

Proyecto para generar versiones digitales de obras en papel

Alonso Quijano

31-2-2020

0.1 Descripción

Generaremos unas primeras (es decir, susceptible de correcciones) versiones digitales de la colección bibliográfica del siglo XVIII titulada “Locuras barrocas” (que están en dominio público) Mediante un sistema OCR iremos generando bloques de texto que luego procesaremos con unos programas que vamos a hacer, en lenguajes que ya veremos, pero que serán habituales, para obtener una versión digital.

Incluiremos toda la documentación intermedia, desde la posición de cada carácter, hasta la versión revisada por una persona, indicando qué correcciones ha hecho. Los revisores van a ser unos voluntarios que sacaremos del alumnado de nuestros cursos, puesto que de esta forma podrán compensar notas. Seguro que firmarán la licencia para hacer público su trabajo.

Los datos en bruto generados por el software OCR estarán en formato XML, correspondientes a fotos de las páginas en alta resolución, en formatos habituales; estas imágenes serán propiedad del editor. A lo largo de las distintas fases de proceso con nuestros programas, usaremos formatos propios que documentaremos. El texto final usará la codificación Unicode y un marcado estructural que sea acorde con las directrices indicadas por la Federación de Bibliotecas Digitales.

Seguramente tendremos que mejorar los algoritmos OCR existentes para adaptarlos a las tipografías dieciochescas.

Pensamos liberar software y datos de forma libre usando licencias tipo Creative Commons, la mínima. Estará libre para descarga en un plazo máximo de seis meses después de finalizado el proyecto.

El almacenamiento será en un servidor (con alimentación redundante y seguridad física) gestionado por el Servicio de Informática. Hacen copias

de seguridad incrementales. Esta gente nunca ha fallado con esto, pero de todas formas vamos a añadir un servicio externo de copia en cinta. Nuestros cálculos son que entre todo, no llegaremos al TeraByte.

El acceso será con usuario y contraseña. Pondremos especial cuidado en revisar los registros del servidor, para ver si alguien se cuela.

Tres años después de terminado el trabajo, se lo pasaremos todo a la Biblioteca Universitaria, y lo borraremos de nuestro servidor.