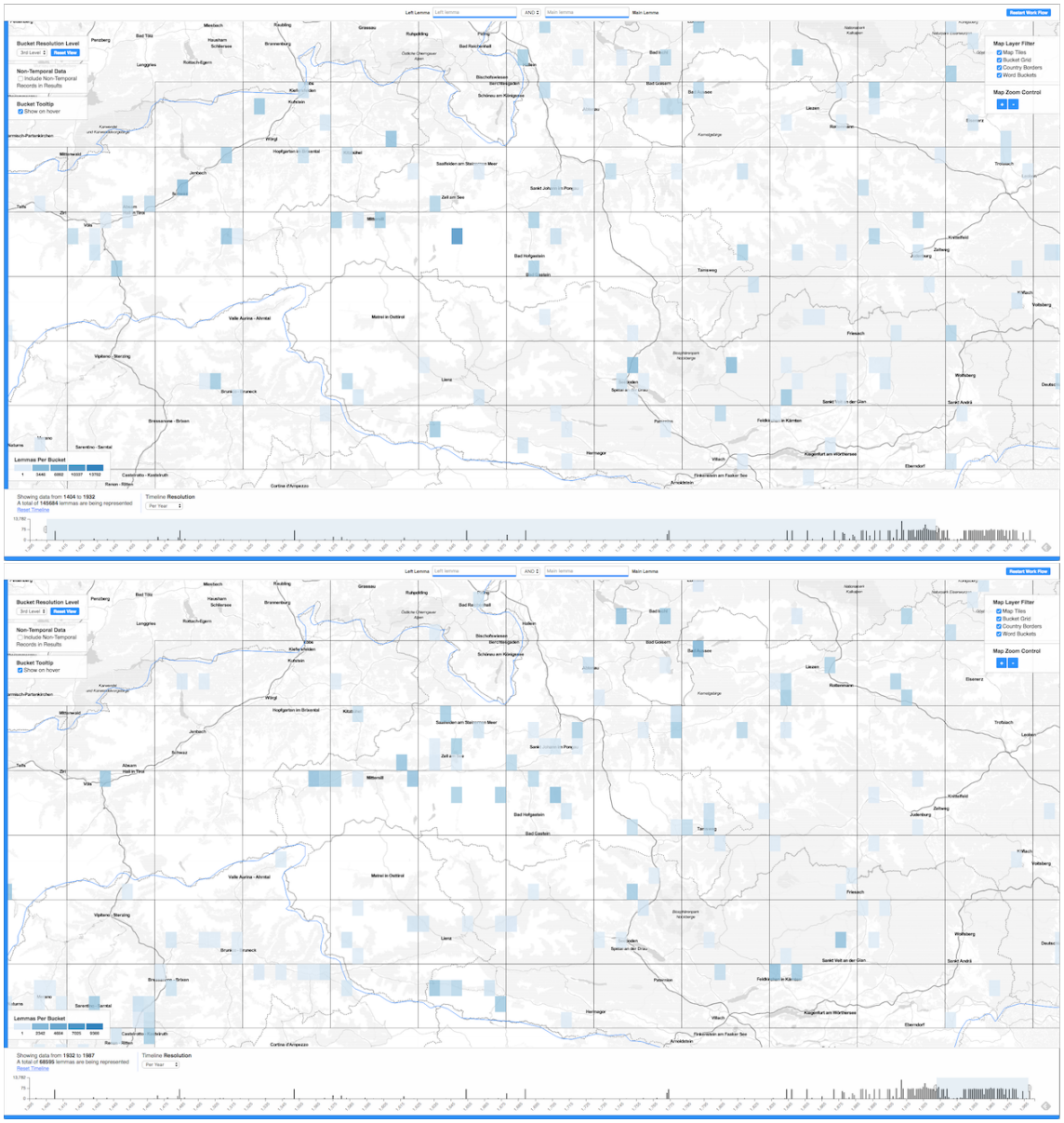


Figura 32: Dos capturas de la interfaz mostrando filtrados temporales en intervalos diferentes para el mismo conjunto de datos.

Combinación de parámetros: En una primera versión del prototipo, soportamos la búsqueda por las diferentes partes del lema, que recordemos habíamos denominado *leftLemma* y *rightLemma*. Para combinar la búsqueda por estos dos campos, se habilita también un selector lógico booleano AND/OR, que permite combinar ambas de dos formas diferentes. En la Figura 33 presentamos un ejemplo de este funcionalidad. En cada una de las cajas de búsqueda empleamos una característica de la sintaxis *Lucene* y las combinamos mediante operadores lógicos diferentes para lograr conjuntos de resultados diferentes.

En el ejemplo presentado se buscan documentos cuyo campo *leftLemma* empiece por la letra "a" (expresado en sintaxis Lucene con el operador estrella "*"). Además, se va a combinar con una búsqueda difusa basada en la distancia de Levenshtein

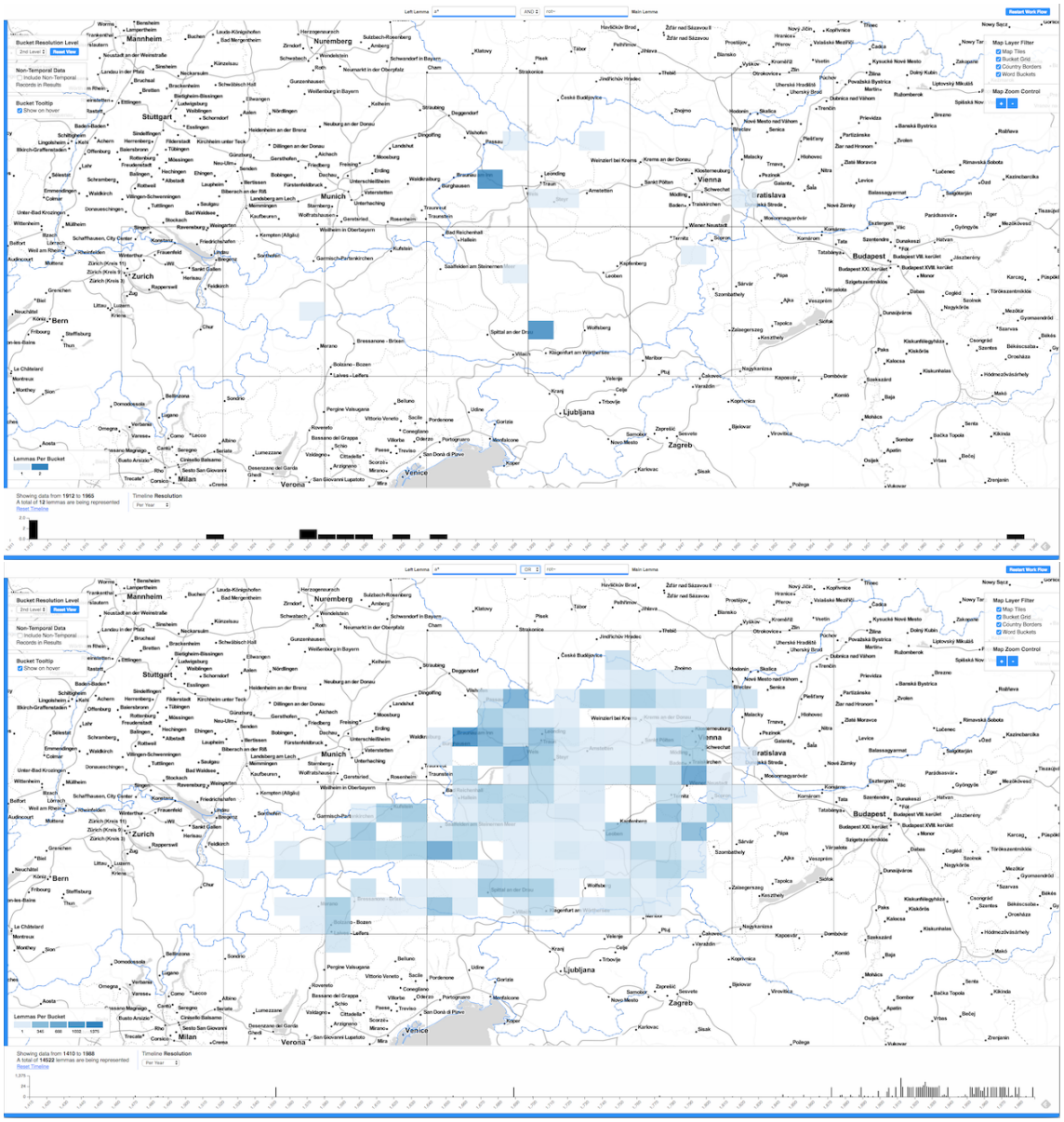


Figura 33: Ejemplos de búsqueda textual combinada mediante el operador AND (arriba) y OR (abajo), produciendo conjuntos de resultados diferentes que son proyectados en el mapa y en la línea temporal.

entre palabras, que será cercana a 1. Esta inclusión será particularmente importante en el estudio de dialectos[52], especialmente en el tratamiento de las palabras homófonas. En la Figura 33 se observa como la combinación de los dos criterios escogidos genera diferentes representaciones visuales de los conjuntos de datos. Como se puede comprobar, resulta muy sencillo para la analista acceder a una funcionalidad compleja del motor búsqueda, así como crear un mapa mental de la situación planteada por la inclusión de parámetros lógicos.

5.2.2.4. Análisis de Redes

Ya se apuntó en los apartados anteriores la posibilidad de crear estructuras visuales para el análisis de redes. Dichas estructuras son generadas a partir de los conjuntos de datos que encajen con los parámetros de búsqueda introducidos por la analista, bien automáticamente o bien a petición de la usuaria, dependiendo del caso. Estas estructuras suponen el final de una iteración en el flujo de trabajo de la analista, ya que son capaces de desvelar relaciones ocultas entre los lemas imposibles de descubrir por medio de otros tipos de análisis (espacial o temporal). A partir de elementos interactivos, y coincidiendo con el precepto "detalles en demanda" del mantra de la visualización, la analista va a poder refinar su flujo de trabajo y adaptar los datos a las coincidencias encontradas mediante la exploración visual de las redes de lemas.

Grafo dirigido de fuerzas: El primer tipo de representación visual que se presenta a la analista para el análisis visual de redes es el grafo dirigido de fuerzas, que es combinado con el enfoque de nube de palabras, para transmitir la idea de dos valiosos conceptos: la **importancia** de un lema y la **pertenencia** a un grupo del mismo. Esta visualización presenta varias características:

1. Genera un grafo de los elementos que entran dentro del criterio de búsqueda de la analista (una combinación de filtrado textual, espacial y textual)

2. En él se representan las relaciones entre lemas, en base a sus partes izquierda y derecha. Un lema que aparece en una entrada a la izquierda de otro aparecerá en el grafo como origen de la arista que los une.

3. Cada lema está representado en el grafo por las letras que lo componen, que formarán el nodo. Este nodo variará en tamaño en base a un escala lineal que relaciona el tamaño con la **importancia** del lema en el conjunto o lo que es lo mismo, en base al número de relaciones (aristas que llegan o parten) de ese nodo.

4. Debido al gran tamaño que pueden presentar estos grafos, especialmente en la búsqueda exploratoria, se genera además un análisis de comunidades sobre dicho grafo, método mencionado en el trabajo de Mayer et al.(2014)[31] para la búsqueda de patrones de colexificación. Ahondaremos más en la validez y detalles de este aspecto en la siguiente sección.

5. Se aplica además un filtrado dinámico en base al tamaño medio de las comunidades detectadas en el grafo, que permite dar una primera visión del grafo lo más adecuada posible a la analista.

Análisis de Comunidades Nuestro flujo propuesto encuentra comunidades no superpuestas de nodos en los grafos con el objetivo de realizar particiones de los grafos que aporten valor a la investigación a la hora de la búsqueda de patrones

reconocibles en los datos. En este tipo de particiones del grafo, la red se divide de forma natural en grupos de nodos densamente conectados internamente y con pocas conexiones con otros grupos. El algoritmo de detección de comunidades fue el llamado *Louvain*[53], ya que era importante que este algoritmo se ejecutase en tiempos lo más cortos posibles, en el intento de crear una interfaz suficientemente responsiva sin sacrificar en demasía la exactitud de los resultados. En implementaciones del algoritmo escogido[54], se llega a generar una estructura comunitaria en grafos con 2 millones de elementos en 2 minutos, por lo que consideramos este método como suficientemente bueno[55] para poder ejecutarse en nuestro entorno. La versión final ejecutada en el prototipo, que es compatible con el navegador, es una modificación de la versión encontrada en [56].

En la Figura 34 mostramos el grafo generado para la región delimitada por el geohash "u20", donde podemos ver las distintas comunidades representadas visualmente mediante la técnica de *Convex Hull*, cada una con un color proveniente de una escala categórica creada a tal efecto.

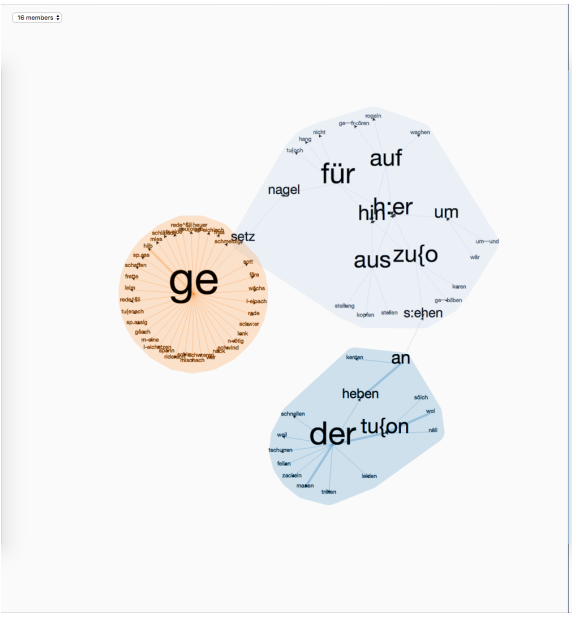


Figura 34: Vista inicial del grafo, con el filtro activado por defecto a 16 miembros.

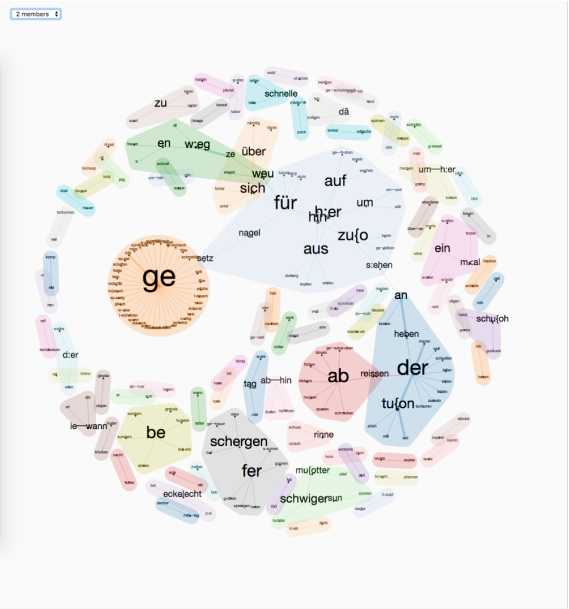


Figura 35: El grafo con el nivel de filtro activado a 2 miembros, muestra comunidades menos relevantes

Como se aprecia en las dos imágenes adjuntas, el grafo es filtrado mediante el control de la esquina superior izquierda. Este control se activa por defecto al valor medio del tamaño de las comunidades encontradas, de manera que las comunidades menos importantes son ocultadas al inicio. En la Figura 35 se ha modificado este valor a petición de la analista, mostrando el resto de elementos del grafo. Las comunidades más grandes, ocupando más espacio, desplazan a las menos importantes hacia los extremos del lienzo sin comprometer la expresividad de la visualización. Este tipo de representación permite también interactuar con los nodos, así como hacer zoom y desplazamiento, como en el caso del mapa. En esta vista la analista va a poder analizar de manera visual las diferentes comunidades creadas y sus relaciones,

creando un mapa mental del conjunto analizado que va a ayudar a la investigación y por tanto a la llegada a conclusiones significativas.

5.2.3. Detalles en demanda después

Hasta ahora hemos hecho referencia a las dos primeras partes del mantra de la visualización, quedando la última y no menos importante por comentar. Es en esta última parte del flujo de trabajo donde la usuaria, a través de las vistas globales, identifica un hecho significativo, a saber en nuestro caso: Cierta predominancia de una clase de lemas a originarse en partes concretas del mapa, ciertas relaciones entre lemas que tienden a repetirse, etc. la usuaria, inconscientemente, ha fijado su estado mental en este hecho y es por tanto necesario proporcionarle la opción de actualizar la interfaz en concordancia con el mismo. En nuestro enfoque existen diferentes maneras específicas de lograr esto (además de modificar alguna de las ya citadas más generales), con la peculiaridad de que todas ellas necesitan de la interacción de la usuaria (a diferencia de las anteriores, que sucedían de un modo más o menos automático). Continuamos por tanto en esta sección donde terminamos la otra, para remarcar el hecho de que es precisamente el análisis de redes el que va a servir de nexo de unión entre las dos fases y el que va a marcar también el inicio de una nueva iteración del ciclo de trabajo propuesto.

5.2.3.1. Análisis de redes

Gráfico de árbol: Aparte del análisis de comunidades, que se aplica a grupos de lemas, se añadió a mayores la opción de realizar otro análisis visual orientado a un sólo lema, que también representase la red formada por el mismo en relación a otros. Recordemos cómo en la Figura 32.1 se mostraba, al seleccionar un *bucket*, una vista de detalle que incluía en orden decreciente de importancia los lemas encontrados en el mismo. Mediante la interacción de la analista se van a poder realizar las siguientes acciones sobre cada uno de ellos:

1. Generar gráfico de árbol para la red en la que el lema seleccionado es el nodo principal, teniendo en cuenta todo el conjunto de datos (se ignoran los filtros previos).

2. Generar gráfico de árbol para la red en la que el lema seleccionado es el nodo principal, respetando las opciones de filtrado previamente aplicadas.

3. Realizar una búsqueda textual de ese lema y proyectar los resultados en el mapa.

Una característica añadida a este análisis individual de lemas, es que el gráfico de árbol permite el intercambio de posición del lema escogido, mostrando coincidencias de resultados en los que éste aparece en la parte izquierda o derecha, a petición de la analista. En la Figura 36, se expande la red para relaciones con 20, 30 y 50 resultados.

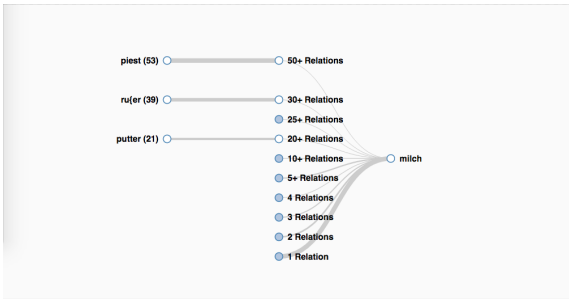


Figura 36: El gráfico de árbol que visualiza la red para el lema *milch* en la parte derecha o principal.

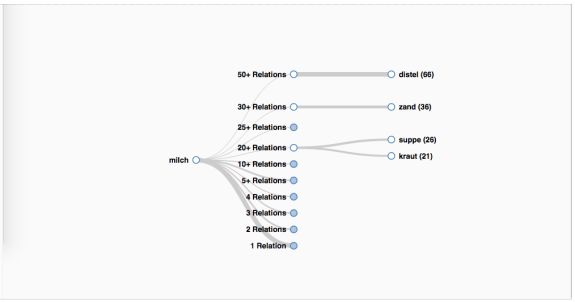


Figura 37: El gráfico de árbol con la red formada con *milch* en la parte izquierda del lema.

Se ve que los casos con más relaciones forman los términos *puttermilch* y *piestmilch* que hacen referencia al suero (que se obtiene en el proceso de hacer queso) y a la leche animal, respectivamente. Cuando se cambia el orden en el que aparece *milch* dentro del lema, y de manera análoga se navegan las conexiones más pobladas, se forman otros lemas derivados, que son también los más comunes: *milchdistel* (cardo de la leche) y *milchzand* (diente de leche). Para completar el proceso de análisis, se da la posibilidad de que la analista lance nuevas búsquedas interaccionando con los elementos del árbol, una vez que éste ha sido desplegado hasta un nodo hoja. En el caso presentado, al seleccionar una rama, se lanzaría una nueva búsqueda textual con *milch* como lema principal o izquierdo, según corresponda, y el otro término como parámetro complementario, comenzando así una nueva búsqueda. Simplemente con desplazar el cursor sobre las diferentes ramas, se emplea también *highlighting* en la proyección temporal, dando una idea del reparto de la relación en el tiempo.

Grafo dirigido de fuerzas: Volvemos ahora al grafo dirigido de fuerzas presentado en la sección anterior, que también cuenta con capacidades para lanzar nuevas búsquedas en demanda de la analista o generar árboles para un lema en concreto. Se ofrecen, a través de la interacción con los elementos del grafo, 3 opciones:

1. Interaccionando con los **nodos** del grafo se podrá:

   a) Lanzar una búsqueda textual con el lema en la posición en la que se encuentre en el grafo. Si éste se encuentra en ambas, se lanzará una búsqueda OR con el mismo término en ambos lados.

   b) Generar un árbol con las relaciones encontradas en el grafo para el lema escogido

2. Si se interacciona con las **comunidades**, se podrá lanzar una búsqueda textual con todos los elementos de la comunidad en los lados que corresponda. la analista podrá modificar el operador lógico que une ambas partes del *query* posteriormente.

5.2.3.2. Registro original

En ocasiones, también puede resultar interesante ver el contenido original del registro TUSTEP con el objeto de complementar la información de la que dispone para un registro en concreto. Este tipo de información se muestra cuando la búsqueda ha reducido tanto el conjunto de datos que existen *buckets* con una sola instancia. Es entonces que al seleccionarlos se muestran los campos a la usuaria, como se recoge en la Figura 38.

Figura 38: Vista detalle que muestra los campos originales TUSTEP de un elemento de la visualización.

5.2.4. Cerrando el ciclo de trabajo

Después de una serie de iteraciones del ciclo propuesto, la usuaria considerará la sesión como terminada, bien porque se ha llegado a conclusiones sobre los datos estudiados y se desea continuar la investigación a partir de esta nueva base de conocimiento, o bien porque no se ha podido llegar a una conclusión y se considera necesario emprender otro análisis empleando técnicas diferentes. Sea como fuere, lo que se haga a partir de este momento termina con un ciclo en el proceso mental de la usuaria, y por tanto se han de descartar los estados previos que pueda reflejar aún la interfaz. Para explicitar este deseo de la usuaria, se dispone de un botón en

la parte superior derecha que limpia la interfaz y la lleva de vuelta al estado inicial, siendo a todos los efectos, equivalente a abrir la aplicación de nuevo. En el proceso suceden, de manera paralela, los siguientes eventos:

1. Se limpia el área de análisis de redes y se cierra el panel correspondiente a esta zona.
2. La búsqueda textual se devuelve al estado de *match all*, que es la búsqueda que recupera todos los resultados.
3. Se eliminan los filtros temporales y se proyectan los resultados de la búsqueda explicada en 2.
4. Se limpia el mapa de *buckets*, mostrando el conjunto inicial de acuerdo a la búsqueda explicada en 2.
5. Se restaura la proyección original del mapa, así como la resolución, que vuelve a su nivel más bajo.

Una vez realizadas estas tareas, se considera que la interfaz está lista para emprender un nuevo flujo de trabajo distinto y la usuaria puede continuar con la investigación.

6. Casos de estudio

En este apartado presentaremos dos ejemplos de casos de estudio desarrollados en conjunto con el equipo de expertas en lexicografía que ha colaborado con esta investigación. Después de las sesiones necesarias para instruirles en el uso del prototipo desarrollado, les pedimos que intentasen realizar algunas de las tareas de su proceso de investigación habitual, y que anotasen en qué manera el prototipo ayudó a mejorar y acelerar la extracción de conocimiento. También les pedimos que nos proporcionasen información sobre las partes del proceso de exploración visual propuesto que contribuyeron en mayor medida a crear un mapa mental más fidedigno de la situación analizada, y que fueron capaces de aportar conocimiento añadido en relación a su metodología habitual de trabajo. Por último también se les solicitó que reseñasen en qué momentos notaron que la herramienta no fue capaz de reflejar correctamente la intencionalidad o dirección de la investigación, y que apuntasen aquellas carencias que creían que se hicieron más fehacientes durante la utilización del prototipo. Esta información es detallada en la sección 7 a continuación de la presente.

En las variaciones dialectales de una lengua se producen una gran cantidad de fenómenos linguísticos que guardan relación con la manera en la que los términos son transmitidos entre territorios. Las palabras, a lo largo de la historia, viajan de unas partes del área de influencia de una cierta lengua. Con el tiempo y en cada zona, estos lemas degeneran en relación a la pronunciación y escritura originales, provocando alteraciones en el significado, en la escritura y en la pronunciación de cada una. Es por tanto interesante estudiar las generalidades y particularidades de estos eventos, en un intento de datar el origen de los términos empleados en cada región.

Algunas de las tareas más habituales en los procesos de estudio de diccionarios históricos y muy en especial en el de diccionarios dialectales, como es nuestro caso, incluyen el análisis temático del léxico en zonas geográficas concretas o la datación de tendencias en el uso del lenguaje, así como también comprobar el efecto de la degeneración de los lemas originales en la composición de nuevas palabras homófonas y otros accidentes.

6.1. Apariciones de un lema en una posición concreta

Este caso de estudio se centra en la temática del color rojo (los colores es un aspecto empleado con frecuencia por el grupo de expertas). Se parte de la siguiente pregunta de estudio: **¿Qué palabras compuestas contienen a *rot* (rojo) como lema principal y cuál es el lema que aparece más frecuentemente como lema prefijo en las mismas en las regiones del oeste de Austria?** Es así que el estudio comienza por una búsqueda dirigida, en la que la usuaria ya expresa la intencionalidad de centrarse en un subconjunto de los datos, en concreto en aquél en el que el lema sufijo es el color rojo. Ya que en un principio la interfaz no va a reflejas este estado, es necesario que la usuaria interaccione con la aplicación a

través de las cajas de búsqueda textual, introduciendo la palabra buscada (nótese que además se busca una pronunciación concreta). Una vez realizada la acción, el sistema lanza la búsqueda hacia el motor de búsquedas, que devuelve un subconjunto de los resultados que se mostraban en el paso anterior. Véase la Figura 39 para comprobar cuál es el estado de la interfaz en este punto.

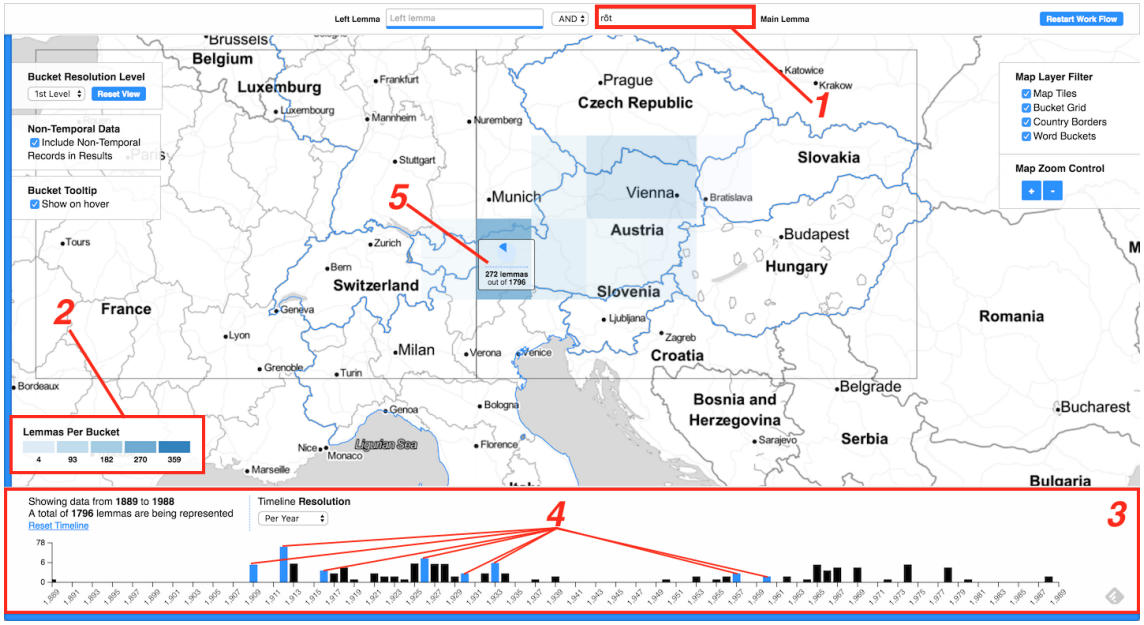


Figura 39: Visualización del conjunto de datos para el lema "rôt", mostrando las proyecciones geográficas y temporales del mismo al mínimo nivel de resolución.

6.1.1. Visión general primero

Ahora que la analista ha fijado la interfaz en el estado deseado mediante la interacción con la caja de entrada de texto para búsquedas sobre el lema principal (Figura 39.1), puede hacer uso de las diferentes ayudas visuales que brinda el prototipo en esta vista general. Ya que la tarea de análisis acaba de comenzar, las resoluciones temporales y visuales encajan con el nivel de zoom aplicado, que es el mínimo, permitiendo que se reciba una visión global de la distribución en ambas dimensiones del conjunto de resultados.

En una **primera inspección** de esta interfaz, la usuaria fijará su vista primeramente en la parte más notoria de la interfaz, que es el mapa. En él puede observar la distribución de los lemas con "rôt" en su parte derecha o principal. Además, en una primera mirada y sin requerir ningún tipo de acción manual, se buscará la escala (Parte inferior de la izquierda, Figura 39.2) que relaciona la tonalidad del color azul de los *buckets* con el número de resultados que han "caído" en los mismos. Esta escala a simple vista indica un mínimo de 4 y un máximo de 359 ocurrencias por *bucket*. La distribución de los colores denota dos zonas con un número elevado de registros (Viena y su colindante por el oeste), mientras que los *buckets* de los extremos del país contienen un número mucho menor. Una **primera conclusión** a la que

llega la analista en este punto sin requerir de más interacción por su parte, es que **los resultados no están uniformemente distribuidos en el espacio**. Además, ya que la intención de la analista es centrar su estudio en las zonas al oeste del territorio, buscará esa parte del mapa espontáneamente, observando que el *bucket* posicionado al sur de Munich contiene un número medio-alto de ocurrencias.

La analista después se centra en la parte inferior de la pantalla, donde reside la barra temporal 39.3) Entre los elementos analizados, los primeros en interiorizar serán aquellos visuales que denoten también mínimos, máximos y ausencia de datos. Entre el conocimiento inmediato que se transmite está: 1) La existencia de una continuidad temporal en la recopilación de registros, excepto por un **período de ausencia** de los mismos que, fijándonos en más detalle podemos comprobar que se extiende desde 1939 a 1950 (Una posible explicación histórica a este hecho puede ser la 2ª Guerra Mundial, comprendida entre los años 1939-45 y que llevaría a la suspensión de la actividad académica). 2) El **máximo de ocurrencias tiene lugar en el año 1912**, el mínimo sin embargo no está claro y requiere de un análisis más concienzudo de la visualización.

En un **segundo paso**, la información es adquirida también visualmente pero por otro canal cognitivo: ahora que ya se han comprendido las generalidades básicas, es turno de buscar detalles. Este tipo de detalles necesitan de la **lectura** de información textual presente en la interfaz: La usuaria podría ahora leer la leyenda de la línea temporal, repasando el conjunto de años sobre el que se proyectan los datos (de 1889 a 1988) o desplazar el razón sobre los diferentes *buckets*, que mostrarían un vista de resumen/detalle como la mostrada en 39.5. Además, mediante la técnica de *highlighting* empleada en el prototipo la usuaria va a recibir información visual sobre la distribución temporal de los elementos de cada *bucket*. En nuestro ejemplo (39.4) vemos que la mayoría de los registros del *bucket* se sitúan en un período de 30 años que va de 1909 a 1939.

En un **tercer y último paso** la usuaria ha de decidir qué subconjunto de los datos mostrados es de su interés. Es ahora que podría modificar la búsqueda textual original para reducir aún más el número de resultados en caso de disponer de más criterios para ello, o realizar alguna acción que transmita el estado mental al que ha sido movida por la información adquirida en los pasos anteriores, efectuando un seccionamiento de los datos a través de los controles disponibles en la interfaz (el filtro temporal o la interacción con los *buckets*). En este caso de estudio, centra su atención en el *bucket* que se señala en la figura, interactuando con el mismo a través de la acción de *click*.

6.1.2. Zoom y filtrado

Como ya se explicó en secciones anteriores al comentar la funcionalidad ofrecida por el prototipo, al seleccionar un *bucket* la usuaria expresa el deseo de interactuar sólo con un segmento de los datos debido a una serie de circunstancias o características del mismo que le hayan resultado interesantes para el objeto de su estudio. Al recibir esta acción, la aplicación realiza una animación de traslación y zoom, en la

que la proyección del mapa cambia para centrar los datos seleccionados en pantalla. A la vez que esto sucede se cambia el nivel de resolución de los *buckets*, ofreciendo una visión más detallada adecuada al nivel de zoom aplicado y se actualiza la escala de colores de los mismos. En los **primeros momentos** después de detenerse la animación, la usuaria va a repartir su atención en las tres partes fundamentales de la interfaz, que se ilustra en la figura siguiente:

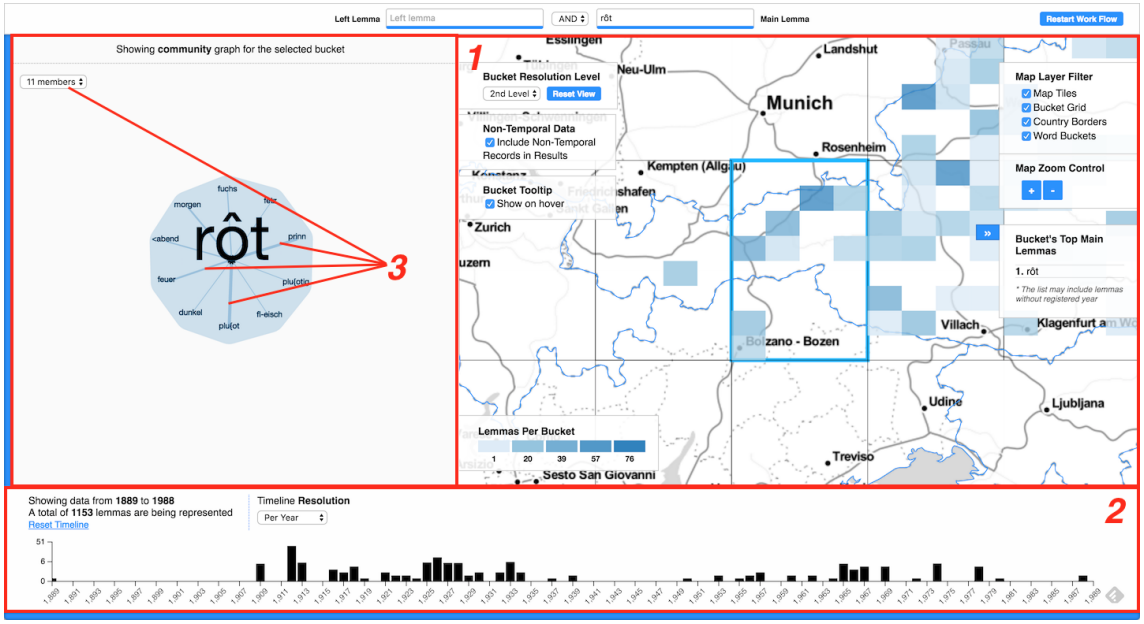


Figura 40: Detalle de la interfaz del prototipo al seleccionar el *bucket* objeto del estudio propuesto. 1)Mapa centrado en las coordenadas centrales del *bucket* mostrando la distribución espacial de sus componentes. 2)Línea temporal actualizada para reflejar la dimensión temporal del subconjunto de datos. 3)Área de SNA mostrando el grafo de relaciones de los lemas resultantes del filtrado espacial.

Ahora la analista puede observar en más detalle la naturaleza espacial de los datos (40.1), a través de una nueva escala que ha reemplazado a la anterior y, de manera análoga a lo visto en el anterior paso, también la repartición de los datos en la línea temporal(40.2).

También entra en juego ahora el **análisis de redes más general** de los dos ofrecidos por el prototipo: El grafo que muestra las diferentes comunidades encontradas en el espacio geográfico seleccionado, da una idea de las relaciones entre los diferentes lemas del subconjunto de datos analizado (40.3). Este grafo ocupa ahora parte de la pantalla que antes era dedicada al mapa, ya que nuestro flujo de trabajo entiende que esta visión es interesante ya a este nivel de detalle más específico y no así en los niveles superiores, donde un grafo hiperpoblado no sería tan útil como para ocupar la parte de la pantalla que se le dedica.

A través de las visualizaciones empleadas, la analista puede constatar ciertos hechos de manera casi inmediata: 1. **La distribución espacial de los término no es equitativa** a lo largo del territorio seleccionado, sino que se concentra en ciertos puntos del mismo. En este punto se podría aplicar en demanda un mayor nivel de

zoom que aclarase, mediante la leyenda, qué localidades aglutinan los resultados e inspeccionar otras características espaciales. 2. **La palabra más importante** del subconjunto de datos es "rôt". Éste hecho se refleja claramente por medio del tamaño de la fuente empleada para representar el lema. Y es más, este lema es la palabra central de la única comunidad encontrada en los datos. es Como no podía ser de otra manera, la visualización refleja el sentido de la búsqueda dirigida efectuada por el usuario y, a pesar de que esta conclusión pudiese parecer demasiado demasiado obvia en este caso de estudio, en otros flujos de trabajo con búsquedas dirigidas más complejas interesa conocer esta información en una primera impresión, como veremos en el siguiente caso de estudio.

Sin embargo, no toda la información codificada en el grafo resulta tan obvia. Algunos detalles visuales hacen que el usuario adquiera, en **segunda instancia**, más conocimiento añadido: 1. Como se puede observar en la Figura 40.3, el grafo marca en su selector de niveles de filtrado situado en la esquina superior izquierda, el tamaño mínimo de elementos de las comunidades que aparecen en él. Por tanto, la usuaria sabe que esta comunidad cuenta con exactamente 11 miembros. 2. También se sabe simplemente con prestar atención al grosor de las aristas del grafo que hay tres palabras más referenciadas que otras en el subconjunto de datos, y que además éstas son las que contienen "plu{ot", "prinn" y "morgen" como prefijos de "rôt". En este momento, y sólo mediante el uso de dos interacciones (una entrada de teclado y un *click*) se habría contestado a la pregunta inicial planteada en el caso de estudio.

Por otro lado, resulta en la mayoría de los casos analizar en más detalle algunas de estas relaciones, en un intento de desvelar otro tipo de patrones más generales hallados en otros flujos de trabajo. Se trabajará ahora en este flujo con el detalle mediante la interacción no ya con conjuntos de lemas, sino con representaciones unitarias de los mismos.

6.1.3. Detalles en demanda después

El estado mental de la analista en este punto refleja el descubrimiento de conocimiento del que no se disponía antes de comenzar el proceso de exploración visual. Este nuevo conocimiento recién adquirido desencadena, luego de haberse interiorizado, una serie de procesos mentales conscientes e inconscientes que pueden resultar en la generación de nuevas conjeturas, conclusiones o hipótesis de todo tipo. En nuestro caso de estudio, se consideró interesante comparar los resultados encontrados en la región seleccionada con los globales de todo el conjunto de datos con el fin de encontrar patrones y coincidencias que pudiesen dar pie a nuevas preguntas de investigación.

Primero, la usuaria va a visualizar las relaciones de la palabra rôt en el gráfico de árbol dedicado que se dedica a la visualización de redes centradas en un lema. Para ello, va a seleccionar la palabra de interés (rôt) y se selecciona la opción correspondiente. De manera análoga a como se dividió inicialmente el espacio disponible para hacer sitio a la vista de SNA, ahora es ésta la que se vuelve a dividir para repartir su espacio entre el grafo de comunidades y el grafo de árbol recién genera-

do. De momento la aplicación mantiene el contexto seleccionado del *bucket* ya que el usuario no ha expresado otra intención (que se haría fehaciente empleando los controles fuera del grafo de comunidades), de manera que la red empleada va a ser la misma en ambas visualizaciones. En esta ocasión, la usuaria ve más claramente la ordenación numérica por número de fuentes encontradas de cada palabra, ya que son presentadas en orden descendente. La misma información que vimos antes vuelve a aparecer: los lemas que habiendo sido combinados por la derecha con "rôt" tienen más ocurrencias en esa región son "plu{ot" (20 referencias), "prinn"(17), feuer (11) y "morgen"(10)... (Ver Figura 41 arriba). La analista puede anotar estas palabras y continuar con el análisis comprobando qué ocurre cuando se coloca "rojo" como prefijo en el mismo subconjunto. Los resultados se muestran en la parte inferior de la Figura 41. Para completar el proceso, se emplea el botón a la derecha de la pantalla en la lista de lemas principales más usados del *bucket*, y se repite el mismo proceso pero en este caso empleando redes generadas a partir de todos los resultados disponibles sin filtrar.

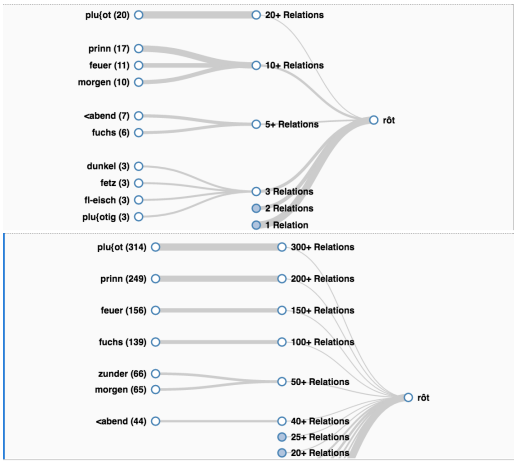


Figura 41: Relaciones por la izquierda del lema "rôt" en el *bucket* seleccionado (arriba) y en todo el conjunto de datos

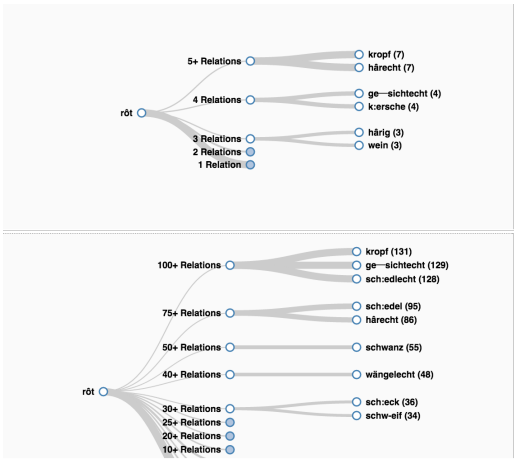


Figura 42: Relaciones por la derecha del lema "rôt" en el *bucket* seleccionado (arriba) y en todo el conjunto de datos

Si se comparan las palabras más usadas tanto en el *bucket* como globalmente, la analista puede observar coincidencias evidentes para ambas posiciones del lema. En el caso izquierdo, los tres lemas más usados aparecen en el mismo orden: "plu{ot" (sangre), prinn (ardiente) y feuer (fuego). También aparecen, en distinto orden pero entre los usos más comunes: "morgen" (mañana), "¡abend" (noche) y fuchs (zorro) por lo que se podría decir que los conjuntos están bastante correlados ya que se ha encontrado un patrón de coincidencia bastante claro. En este caso se concluye la investigación afirmando que los usos como sufijo del lema objeto del análisis son comunes a los de otros dialectos.

Para la situación de prefijo, llama la atención el cuarto lema más usado en el *bucket*: "k:ersche" (cereza), ya que no es posible encontrarlo entre los análogos del conjunto global. Habría que descender mucho más abajo, hasta la posición 21 para encontrarlo. A raíz de este nuevo descubrimiento, la usuaria va a fijar su atención

en este caso particular de combinación de lemas y por medio de la acción de *click* va a lanzar una nueva búsqueda en base a estos términos, comenzando una nueva iteración en el proceso.

6.2. Detección de la homonimia

Continuando con el caso de estudio anterior, se presenta una nueva aplicación del prototipo en la búsqueda de patrones en la formación de nuevas palabras en los diferentes dialectos. Un fenómeno lingüístico que se da con cierta frecuencia y suele ser objeto de estudio en lexicografía dialectal es el de la degeneración fonética de los lemas. Dependiendo del acento que se emplee en un zona de influencia de cierta lengua, una palabra puede ser pronunciada de forma diferente a la original. Con el paso de los años, esta nueva pronunciación puede adquirir tal importancia que el término original es reemplazado y el concepto al que éste se refería es verbalizado mediante una nueva palabra con entidad propia resultante de esta variación fonética. Esta nueva palabra reflejará en su escritura esta nueva pronunciación, que pasará a formar parte del léxico del dialecto hablado en dicha zona.

Existen ciertos métodos computacionales[52] que pueden ayudar a las investigadoras en esta tarea de detectar y estudiar estos accidentes lingüísticos. Éstos, en combinación con los métodos de SNA propuestos en el prototipo van a proporcionar una serie de elementos visuales que acelerarán la detección de patrones típicos de estos fenómenos y acelerarán por tanto la extracción de conocimiento de los datos.

6.2.1. Visión general primero

Estos patrones típicos suelen denotarse por la aparición de términos en los que los lemas original y degenerado aparecen en la misma posición junto a los mismos lemas, generando palabras parecidas en escritura pero con significados iguales. Continuando con el estudio del color rojo, la analista va a comenzar la sesión de estudio mediante una búsqueda dirigida en esta ocasión más compleja que en el caso anterior, que explicamos brevemente antes de pasar al detalle del proceso: Cuando se producen diferencias en la pronunciación de los términos, las palabras sufren variaciones en la escritura para representar este cambio. Sin embargo, como estas nuevas palabras derivan de una diferencia fonética, su escritura no resultará ser muy diferente y por lo tanto, la analista puede lanzar una **búsqueda difusa** en sintaxis *Lucene* que emplee la distancia de Levenshtein para encontrar palabras "parecidas" a la dada. En respuesta a esta petición del usuario, que es realizada empleando los mismos métodos y elementos gráficos que en el caso anterior, el motor de búsquedas va a devolver ocurrencias de lemas que tengan una distancia de Levenshtein de 1 (que necesiten de una edición para ser iguales a la proporcionada) con respecto a la original. En el caso del lema sometido a estudio "rôt", la analista sabe también, que éste aparece con mucha más regularidad en la parte izquierda que en la derecha (ver figuras 41 y 42), así que comenzará lanzando la búsqueda con esta premisa. En las Figuras 43 y 44 presentamos dos detalles del mapa de la interfaz, el primero

capturado al introducir la palabra “rôt” y el segundo instantes después de añadir el carácter “˜” al término (y eliminando el carácter fonético de la letra “o”), expresando su intención de realizar una búsqueda difusa. Además, gracias a la técnica de UX de búsqueda instantánea, la usuaria del sistema va a poder ir observando estos cambios de distribución espacial a la vez que teclea.

Figura 43: Detalle de la distribución espacial de la búsqueda de “rôt”

Figura 44: Distribución espacial de la búsqueda difusa de “rot”

Al ampliar la búsqueda inicial del término para incluir la componente difusa de distancia entre palabras, se pueden apreciar dos hechos importantes en la visualización espacial:

1. Se amplia el número de resultados, que pasa de 638 a 1099.
2. Aunque no se muestre en las imágenes adjuntas, la distribución temporal cubre un período más amplio (aproximadamente 300 años mayor).
3. Existe una zona al norte del país que ha variado considerablemente su tonalidad al incluir el parámetro difuso en la búsqueda. Esto denota que esa zona ha sido especialmente afectada por la introducción de este cambio. La usuaria va a desplazar el cursor sobre esta zona para comprobar este hecho.

6.2.2. Zoom y filtrado

La analista, a vista de estos resultados y especialmente de la última conclusión de las apuntadas, va a aumentar la resolución de los *buckets* manualmente para poder observar en más detalle la distribución espacial de los resultados (Figura 45).

La analista analiza nuevamente la situación y observa la distribución de los datos dentro del *bucket* seleccionado. A pesar de no ser la zona que cuenta con más ocurrencias, a este nivel de detalle sí permite afirmar que la distribución espacial de las mismas es la más amplia de todas. Ésto, ligado al hecho de ser la zona más afectada por la introducción del parámetro difuso hace que la usuaria se decante por centrar su análisis en ella.

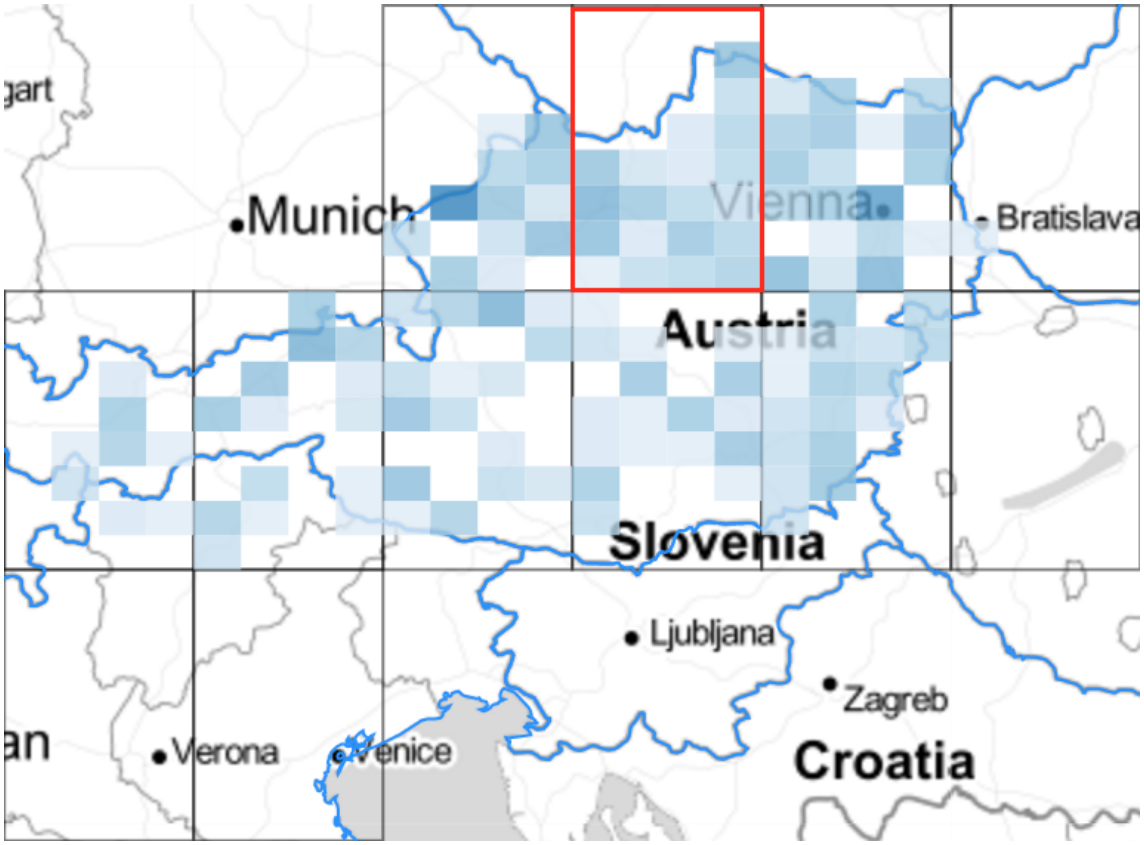


Figura 45: Detalle del mapa mostrando la distribución espacial de los resultados a resolución más alta.

Cuando la animación termina, se ha generado un nuevo grafo que va a permitir realizar un análisis de redes sobre el subconjunto de los datos seleccionado que se muestra en detalle en la Figura 46.

Lo primero que llama la atención del grafo recién generado son las dos grandes comunidades dominadas por los términos "rôt" y "rotz" mucho más resaltados que el resto y que resultan de haber encajado en los términos de la búsqueda difusa (Ambas palabras tienen una distancia de un carácter con "rot"). El resto de nodos más pequeños representan las apariciones de otros lemas en las partes derechas de las mismas. Entre ellos destaca el lema "maulecht" que, a pesar de permanecer a la comunidad de "rôt", también se asocia con "rotz". Esto es indicador para la analista de un hecho curioso, que ha sido detectado gracias a métodos exclusivamente visuales. La analista decide seguir investigando este hecho modificando el selector del nivel de filtro para mostrar comunidades menos pobladas en un intento de encontrar más relaciones con el término "maulecht". El grafo se modificado de acuerdo a esta nueva selección y aparece como se refleja en la Figura 47.

Al aparecer más resultados pertenecientes a comunidades más pequeñas (los términos que encajaron en la búsqueda participan en menos palabras), se ven otros términos centrales, como "r¡ot", que es una pronunciación diferente de la estándar "rot" y participa del término "pelirrojo" ("r¡otkopf) y otras derivaciones sin relación

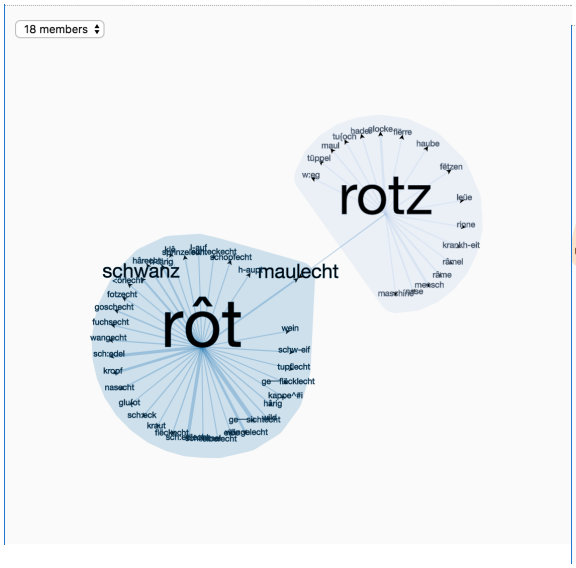


Figura 46: Detalle del grafo filtrando comunidades de menos de 18 elementos.

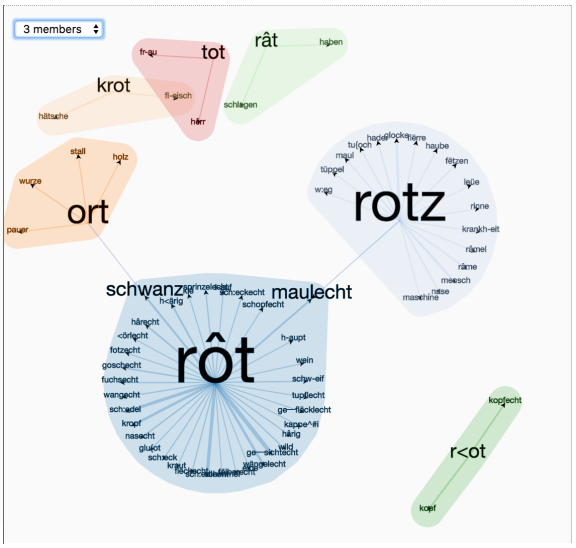


Figura 47: El mismo grafo mostrando ahora comunidades menos pobladas.

con el color rojo.

6.2.3. Detalles en demanda después

Al no encontrar más relaciones, se continua el estudio analizando las palabras "rotzmaulecht" y "rôtmaulecht". Las preguntas que la analista se hace ahora se refieren a las particularidades de ambos términos: ¿Cumplen "rot" y "rotz" la misma función dentro de la palabra? ¿expresan ambos lo mismo? Es por tanto que va a necesitar el detalle de las mismas. Desde el grafo se van a buscar ambos términos, primero uno y luego otro, llegando hasta el registro que los origina. Analizando el registro original TUSTEP, se va a poder comprobar que "rôtmaulech" hace referencia al "sonrojo", mientras que "rotzmaulecht" significa "estirado", en sentido peyorativo. En este caso la analista ha descubierto que, a pesar de existir una relación entre términos parecidos a través de "maulecht", estos no guardan ninguna relación en significado, al menos de manera aparente, dando lugar al fenómeno de la **homofonía**. Si la analista repite la búsqueda, esta vez centrándose en otras zonas geográficas de la Figura 44, puede encontrar el caso contrario (Ver Figura 48).

En el nuevo grafo de comunidades se ha identificado otra pronunciación de "rot", vista en el ejemplo previo también: "r¡ot". Este término se asocia con "kopf", que a su vez se relaciona con "rôt". Si la usuaria repite el proceso y accede a los registros individuales de cada uno, podrá comprobar que "r¡otkopf" se refiere a "Caperucita Roja", mientras que "rôtkopf" hace referencia a la persona pelirroja. Este fenómeno se conoce como **homografía**. En este ejemplo concreto, ambas derivaciones del término de búsqueda textual empleado "rot" mantienen su significado (el color rojo) dentro del concepto al que definen, sin embargo en esta ocasión también derivan en términos con significados diferentes, deduciendo que el lema que experimenta

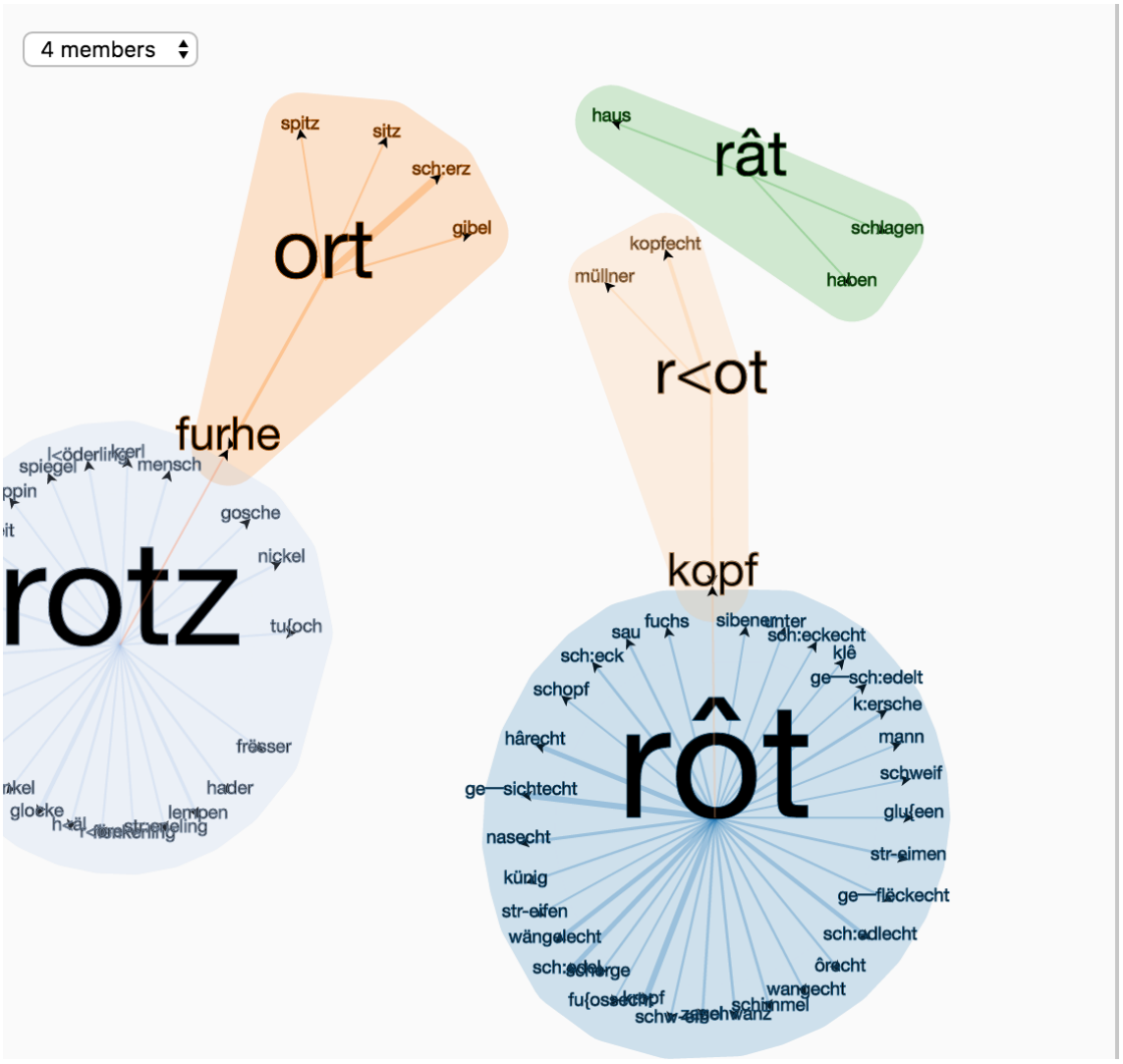


Figura 48: Un grafo en el que dos pronunciaciones de “rot” se asocian con el mismo lema “kopf” produciendo palabras con significados diferentes.

la polisemia es en este caso “kopf”. Un estudio posterior más amplío podría por ejemplo determinar si esta manera de denominar a “Caperucita Roja” es particular de cierto dialecto (y de ahí la variación pronunciación) o si por el contrario ésta es la forma habitual en todos los territorios.

7. Realimentación de expertos

Empleando el flujo de desarrollo del prototipo citado en 4.1.1 se producen, al final de las distintas iteraciones realizadas sobre el mismo, reuniones con el equipo de expertas encargado de validar el prototipo. En las primeras ocasiones, se buscaba orientación sobre la naturaleza de los conjuntos de datos y otros tipos de información interesante que se consideró podían contribuir a la creación de un prototipo más ajustado a las necesidades reales del investigador y por tanto, mejor.

El prototipo presentado en esta memoria es el resultante de la última iteración realizada, en la que al final de la misma se puso a disposición del equipo de expertas para su validación a través de una serie de pruebas y casos de estudio prácticos. Como hemos visto en la sección 6, el prototipo resultó adecuado y consiguió acelerar la extracción de conocimiento y la llegada a conclusiones significativas en comparación a otros métodos de exploración no visuales. Es gracias a este proceso iterativo e incremental de recogida de realimentación y sucesivas implementaciones que ha sido posible llevar a cabo la investigación y el prototipo asociado a ella. El flujo de trabajo propuesto ha probado ser muy efectivo en el ámbito de la colaboración entre expertos en HD y creemos sería muy acertado mantenerlo en futuras posibles colaboraciones que diesen lugar a una evolución del sistema propuesto.

Volviendo sobre el prototipo, éste recibió muy buenas críticas por la visión renovada que ofreció del conjunto de datos en su totalidad. La manera de proyectar los resultados sobre las dos dimensiones principales que maneja el prototipo, tiempo y espacio, fue especialmente útil a la hora de crear un mapa mental nuevo de la naturaleza de los datos. También el equipo de evaluadoras que colaboró en esta investigación hizo hincapié en la responsividad de la interfaz, que consiguió mantenerse en tiempos de interacción en casi la totalidad de las pruebas realizadas.

Por otra parte, uno de los problemas que presentó este prototipo fue la falta de una integración más amplia de todos los flujos de trabajo distintos que manejan en su día a día las investigadoras. A pesar de su validez como herramienta de exploración del diccionario, se detectaron durante las pruebas de validación algunas carencias en las diferentes partes del prototipo que comentamos a continuación:

Los grafos generados por el prototipo están basados en las relaciones existentes entre los diferentes lemas por su posición o aparición en conjunción con otros lemas en términos compuestos. Como se ha explicado, esta decisión se tomó por su idoneidad para crear un número de relaciones suficiente como para afrontar el problema desde un punto de vista visual. A pesar de que este enfoque es correcto, el equipo de expertas nos hizo saber que este tipo de relaciones podrían crearse a partir del **tratamiento de otros campos diferentes** de los registros XML, como el número de cuestionario al que hacen referencia o el autor. De esta manera se podría crear una herramienta que atajase el problema de la exploración de redes en base a más características significativas de los textos y que por tanto sirviese para resolver problemas más complejos.

Además de realizar esta creación de relaciones en base a más campos, se co-

mentó también la idea de crear **grafos más dinámicos**, que fuesen capaces de alterar su estructura en base a los criterios del investigador. Una idea en la que se hizo especial hincapié fue la de colapsar y expandir los nodos de los grafos bajo demanda en base a las comunidades, eligiendo como representante al lema principal o izquierdo más referenciado de las mismas.

En relación a la visión que da el prototipo del conjunto de datos, se criticó la **falta de más métricas globales que facilitasen la creación de búsquedas guiadas** por la aplicación sobre el conjunto de datos. En la actual implementación no se tratan por ejemplo datos estadísticos sobre variables cualitativas (como por ejemplo el tipo de palabra) ni se generan informes sobre cuáles son las palabras más usadas del conjunto de datos. Éstas métricas se dijo, podrían ser el origen de muchos casos de estudio de gran utilidad para la investigación.

La **falta de tratamiento de algunos campos**, como "BD/LT1" (significado) o "NR" (temática del cuestionario / contexto), hace en ocasiones complicado realizar búsquedas temáticas, en las que los resultados estarían relacionados precisamente por estos campos. Ofrecer resúmenes y proyecciones espacio-temporales de los conceptos englobados por alguna de estas variables sería de gran utilidad para el estudio de los diccionarios. Por último, se remarcó la necesidad de ofrecer **sesiones de trabajo** en las cuales la investigadora podría guardar temporalmente resultados interesantes que fuese encontrando a lo largo de las diferentes interacciones del ciclo de trabajo para añadirlos en un paso final a otro tipo de visualización especialmente enfocada a la comparación de registros textuales o a alguna de las ya existentes.

8. Conclusiones y líneas de trabajo futuras

8.1. Líneas de trabajo futuras

La lexicografía es una disciplina del estudio humanístico que presenta una estructura sumamente compleja. Se necesitan años de estudio y preparación, aparte de una tremenda capacidad de esfuerzo y sacrificio, para que un investigador sea capaz de extraer conocimiento científico veraz de las millones de fuentes, referencias, cuestionarios y documentos de los que dispone.

El trabajo descrito en esta memoria intenta proporcionar una metodología de análisis visual que de soporte a las tareas asociadas a la resolución de problemas típicos asociados al estudio de diccionarios históricos. Problemas, por otra parte, que han sido fruto de estudio para reconocidos investigadores en la materia a lo largo de numerosos años y han generado incontables artículos, tesis y trabajos a lo largo de su búsqueda de soluciones. Es por esta razón que este estudio no pretende ser la solución definitiva a estos problemas, sino crear más bien una base sólida de cooperación y buena praxis en el ámbito de la aplicación de técnicas y herramientas computacionales al estudio de la evolución de los dialectos en particular.

A pesar de esto, existen soluciones válidas que no fueron aplicadas en el prototipo final, bien por ser demasiado complejas y escapar de los límites de este estudio, o bien porque el autor no pudo comprender ciertos conceptos suficientemente a tiempo para que entrasen en los tiempos razonables para esta investigación. Estas potenciales soluciones, después de la puesta en común con el equipo de expertos que evaluó este trabajo, desembocaron en una serie de propuestas que recogemos aquí como testigo para próximas investigaciones en la materia.

8.1.1. Tratamiento de la incertidumbre

Una de estas tareas altamente exigentes, y que requerirían de un estudio mucho más largo y elaborado es, sin duda, el **tratamiento visual del concepto de incertidumbre**. Ya no sólo por su complejidad en términos matemáticos y visuales, sino también por las enormes implicaciones que tiene en el estudio humanístico de los diccionarios históricos. La historia es, por definición, borrosa; y es por ello que, como hemos observado en ciertas partes de los datos, existen lagunas en ciertas épocas o en ciertos lugares, o en ciertas épocas en sólo ciertos lugares, o sólo en algunos lugares en ciertas épocas. El trabajo de estos investigadores es, desentrañar los porqués de esas lagunas y rellenar los vacíos de conocimiento con los frutos de sus investigaciones. Sin embargo, no siempre es posible obtener verdades totales sobre hechos concretos, ya sea porque las fuentes disponibles no lo permiten, o porque se necesita desvelar antes otros misterios que aún permanecen ocultos, entre otras muchas razones.

Es por esto que se acostumbra a realizar **aproximaciones**, en las que la dimensión que se intenta acotar no es nítida. Es el caso por ejemplo de los registros en los

que los años de adquisición de la fuente no han podido obtenerse y no se asocian a un año en concreto, sino a períodos más o menos largos (desde dos años consecutivos hasta el siglo). Este tipo de acotaciones no se dan sólo por motivos puramente epistemológicos, sino que también suceden por la propia naturaleza de las fuentes: Éstas, de manera natural, pueden extenderse en el tiempo superando la barrera imaginaria del año. Imaginemos por ejemplo, una fuente literaria. Por supuesto, toda fuente literaria tiene una fecha de publicación, pero ¿qué ocurre si esta información no es la importante en el contexto del problema que se quiere solucionar, sino el período de años que recoge en su contenido? o ¿qué ocurre entonces si la fuente literaria es un volumen de diccionarios publicados entre unos años determinados? No tendría sentido pues, tratar este período de la manera antes propuesta, ya que no es esa la interpretación que se la ha de dar. Obviamente, el concepto de desconocer cuándo se genera una fuente y el de saber exactamente cuándo se genera, que es a lo largo de un período superior al año, son completamente diferentes. Y es por esta causa además que requerirían de un análisis visual distinto, que llevase a conclusiones no relacionadas entre sí. En nuestro enfoque ya tratamos la recogida de estos períodos (sección 5.1.2) para un tratamiento visual posterior en la herramienta que no llegó a materializarse en este estudio.

Para la dimensión espacial de los datos, recordemos, también existen diferentes niveles de incertidumbre: En nuestro estudio consideramos sólo el análisis espacial de los puntos que ofrecen el máximo nivel de concreción: El punto. No obstante existen fuentes en los datos que asocian fuentes a regiones o áreas del espacio más grandes (Ver Figura 18) y que requieren de otro tipo de análisis visual. Si bien es cierto que estas regiones podrían agruparse también por la técnica del *geohash*, no está claro a priori cómo éstas contribuirían al aspecto del *bucket*, o cómo estas zonas se tratarían en una vista general de otro tipo. Es así que no se pudo probar a lo largo de esta investigación que el enfoque propuesto fuese el adecuado para esta tarea y por tanto se descartó la inclusión de estas características.

Existen estudios[57][58] sobre el tratamiento visual de la incertidumbre: tanto la objetiva (entendida como falta de información, la imposibilidad física de apreciar), como la subjetiva (el estado mental del observante que denota falta de confianza en la información que se está recibiendo a través de los sentidos). La línea de trabajo en este campo comprendería plantear el problema de la incertidumbre desde este punto de vista visual, para permitir a la analista identificar patrones también en la ausencia de datos o concreción en alguna de las dimensiones analizada. Esta búsqueda de patrones, podría desembocar en la asociación exitosa de subconjuntos de datos similares en base a alguna otra característica en común, que a su vez pudiera llevar a rellenar estos vacíos de información en los datos. La aplicación de análisis heurísticos que condensasen más conocimiento experto, y la generación previa de estructuras de datos descriptivas de los datos más complejas serían también de ayuda a la hora de crear visualizaciones más efectivas a la hora de resolver este tipo de problemática.

8.1.2. Búsquedas difusas

Directamente relacionado con el concepto de incertidumbre previamente expuesto, están las búsquedas difusas. Una versión simple de las mismas es soportada en el prototipo actual, como se vio en la Figura 33. Por medio de la sintaxis *Lucene* se pudieron encontrar conjuntos de lemas que tuviesen una cierta distancia máxima de Levenshtein entre ellas. No obstante, la representación visual de estas características requiere de un tratamiento mucho más particular. En la implementación actual esta distancia no goza de representación visual de ningún tipo y, aunque existen trabajos de referencia en la materia cuyo estudio y puesta en aplicación serían sin duda interesantes desde el punto de vista gráfico, más pruebas son necesarias para dictaminar si esta característica tiene sentido o no dentro del flujo de trabajo de la usuaria final de la herramienta.

Dentro de la misma temática, la búsqueda dirigida se podría complementar con otros parámetros difusos espaciales o temporales, esto es: lanzar búsquedas dentro de una selección de ciertas regiones, o por cercanía (encontrar resultados que disten entre ellos un máximo y un mínimo definidos por la usuaria). En la dimensión temporal, estas búsquedas funcionarían de manera análoga en base a intervalos temporales difusos: Buscar lemas que disten de uno dado +/-5 años, asignando puntuaciones más bajas a los resultados más lejanos del centroide elegido. Para completar esta funcionalidad, se podría dar a la usuaria también la opción de priorizar los criterios de búsqueda dentro de una escala prefijada: Dar más prioridad a resultados que encajen más exactamente con cierto parámetro, etc. ElasticSearch ya soporta internamente esta funcionalidad que no ha sido explotada en nuestro prototipo propuesto, la cual sería interesante discutir con el equipo de expertos en próximas reuniones.

8.1.3. Soporte de más campos en búsqueda textual

Según la realimentación recibida por el equipo de expertos, la búsqueda textual por más campos (no sólo por lema, sino también por significado, autor o número de cuestionario) se hace necesaria una vez vistos los resultados obtenidos en el actual prototipo. Según lo que se nos explicó, este tipo de búsqueda *multicampo* se acerca más a su modo de trabajo habitual y por tanto es más representativa de la naturaleza de sus investigaciones. Ofrecer un mayor control del que disponen actualmente sobre qué buscar en cada campo, y visualizar correctamente el conjunto de resultados recibido en base al nivel de similitud de cada elemento con los distintos parámetros de búsqueda supondrían el mayor reto en una hipotética implementación de estos flujos de trabajo. El soporte de la búsqueda diacrítica también es importante, ya que algunos lemas contienen símbolos de pronunciación que a menudo hacen imposible el correcto análisis de los mismos según las tesis planteadas en este estudio.

8.1.4. Análisis de redes

8.1.4.1. SNA geográfico

En base a lo observado en otros trabajos mencionados en esta memoria, creemos que sería interesante realizar algún tipo de proyección de las estructuras de análisis de redes en el mapa. Es el caso de Orbis[38], que hace uso extensivo de este tipo de técnicas: Los grafos se proyectan sobre el mapa, permitiendo dar una dimensión geográfica directamente (en nuestro enfoque esto se consigue indirectamente mediante el lanzamiento de nuevas búsquedas desde los elementos de los gráficos de análisis de redes) a las relaciones creadas en los mismos. Los algoritmos de *pathfinding* en este contexto podrían ser de especial utilidad para encontrar distancias entre apariciones del mismo lema en puntos diferentes. Ejecuciones simultáneas de los mismos darían lugar a pistas visuales para encontrar coincidencias entre apariciones de los mismos términos en las mismas épocas o lugares. Si además se combinasen con animaciones en base a la dimensión temporal, servirían a la analista para buscar patrones de propagación de los lemas en base al tiempo y el espacio. La aplicación de algoritmos de inundación podría ser empleada para identificar "caminos" culturales por los que se transmitían los conceptos de unas partes a otras del mapa. La inclusión también de **fronteras geográficas históricas** para contextualizar esta información sería muy probablemente adecuada también. El Figura se muestra un ejemplo de este concepto de inundación de un grafo sobre el mapa, capturado en la aplicación *Orbis*. Como se ve, las aplicaciones son muchas y el desafío está en transmitir estas ideas al especialista en humanidades, que es el que finalmente dará sentido a estas interpretaciones más o menos aventuradas.

8.1.4.2. Clustering

En el grafo dirigido de fuerzas mostrado en el prototipo se aplica, como se comentó en la sección correspondiente, un algoritmo de detección de comunidades. A pesar de que el rendimiento de este algoritmo es muy bueno, con tiempos de ejecución menores al segundo en la mayoría de los casos, se ha detectado la presencia aún de grafos demasiado poblados. Este contratiempo se trató de solventar, como así fue en muchos casos, con un filtrado dinámico que hace que se muestre un nivel aceptable de nodos por defecto. No obstante, en ocasiones ésto no es suficiente y se siguen obteniendo grafos que no permiten la correcta identificación de las partes por estar muy densamente poblados. Una posible solución a este problema es la creación de nodos que agrupen conjuntos de lemas (*clusters*) en base a las comunidades detectadas en la ejecución del algoritmo Louvain. El problema de este enfoque es que la elección del representante del grupo, en situaciones en las que la comunidad es muy grande, puede no ser del todo representativa y en el peor de los casos, puede llegar a ocultar información valiosa al investigador. Esta fue la primera causa por la que esta funcionalidad no fue implementada, y sin duda habrá de ser discutida con el equipo de lexicógrafos en futuras revisiones del prototipo.

Por otra parte, la creación de grafos en base a otras características (relación de

Figura 49: Inundación del grafo con base en Alejandría en la aplicación Orbis. Muestra todas las posibles rutas con origen en dicha ciudad que se podían realizar durante la primavera.

significado, de distancia, de tipo de palabra) también se ha contemplado y se espera recibir más información al respecto cuando el equipo de expertos proceda en su investigación empleando el prototipo actual.

8.1.5. Vista de detalle mejorada

Como se mostró en la secciones introductorias de este trabajo, algunos de los registros TUSTEP empleados como conjunto de datos en la investigación hacen referencia a ediciones digitalizadas de los manuscritos originales de los que provienen. En este aspecto, una conexión de las visualizaciones con este repositorio de imágenes sería de alto valor para contextualizar los datos y ahorrar trabajo al investigador.

8.1.6. Ciencia Ciudadana

Por último, introducir técnicas de Ciencia Ciudadana como la gamificación para conseguir hacer partícipe a la población del proyecto de estudio es algo que es de

interés para el investigador en humanidades. En este tipo de enfoques, las usuarias, desde sus casas, aportan conocimiento experto y valor añadido al proyecto al realizar tareas que no es posible automatizar (como la tarea de la geocodificación histórica descrita en la sección de adquisición de los datos), pero que un humano sí puede desempeñar. Sin duda la plataforma está bien orientada hacia este fin gracias a su estructura orientada hacia la web. Sin embargo existen aún muchos interrogantes sobre cómo ésto se podría llevar a cabo al considerarse una de las últimas etapas del proyecto y encontrarse este aún en sus etapas iniciales.

8.2. Conclusiones

Como se comenzó diciendo al principio de esta memoria, la lexicografía es una rama sumamente complicada de las Humanidades en la que intervienen multitud de disciplinas y factores diferentes: históricos, políticos, sociológicos, antropológicos... Las Ciencias de la Computación tienen también una larga historia en su relación con esta materia y no es probable que esta tendencia termine en un futuro cercano. Existen aún multitud de campos en los que este tipo de colaboraciones entre las dos ciencias son muy escasas y se necesitará del esfuerzo de muchas investigadoras más para cubrir por completo estos campos del conocimiento humano.

En nuestra investigación tratamos por medio de una de estas colaboraciones crear un método y una herramienta de trabajo que consiguiesen aligerar el esfuerzo que tradicionalmente ha demandado el estudio de estas cuestiones. No obstante, la sensación final es que a pesar de los logros conseguidos y aún queda mucho camino por recorrer. Aunque cada día actores de ambas disciplinas encuentran puntos en común sobre los que realizar trabajo significativo, éstas siguen estando bastante alejadas tanto en discurso como en método. El perfil de un investigador en al área de las C. de la Computación sigue asociado con un tipo de mentalidad que dista mucho de aquella típica de un humanista y viceversa, y a pesar de la cada vez más rápida evolución de los avances tecnológicos y los sistemas educativos, los diálogos y el intercambio de ideas entre las partes siguen siendo complicados y en ocasiones incluso infructíferos.

La solución a gran parte de estos problemas, a juicio del autor, pasa por la integración de todo el conjunto de la sociedad en los procesos investigadores. Como hemos visto, algunos investigadores de prestigio en la temática ya han apuntado metodologías y prácticas en Ciencia Ciudadana que consiguen acelerar el proceso de adquisición e intercambio de conocimiento entre las diferentes partes involucradas. En esto juegan un papel especialmente importante dos temáticas también íntimamente interconectadas: La visualización de la información y la educación. En este trabajo se ha tratado de aportar métodos visuales capaces de ayudar al observador en la tarea de la interpretación de grandes conjuntos de datos como los empleados. Es así que la visualización juega un papel importante no sólo en el proceso de extracción de conocimiento centrado en la investigación académica, sino también en la divulgación y transmisión del mismo a todo el conjunto de la sociedad. Como resultado de las metodologías aplicadas, se ha conseguido crear un sistema abierto y

orientado hacia la web que en un futuro podría ser la base de intentos más ambiciosos de implicar a más personas en el estudio de todas las disciplinas humanísticas y científicas de manera conjunta. Este tipo de prácticas sólo pueden resultar en el beneficio de todas las partes implicadas por los motivos ya explicados al inicio de este trabajo: Las Humanidades necesitan de las Ciencias para poder afrontar muchos de los retos que se plantean hoy en día, pero las Ciencias también necesitan de las Humanidades, para mejorar los métodos de enseñanza de la Ingeniería y adaptarlos a las nuevas generaciones, para saber recoger el testigo dejado por las pasadas y darle una dimensión actual con significado y sentido más allá de la simple memorización y repetición de conceptos.

En este trabajó se defendió el papel protagonista de la visualización en la liberación de la potencia de los métodos matemáticos y computacionales de los que se beneficia la lingüística, como los ya mencionados análisis de redes, procesamiento del lenguaje natural y sistemas de información geográfica. En este trabajo se vieron casos reales de puesta en práctica de muchos de los recursos ofrecidos por cada uno de ellos con el objetivo común de desempeñar una tarea específica de investigación bien definida. El cambio de paradigma propuesto con respecto a muchas de las soluciones de visualización de datos preexistentes se ha considerado uno de los logros más importantes de esta investigación. La propuesta arquitectónica sobre la que se asienta ha probado que hoy en día es posible visualizar adecuadamente grandes conjuntos de datos en el navegador empleando tecnologías y estándares web abiertos sin comprometer el nivel de rendimiento al que los usuarios están acostumbrados. A través de una metodología de desarrollo iterativa se pudo crear un sistema que no se desviase de los intereses del perfil investigador al que estaba orientado. Los diversos prototipos realizados consiguieron su propósito de dar una perspectiva diferente pero orientada hacia el mismo objetivo final de los conjuntos de datos, contribuyendo en mayor o menor medida a la consecución de los objetivos propuestos al principio de la investigación. En este proceso, y gracias a las distintas fases de prueba y reuniones con el equipo de expertas que potenció este modelo de desarrollo, se fue adquiriendo progresivamente una idea más detallada de la complejidad de un problema que era en un principio totalmente ajeno al investigador y que redundó en una aplicación mejor orientada de las técnicas computacionales de las que hace gala el sistema final.

Las HD son un campo altamente interesante sobre el que aplicar de maneras novedosas soluciones a problemas clásicos de computación en contextos diferentes a los tradicionales. La necesidad de crear métodos y marcos de trabajo que dirijan adecuadamente esta experimentación se hace por tanto más patente que nunca ante el incipiente crecimiento tecnológico del que es partícipe la sociedad de hoy en día. El esfuerzo conjunto y coordinado de los diferentes grupos de investigación y otros actores involucrados en la puesta en práctica de esta colaboración, así como del resto de la sociedad es, por tanto, imprescindible.Para ello, el empleo de estándares abiertos que potencien el intercambio de datos, así como de nuevas técnicas de visualización preparadas para tratar con grandes flujos de información se convertirá en una necesidad cada vez más obvia en el día de mañana.

A. Notas

Se pone a disposición de los evaluadores una dirección de Internet en la que se puede probar el prototipo propuesto en esta memoria:

`http://exploreat.usal.es/exploreat/ex_tustep_map`

usuario: tester contraseña: beware.garlic.alum

B. Referencias Bibliográficas

Referencias

[1] R. Theron Sanchez and E. Wandl-Vogt, "The fun of exploration: How to access a non-standard language corpus visually," 2014. [*Citado en págs.* VII*, 23, 30, 39 y 57.*]

[2] M. K. Gold, *Debates in the digital humanities*. U of Minnesota Press, 2012. [*Citado en pág. 1.*]

[3] E. Gardiner and R. G. Musto, *The digital humanities: a primer for students and scholars*. Cambridge University Press, 2015, p. 5. [*Citado en pág. 1.*]

[4] E. Meeks, "Digital literacy and digital citizenship," Último acceso: 20-06-2016". [Online]. Available: https://dhs.stanford.edu/algorithmic-literacy/digital-literacy-and-digital-citizenship/ [*Citado en pág. 1.*]

[5] "Sitio web de wikipedia en español para openstreetmap," Último acceso: 20-06-2016. [Online]. Available: https://es.wikipedia.org/wiki/OpenStreetMap [*Citado en pág. 6.*]

[6] "Sitio web de la librería d3.js," Último acceso: 20-06-2016. [Online]. Available: http://d3js.org/ [*Citado en pág. 7.*]

[7] K. Verbert, "On the use of visualization for the digital humanities," Último acceso: 20-06-2016. [Online]. Available: http://dh2015.org/abstracts/xml/VERBERT_Katrien_On_the_Use_of_Visualization_for_t/VERBERT_Katrien_On_the_Use_of_Visualization_for_the_Dig.html [*Citado en pág. 7.*]

[8] "Word clouds en wordle," Último acceso: 20-06-2016. [Online]. Available: http://www.wordle.net/ [*Citado en pág. 8.*]

[9] A. Inselberg, "The plane with parallel coordinates," *The visual computer*, vol. 1, no. 2, pp. 69–91, 1985. [*Citado en págs. 9 y 24.*]

[10] J. D. Thatcher, "Computer animation and improved student comprehension of basic science concepts." [*Citado en pág. 9.*]

[11] S. Berney and M. Betrancourt, "When and why does animation enhance learning: A review," in *Proceedings of the EARLI biennial conference, Amsterdam*, 2009, pp. 25–29. [*Citado en pág. 9.*]

[12] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 13, no. 6, pp. 1224–1231, 2007. [*Citado en pág. 9.*]

[13] J. Unsworth, "What is humanities computing and what is not." [*Citado en pág. 11.*]

[14] R. Busa, "The annals of humanities computing: The index thomisticus," *Computers and the Humanities*, vol. 14, no. 2, pp. 83–90, 1980. [*Citado en pág. 11.*]

[15] R. Wisbey, "The analysis of middle high german texts by computer—some lexicographical aspects," *Transactions of the Philological Society*, vol. 62, no. 1, pp. 28–48, 1963. [*Citado en pág. 12.*]

[16] S. M. Parrish, "Problems in the making of computer concordances," *Studies in Bibliography*, vol. 15, pp. 1–14, 1962. [*Citado en pág. 12.*]

[17] G. Gorcy, "L'informatique et la mise en œuvre du trésor de la langue française (tlf), dictionnaire de la langue du 19e et du 20e siècle (1789–1960)," in *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries: Proceedings of the European Science Foundation Workshop, Pisa*, 1981, pp. 119–44. [*Citado en pág. 12.*]

[18] F. De Tollenaere, "The problem of context in computer-aided lexicography," *Aitken, AJ, Bailey, RW and Hamilton-Smith, N.(eds)*, pp. 25–35, 1973. [*Citado en pág. 12.*]

[19] A. Q. Morton, *The authorship of the Pauline Epistles: a scientific solution.* [Saskatoon, Sask.]: University of Saskatchewan, 1965. [*Citado en pág. 12.*]

[20] "Sitio web de ach," Último acceso: 20-06-2016. [Online]. Available: http://ach.org/ [*Citado en pág. 12.*]

[21] W. Ott, "Strategies and tools for textual scholarship: the tübingen system of text processing programs (tustep)," *Literary and linguistic computing*, vol. 15, no. 1, pp. 93–108, 2000. [*Citado en pág. 13.*]

[22] L. D. Burnard, "Report of workshop on text encoding guidelines," *Literary and Linguistic Computing*, vol. 3, no. 2, pp. 131–133, 1988. [*Citado en pág. 15.*]

[23] R. J. Finneran, *The literary text in the digital age.* University of Michigan Press, 1996. [*Citado en pág. 17.*]

[24] G. Bornstein and T. L. Tinkle, *The iconic page in manuscript, print, and digital culture.* University of Michigan Press, 1998. [*Citado en pág. 17.*]

[25] J. Price-Wilkin, "Using the world-wide web to deliver complex electronic documents: Implications for libraries." *Public Access-Computer Systems Review*, vol. 5, no. 3, 1994. [*Citado en pág. 17.*]

[26] M. Neuman, M. Keeler, C. Kloesel, J. Ransdell, and A. Renear, "The pilot project of the electronic peirce consortium," in *ALLC-ACH*, vol. 92, 1992, pp. 25–27. [*Citado en pág. 17.*]

[27] "The xml leningrad codex," Último acceso: 20-06-2016. [Online]. Available: http://www.leningradensis.org [*Citado en pág. 17.*]

[28] "Infographic: Quantifying digital humanities," Último acceso: 20-06-2016. [Online]. Available: http://blogs.ucl.ac.uk/dh/2012/01/20/infographic-quantifying-digital-humanities/ [*Citado en pág. 18.*]

[29] C. Rohrdantz, "Visual Analytics of Change in Natural Language," Ph.D. dissertation, University of Konstanz, 2014, ph.D. Dissertation. [*Citado en pág. 18.*]

[30] R. Theron and L. Fontanillo, "Diachronic-information visualization in historical dictionaries," *Information Visualization*, vol. 14, no. 2, pp. 111–136, 2015. [*Citado en págs. 20 y 30.*]

[31] T. Mayer, J.-M. List, A. Terhalle, and M. Urban, "An interactive visualization of crosslinguistic colexification patterns," *09: 00–10: 30–Morning Session, Part I 09: 00–09: 10–Introduction 09: 10–09: 40 Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, An Interactive Visualization of Crosslinguistic Colexification Patterns*, vol. 11, no. 15, p. 1. [*Citado en págs. 20 y 69.*]

[32] "Sitio web de wikipedia para el grafo dirigido de fuerzas (en inglés)," Último acceso: 20-06-2016. [Online]. Available: https://en.wikipedia.org/wiki/Force-directed_graph_drawing [*Citado en pág. 20.*]

[33] "Prototipo online de visualización de patrones de colexificación propuesto por mayer et al." Último acceso: 20-06-2016. [Online]. Available: http://clics.lingpy.org/browse.php [*Citado en pág. 21.*]

[34] F. Wanner, W. Jentner, T. Schreck, A. Stoffel, L. Sharalieva, and D. A. Keim, "Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis," *Information Visualization*, vol. 15, no. 1, pp. 75–90, 2016. [*Citado en pág. 21.*]

[35] E. Wandl-Vogt, "Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries," 2010. [*Citado en pág. 24.*]

[36] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on.* IEEE, 1996, pp. 336–343. [*Citado en págs. 24 y 41.*]

[37] E. Meeks, "Digital humanities specialist en stanford.edu," Último acceso: 20-06-2016. [Online]. Available: https://dhs.stanford.edu/ [*Citado en pág. 26.*]

[38] ——, "Sitio web de orbis en stanford.edu," Último acceso: 20-06-2016. [Online]. Available: http://orbis.stanford.edu/ [*Citado en págs. 26 y 91.*]

[39] "Sitio web de transvis. swansea university," Último acceso: 20-06-2016. [Online]. Available: http://othellomap.nand.io/ [*Citado en pág. 27.*]

[40] "Manifest destiny. the story of the us told in 141 maps." Último acceso: 20-06-2016. [Online]. Available: http://michaelporath.com/projects/manifest-destiny/#overview [*Citado en pág. 28.*]

[41] E. Wandl-Vogt and T. Declerck, "Mapping a traditional dialectal dictionary with linked open data," in *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, 2013, pp. 460–471. [*Citado en pág. 29.*]

[42] "Tuebingen system of text processing tools," Último acceso: 20-06-2016. [Online]. Available: http://www.tustep.uni-tuebingen.de/tustep_eng.html [*Citado en pág. 30.*]

[43] "Extensions for spatial data. sitio de web de mysql para desarrolladores," Último acceso: 20-06-2016. [Online]. Available: http://dev.mysql.com/doc/refman/5.7/en/spatial-extensions.html [*Citado en pág. 32.*]

[44] J. Bernard, D. Daberkow, D. Fellner, K. Fischer, O. Koepler, J. Kohlhammer, M. Runnwerth, T. Ruppert, T. Schreck, and I. Sens, "Visinfo: a digital library system for time series research data based on exploratory search—a user-centered design approach," *International Journal on Digital Libraries*, vol. 16, no. 1, pp. 37–59, 2015. [*Citado en pág. 38.*]

[45] "Ranking de los sgbd más usados," Último acceso: 20-06-2016. [Online]. Available: http://db-engines.com/en/ranking [*Citado en pág. 41.*]

[46] T. Hauswedell and M. Wevers, "Reporting the empire: The branding of metropolises and empire in the pall mall gazette 1870-1900." [*Citado en pág. 42.*]

[47] "Lucene: The good parts," Último acceso: 20-06-2016. [Online]. Available: http://blog.parsely.com/post/1691/lucene/ [*Citado en pág. 43.*]

[48] "The future of compass & elasticsearch," Último acceso: 20-06-2016. [Online]. Available: http://thedudeabides.com/articles/the_future_of_compass/ [*Citado en pág. 45.*]

[49] "Elasticsearch: The definitive guide," Último acceso: 20-06-2016. [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/guide/current/index.html [*Citado en pág. 45.*]

[50] "Glosario de elasticsearch 2.x," Último acceso: 20-06-2016. [Online]. Available: https://www.elastic.co/guide/en/elasticsearch/reference/2.x/glossary.html [*Citado en pág. 46.*]

[51] G. Niemeyer, "Geohash," 2008. [*Citado en pág. 58.*]

[52] W. J. Heeringa, "Measuring dialect pronunciation differences using levenshtein distance," Ph.D. dissertation, Citeseer, 2004. [*Citado en págs. 68 y 81.*]

[53] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008. [*Citado en pág. 70.*]

[54] "Código fuente de la implementación original del algoritmo louvain," Último acceso: 20-06-2016. [Online]. Available: https://sites.google.com/site/findcommunities/ [*Citado en pág. 70.*]

[55] G. K. Orman, V. Labatut, and H. Cherifi, "On accuracy of community structure discovery algorithms," *arXiv preprint arXiv:1112.4134*, 2011. [*Citado en pág. 70.*]

[56] "Código fuente javascript del algoritmo louvain," Último acceso: 20-06-2016. [Online]. Available: https://github.com/upphiminn/jLouvain [*Citado en pág. 70.*]

[57] S. BARTHELME, "Visual uncertainty (a bayesian approach)," 2010. [*Citado en pág. 89.*]

[58] A. Dasgupta, M. Chen, and R. Kosara, "Conceptualizing visual uncertainty in parallel coordinates," in *Computer Graphics Forum*, vol. 31, no. 3pt2. Wiley Online Library, 2012, pp. 1015–1024. [*Citado en pág. 89.*]