CONSTRUCCIÓN DE UNA BASE DE DATOS Y UN REPOSITORIO DE DOCUMENTOS DE INVESTIGACIÓN PARA EL PROYECTO TRACE

BUILDING A DATABASE AND A RESEACRH DOCUMENT REPOSITORY FRO PROJECT TRACE

Alejandro Bia Platas
Universidad Miguel Hernández de Elche
abia@umh.es

Jesús Javier Rodríguez Sala
Universidad Miguel Hernández de Elche
jesuja.rodriguez@umh.es

Resumen

El presente trabajo describe el proceso de creación de un archivo digital para investigadores, que comprende una base de datos con datos de investigación de humanidades, y un repositorio de documentos de investigación asociados a estos datos. En este trabajo vamos a discutir las decisiones de diseño que debieron tomarse para evitar o minimizar los problemas y la experiencia que hemos adquirido en el proceso. Uno de los aspectos más interesantes del desarrollo de esta aplicación fue la realimentación cruzada entre los humanistas y los ingenieros de software.

Palabras clave: Bibliotecas digitales. Bases de datos. Repositorios de documentos. Humanidades Digitales. Ingeniería de software.

Abstract

This paper describes the process of creating a digital archive for researchers, comprising a database containing humanities research data, and a repository of research papers associated with this data. In this paper, we will discuss the design decisions that must be taken to avoid or minimize problems and the experience we have gained in the process. One of the most interesting aspects of the development of this application was the cross feedback between humanists and software engineers.

Keywords: Digital Libraries. Databases. Document Repositories. Digital Humanities. Software Engineering.

1. INTRODUCCIÓN[1]

El presente trabajo describe el proceso de creación de un archivo digital para investigadores, que comprende una base de datos con datos de investigación de humanidades, y un repositorio de documentos de investigación asociados a estos datos.

A diferencia de una biblioteca digital convencional, la base de datos de esta aplicación abunda en campos de datos con datos de investigación especializados y detallados. Las relaciones entre las tablas de la base de datos también son complejas, lo que refleja las necesidades de los investigadores y la complejidad inherente a la realidad que estos datos representan.

Los usuarios objeto de esta aplicación son investigadores del proyecto TRACE[2], que trabajan en el análisis contrastivo de corpus bilingües paralelos. Su investigación se centra en varios aspectos de la traducción en el que comparan traducciones entre diferentes idiomas creadas bajo la censura durante la época franquista. Los idiomas incluidos son el español, inglés, francés y el euskera. Además, se tratan varios géneros y tipos de obra, desde narrativa a obras escénicas (teatro, películas y guiones de televisión), y en algunos casos incluso hasta letras de canciones.

En este trabajo vamos a discutir las decisiones de diseño que debieron tomarse para evitar o minimizar estos problemas y la experiencia que hemos adquirido en el proceso. Uno

[1] Este trabajo ha sido desarrollado dentro del proyecto TRACEsofTools (*Herramientas informáticas para análisis contrastivo de textos en corpus bilingües paralelos: alineación y marcado automático, y minería de datos semiestructurados y metadatos*), y financiado con la ayuda FFI2012-39012-c04-02 del VI Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica del Ministerio de Economía y Competitividad de España.

[2] Accesible desde http://trace.unileon.es/.

de los aspectos más interesantes del desarrollo de esta aplicación fue la realimentación cruzada entre los humanistas y los ingenieros de software.

2. LA BASE DE DATOS DE TRACE

Los corpus textuales digitalizados por el proyecto TRACE permiten la aplicación de herramientas avanzadas de análisis textual sistemático. El grupo TRACE-ULE ha desarrollado en proyectos previos algunos programas básicos para limpiar textos escaneados y digitalizados y los programas de alineamiento y conversión de los corpus a bases de datos SQL. La finalidad ahora es dar un salto cualitativo innovador con la integración de todos los subcorpus desarrollados por TRACE con programas de análisis-textual más sofisticados.

Entre los objetivos del proyecto están: la creación, publicación online, ampliación y mantenimiento de la base de datos TRACE-ULE (1939-1985) y del corpus paralelo inglés-español, el desarrollo de las herramientas y el software ya existentes y la creación de nuevas herramientas de análisis textual, así como la creación de un archivo digital de recursos online, que es el tema del presente artículo. El propósito de este archivo digital es desarrollar una plataforma web que soporte tanto las bases de metadatos de TRACE, como el corpus bilingüe para su consulta online.

El archivo digital de TRACE, descrito aquí, es una aplicación orientada a la investigación que tiene que incluir no sólo las obras originales, sino también sus múltiples traducciones a diferentes idiomas.

Debido a la variada naturaleza de las obras (teatro, cine, libros) y el gran número de investigadores, la situación de partida era en cierto modo caótica, ya que cada investigador poseía una base de datos de carácter personal que, a pesar de ser similar a las del resto de los investigadores del grupo, presentaba algunas diferencias derivadas de la naturaleza particular de su propio trabajo. Para dificultar más las cosas, el grupo utilizaba dos tecnologías de bases de datos diferentes: Microsoft Access y FileMaker. Esta situación hizo que la integración de las diferentes bases de datos individuales en una única base de datos integrada para la aplicación (MySQL), fuera una tarea muy difícil.

3. DESARROLLO DE LA APLICACIÓN

La aplicación permite tres tipos de usuarios: no registrados (invitados), que pueden realizar búsquedas; registrados (investigadores), que pueden ver (y eventualmente editar) todos los registros y documentos; y administradores, que pueden crear y gestionar grupos, entre otras tareas críticas. Las Figuras 1 y 2 son muy parecidas, salvo que la primera es para un usuario invitado, no identificado, que sólo puede realizar consultas de búsqueda, además

de identificarse en el sistema (*log in*). Después de identificarse, un usuario registrado podrá ver una serie de otras opciones que aparecen en el menú de la izquierda (Figura 3).

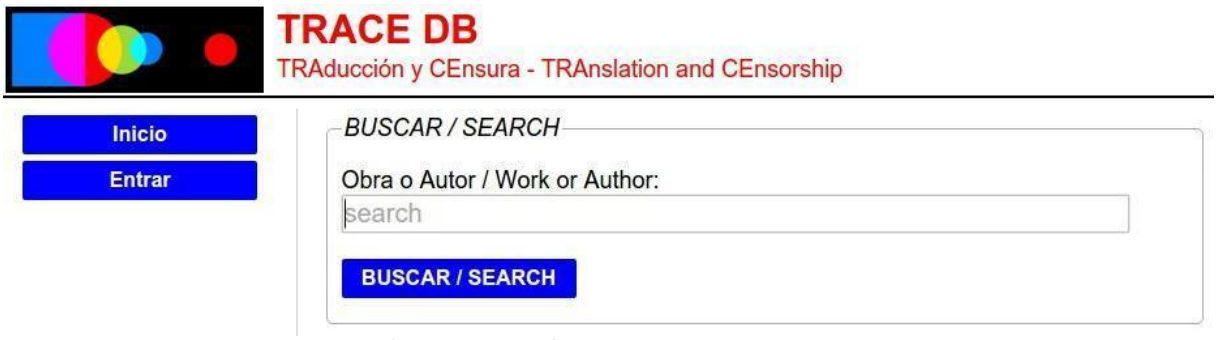

Figura 1. Un usuario invitado (no registrado) sólo puede realizar búsquedas en la base de datos.

Si el usuario registrado tiene además privilegios de administrador, podrá realizar operaciones especiales como la creación de nuevos usuarios o editar las listas de opciones de los paneles de selección utilizados para seleccionar los valores para algunos campos de entrada de datos. Los usuarios pueden ser organizados en grupos para facilitar la gestión de sus niveles de acceso y permisos.

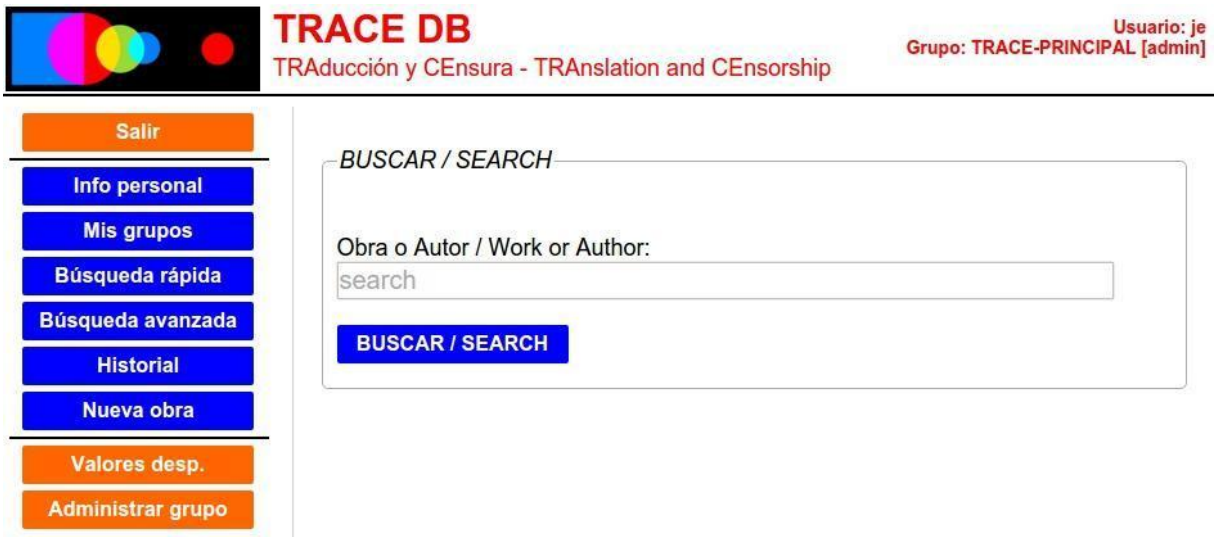

Figura 2. Un usuario registrado puede realizar más operaciones, según el nivel asignado.

La aplicación permite realizar diversas clases de consulta, desde consultas sobre las obras originales, a consultas sobre las obras derivadas (por ejemplo, traducciones). La Figura 3 muestra el formulario de búsqueda avanzada, sólo disponible para usuarios registrados. El usuario puede construir consultas mediante la adición de diferentes criterios de filtrado, además de los más comunes, que se muestran de forma predeterminada. Los resultados pueden ser ordenados por diferentes campos.

Figura 3. Formulario de búsqueda avanzada.

Las Figuras 4 a 8 muestran las diferentes pestañas del formulario de entrada de datos para una obra en particular.

Se pueden ingresar diferentes títulos para una misma obra (incluyendo títulos traducidos): el título real, tal como aparece en el documento original; el título normalizado (decidido por los catalogadores para unificar todas las versiones o derivados de una obra determinada); y el título de la fuente, cuando la obra actual se deriva de otra. La Figura 4 muestra un ejemplo donde un guión de una película, bajo el título en español de *El Cazador implacable* (*The Relentless Hunter*), se corresponde con el título normalizado de *Blade Runner*, la famosa película de 1982 dirigida por Ridley Scott. Esta, a su vez, es una adaptación cinematográfica basada en la novela del año 1968, *Do Androids Dream of Electric Sheep?* de Philip K. Dick (campo *obra original*). Este es un ejemplo interesante para el caso, ya que, para empeorar las cosas, hay otra versión de la película llamada *Blade Runner* (*Director's Cut*), una versión editada especialmente en 1992 para el décimo aniversario de la película. Se utilizó este ejemplo para crear diferentes registros de la novela original, las dos versiones de la película y las diversas traducciones de cada uno de ellos.

Figura 4. Formulario de entrada de datos: información general.

Para simplificar la entrada de datos y ayudar a enlazar correctamente un título a su título normalizado o al título de la obra original, se proporciona una búsqueda rápida por subcadena, en la que escribir unas cuantas letras suele ser suficiente para encontrar el título correcto para un campo determinado.

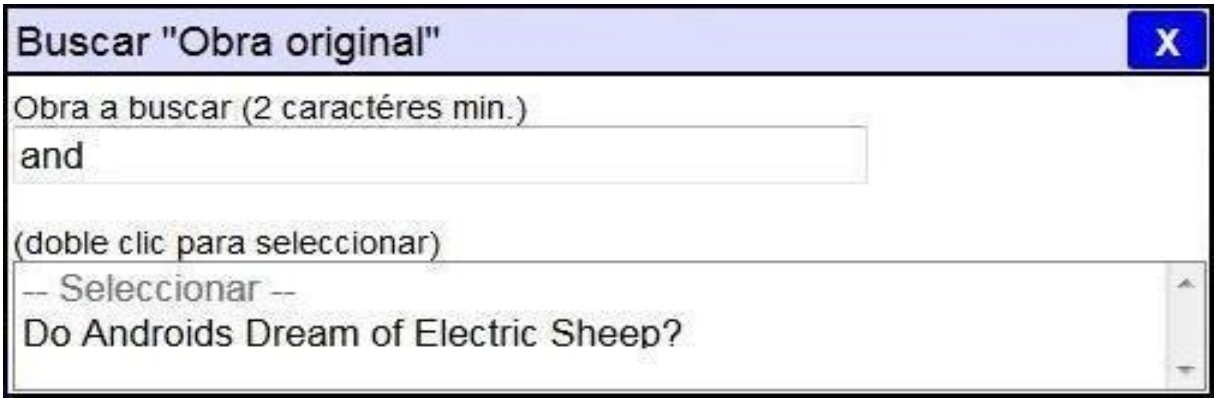

Figura 5. Búsqueda por subcadena para la selección de datos rápida.

Hemos reducido la información redundante en el diseño de la base de datos de la aplicación, mediante la creación de un tipo de registro de usos múltiples que pudiera contener información de los diferentes tipos de obras (libros, guiones escénicos de diferentes clases, etc.). Sin embargo, usando un enfoque adaptativo, las interfaces son diferentes para cada tipo de obra, lo cual permite presentar visualizaciones diferenciadas para cada uno de ellos.

La Figura 6 muestra una pantalla de entrada de datos para una obra escénica, que es diferente de la pantalla utilizada para otros tipos de obra.

Figura 6. Formulario de entrada de datos: información detallada.

Los metadatos de investigación se obtienen de varias fuentes, siendo el Archivo General de la Administración (AGA) la más frecuente. El AGA guarda detalles administrativos e históricos sobre las traducciones, como pueden ser fechas y horas, traductor, valoraciones oficiales y morales, ubicación geográfica (provincia), el nombre del solicitante, correcciones de censura y referencias a otros archivos relacionados.

Figura 7. Formulario de entrada de datos: Información de investigación obtenida del AGA (Archivo General de la Administración).

Una característica de este proyecto es la necesidad de tratar con diferentes roles y responsabilidades relacionadas con los diferentes tipos de obra y géneros literarios.

Dependiendo del tipo de obra, será necesario asignar diferentes responsabilidades (roles): autores, editores, editoriales, directores, actores, productores, letristas, compositores, etc.

El uso de roles definibles, en lugar de un conjunto fijo de campos de roles predefinidos, permite una mayor flexibilidad. De esta manera, el administrador puede añadir nuevos roles a la lista de opciones, siempre que surja la necesidad.

Figura 8. Formulario de entrada de datos: diferentes responsabilidades para diferentes tipos de obra (autor, editor, editorial, productor, director, actor, etc.).

Al igual que en una biblioteca digital, se pueden asociar varios documentos de cualquier formato a las entradas del catálogo. Dichos documentos pueden ser de muy distinta naturaleza, desde la propia obra digitalizada, a cualquier cosa relacionada remotamente con esta: escritos, resúmenes, ediciones paralelas alineadas, archivos del AGA, notas de investigación, facsímiles, biografías, e, incluso, recursos audiovisuales.

Figura 9. Formulario de entrada de datos: documentos asociados.

4. DESARROLLO DEL SOFTWARE

Hemos usado una metodología incremental-prototipada, creando primero una serie de prototipos, para pasar luego al desarrollo definitivo del núcleo de la aplicación, que añade incrementalmente nuevas funcionalidades.

Durante las etapas iniciales del proyecto involucramos a un grupo de estudiantes de ingeniería de software que produjeron varios diagramas de diseño UML, algunos modelos de interfaz (*wireframe mockups*) y desarrollaron nueve prototipos funcionales, tres de los cuales resultaron útiles para probar los requisitos de la aplicación.

El proyecto fue finalmente desarrollado por miembros de nuestro grupo de investigación, que siguieron el popular patrón de arquitectura de software llamado Modelo-Vista-Controlador (MVC), y utilizaron el siguiente software: Apache para el servidor web, PHP para la programación del lado del servidor, Ajax para la programación del lado del cliente, y MySQL para la base de datos, de manera que la aplicación pueda ejecutarse en cualquier plataforma XAMP.

Etapas del desarrollo:

- Captura de requisitos
- Análisis de requisitos y diseño
- Diseño y desarrollo de la plataforma básica
- Construcción base de datos con metadatos de TRACE
- Desarrollo de la Web para Corpus TRACE online

5. CONCLUSIONES

Digna de destacar es la experiencia adquirida en reducir la brecha entre el desarrollo de una biblioteca/archivo digital convencional y una biblioteca con requisitos especializados de investigación en Humanidades.

Uno de los aspectos más interesantes de este desarrollo fue la realimentación cruzada entre los humanistas y los ingenieros de software. En muchos casos hemos tenido que elegir entre lo que los diseñadores de software consideran un diseño de base de datos óptimo y lo que los humanistas necesitan para poder representar su base de conocimientos.

La participación de estudiantes del Grado en Ingeniería Informática fue también una buena idea, aunque el proyecto se retrasó un poco por ello. Fue una interesante experiencia de fusión de la investigación y la enseñanza, donde varios grupos de prácticas de gestión de proyectos compitieron en un entorno creativo para construir los mejores prototipos que han permitido poner a prueba diferentes aspectos del problema.