El proyecto METAe (Meta-data Engine Project): concepto, implementación e integración en bibliotecas digitales

E. Sánchez-Villamil[1], J.M. Iñesta[2], R. C. Carrasco[2], G. Mühlberger[3]

[1] Biblioteca Virtual Miguel de Cervantes
Universidad de Alicante
enrique.sanchez@cervantesvirtual.com
[2] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{inesta,carrasco}@dlsi.ua.es
[3] University Innsbruck Library
Innrain, 52 – 6020 Innsbruck, Austria
guenter.muehlberger@psb1.uibk.ac.at

Resumen. La necesidad de digitalizar documentos impresos requiere la creación de herramientas y estándares que ayuden en esta tarea. El consorcio internacional del proyecto METAe del V programa marco de la Unión Europea ha desarrollado la herramienta Metadata Engine que aporta una solución integrada de digitalización, OCR, y etiquetado XML de textos. Mediante una sencilla interfaz, el usuario gestiona la digitalización, la extracción del formato y el preprocesado del documento impreso, la ejecución del OCR y la extracción del contenido semántico. Todo ello con una mínima supervisión del usuario. La definición de los estándares METS (Metadata Encoding & Transmission Standard) y ALTO (Analyzed Layout and Text Object) para el XML aportan un esquema de etiquetado flexible capaz de generar facsímiles con los que reconstruir el aspecto original a partir de la información almacenada. En este artículo explicamos el funcionamiento de esta herramienta, los estándares utilizados en los documentos XML generados, y por último la integración de la herramienta en una biblioteca digital.

1 Introduccción

El proyecto METADATA ENGINE (METAe) [1] ha desarrollado un software integrado que automatiza en gran medida la obtención de metadatos al aplicar las nuevas tecnologías para el reconocimiento de caracteres, estructuras y documentos, así como la conversión de la información capturada en documentos XML [2]. Estos ficheros XML sirven como base a una gran variedad de aplicaciones (como nuevos buscadores de texto XML [3]), herramientas de navegación, libros electrónicos, libros con sonido digital, y para la producción automática de ficheros HTML, XHTML, PDF o PS.

El proyecto METAe responde a la necesidad de generación automática de metadatos durante la conversión de documentos impresos en la digitalización a gran escala de mate-

rial impreso. La naturaleza europea del proyecto ha hecho que haya sido diseñado para funcionar en seis idiomas correspondientes a socios del proyecto: alemán, inglés, francés, español, italiano y noruego.

El principal objetivo del proyecto METAe es la conversión digital de material impreso, como libros y periódicos, para conseguir una mayor fiabilidad en términos de preservación digital, con una mejor relación eficacia/coste en términos de automatización y una mayor amabilidad y accesibilidad para el usuario. El proyecto tiene los siguientes objetivos generales:

- Elevar la conciencia sobre la necesidad de estrategias de preservación digital mediante generación semiautomática de metadatos, aumentando de esta forma la eficiencia de la digitalización de material impreso, en términos de costes y recursos necesarios.
- Desarrollar software que mejore y automatice en la medida de lo posible la creación de colecciones digitales de material impreso.
- Destacar la necesidad de la preservación digital con la recopilación de metadatos administrativos y descriptivos durante el proceso de conversión.
- Integrar y mantenerse dentro de estándares como DC, RDF, EAD, XML y TEI [4,5].
- Enriquecer la conversión digital mediante tecnologías como el análisis de la distribución de las páginas y la clasificación de documentos.
- Apoyar aplicaciones basadas en la salida del programa METAe tales como buscado res XML, herramientas de navegación y libros electrónicos.
- Aumentar el área de aplicación de los sistemas de OCR *(optical caracter recognition)* para procesar tipos de letra antiguos, como el gótico, y libros europeos antiguos (siglo XIX y principios del XX).
- Obedecer las pautas de un "diseño-para-todos" de forma que se garantice que las personas con discapacidad visual se beneficien también de la conversión digital de material impreso.

Para conseguir estos objetivos el proyecto METAe:

- introduce herramientas de análisis de documentos y su distribución para utilizarse como tecnología clave en el futuro del software de digitalización.
- desarrolla herramientas de conversión y captura para el almacenamiento automático de metadatos administrativos y descriptivos.
- desarrolla un motor OCR general, pero especializado en procesar tipos de letra centroeuropeos del siglo XIX.
- Sigue los estándares en el campo de la preservación, como el XML o el TEI.

Estos objetivos se han materializado en el programa METAe, un paquete integrado para la generación automática de metadatos descriptivos, administrativos y técnicos, durante el proceso de conversión digital y la construcción de los documentos XML. Este programa integra un OCR especializado en tipos de letra europeos del siglo XIX, en especial el Fraktur centroeuropeo, aunque también cubre tipos más modernos.

Este artículo se estructura de la siguiente manera: primero se describirá la aplicación desde un punto de vista global, luego se abordará cada una de las fases desde la digitalización hasta la generación de los ficheros XML y finalmente se discutirán sus ventajas y posibles inconvenientes, a modo de conclusiones.

2 La aplicación informática "Meta-data engine"

Esta aplicación permite la creación de colecciones digitales (facsímiles electrónicos y XML etiquetado) o material impreso mediante el escaneado, la generación altamente automatizada de metadatos y un procesamiento OCR mejorado.

Para alcanzar estos objetivos ha sido necesario investigar y crear un conjunto de reglas que definen el ámbito de los documentos que pueden ser procesados por el METAe de un modo satisfactorio. Asimismo, ha sido necesario investigar la conveniencia de los estándares actuales para los metadatos administrativos, tales como el EAD, DC y RDF. Desde el punto de vista de los formatos de salida del programa, se han desarrollado DTDs de XML para describir del modo más apropiado la distribución y la estructura de los documentos impresos (que están en el ámbito del proyecto), creando una base de conocimiento que provea un conjunto de reglas para la descripción apropiada de las diferentes clases de libros, los tipos de páginas, los tipos de elementos y el vocabulario básico. Por ejemplo: La tabla de contenidos se llama en la mayoría de los casos "índice", pero no es la única denominación y el conjunto de palabras posibles para reconocer esta tabla es distinto para cada idioma.

Para ayudar en la precisión de la extracción de la estructura y de los metadatos, ha sido necesario desarrollar un módulo de análisis de distribución capaz de analizar y clasificar elementos de página, como los números de página, cabeceras, notas al pie, imágenes, frases resaltadas o separadores gráficos. También se ha desarrollado un módulo de aprendizaje capaz de integrar correcciones de los usuarios para mejorar la precisión de los algoritmos de clasificación.

El OCR que forma parte del programa consigue un reconocimiento preciso, con especial interés en tipos de letra utilizados durante el siglo XIX, hasta el primer tercio del siglo XX. Esto se concreta en el caso de los países centroeuropeos en el tipo gótico o Fraktur (ver Fig. 1). La precisión en el reconocimiento se ha basado en el uso de modelos del lenguaje y diccionarios históricos de las seis lenguas cubiertas en el proyecto. En el caso del español se han utilizado los diccionarios de la Real Academia Española en sus ediciones anteriores a 1930.

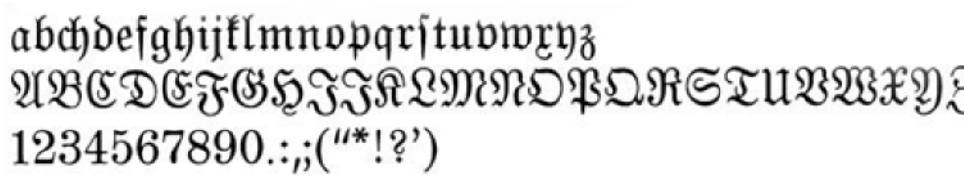

Fig. 1. Ilustración del tipo de letra Fraktur o gótica para la cual se ha desarrollado un OCR específico en el contexto del proyecto METAe

3 Fases en la generación de texto XML

La generación de texto XML conlleva una serie de fases de trabajo que requieren mayor o menor intervención por parte del usuario, pero normalmente esta intervención es simplemente de supervisión, para corregir las imprecisiones de la detección del texto, y de la estructuración del mismo.

A continuación veremos una por una todas las fases y en qué consisten.

1. Escaneado y datos administrativos. En esta primera fase se llevan a cabo dos tareas: por un lado se establecen los datos generales de la obra como son el título, tipo (monográfica o serial), el idioma, la fecha de publicación, etc., y por otro lado se escanean las páginas de las obras o se indican los ficheros de imágenes si ya se han escaneado previamente (ver Fig. 2).

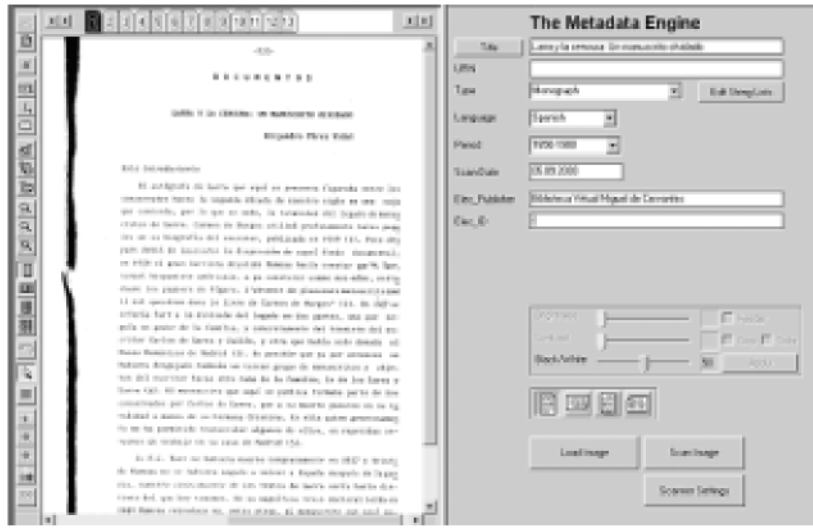

Fig. 2. Ventana inicial del procesado de una obra tras la digitalización. A la izquierda se observan las páginas de la obra y a la derecha se introducen los datos administrativos

2. Detección de marcos. Esta fase es automática y consiste en detectar dónde se encuentra el marco de texto (zona impresa) dentro de la imagen. El software genera un marco con la orientación en la que se estima que se encuentra la página del libro (ver Fig. 3). A continuación el usuario deberá verificar que el reconocimiento haya sido satisfactorio y podrá editar el resultado para precisar mejor los marcos.

Fig. 3. Detección de marcos con las zonas impresas en cada una de las páginas digitalizadas

3. Recorte y detección de bloques de texto y contenidos. En esta fase se recortan los marcos para formar con cada uno de ellos las páginas sobre las que se operará en las fases posteriores. Una vez se tienen las páginas se busca en ellas donde están el texto y las imágenes y se pasa el OCR sólo sobre las zonas de texto. Esta fase también es totalmente automática y solo requiere la supervisión por parte del usuario, que consistirá en ver si los bloques detectados son correctos o si por el contrario existen bloques que hay que juntar o partir (ver Fig. 4).

Fig. 4. Recorte y detección de bloques de texto y contenidos. A la izquierda se observa el árbol jerárquico que las estructuras detectadas y a la izquierda los bloques de títulos y párrafos

4. Detección de números de página. En esta fase se analiza el contenido de los bloques de texto para ver si alguno contiene el número de página. Una vez procesadas todas las páginas se da al usuario la posibilidad de renumerar alguna página si fuera necesario. El programa también indica si ha tenido dudas en la numeración de una página o si ha sido imposible localizar un número, para que el usuario lo introduzca (ver Fig. 5).

Fig. 5. Detección del número de página. A la izquierda se puede contemplar la numeración extraída de las páginas del documento y los espacios para su posible modificación

5. Construcción de la jerarquía básica del documento. En esta fase se identifican los contenidos semánticos de los bloques, y se construye la jerarquía de la página, formada en un primer nivel por la cabecera, el cuerpo y los anexos del documento. A su vez, esta estructura principal se subdivide en bloques más pequeños (párrafos, cabeceras, imágenes, pies de página, etc.) en función del nivel principal en el que se encuentre cada una de esas estructuras. Este es un proceso jerárquico que está sustentado por una gramática que ha sido desarrollada a lo largo del proyecto [6] a partir del análisis de la estructura de cientos de libros y revistas.

En esta fase conviene supervisar que a los bloques se les haya asignado la categoría correspondiente, de entre cabecera, línea de título, nota al pie, etc. (ver Fig. 6).

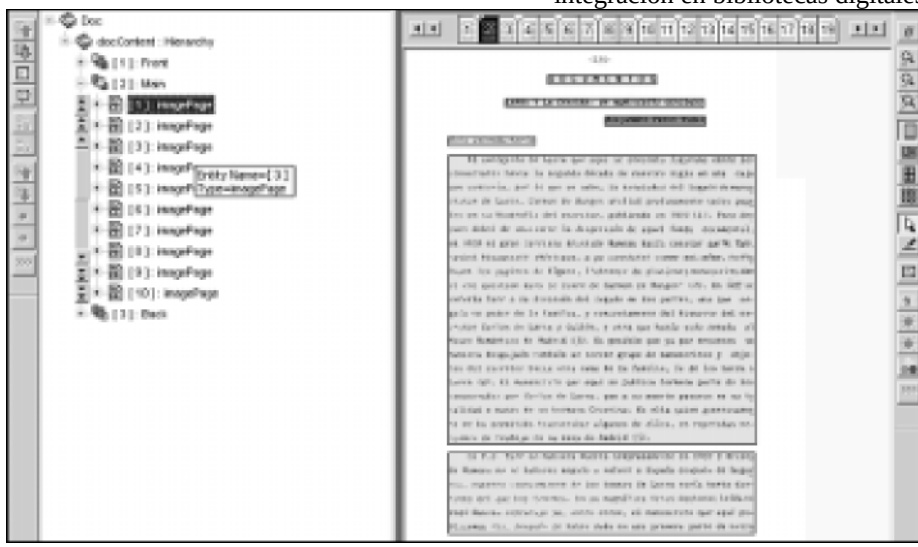

Fig. 6. Construcción de la jerarquía de los bloques de cada página y del documento completo

6. Verificación de la estructura. En esta fase el usuario comprueba que la estructura del documento se haya establecido correctamente. Deben comprobarse la subdivisión en capítulos y el texto generado por el OCR, aunque las palabras dudosas están marcadas para una mayor comodidad del usuario (ver Fig. 7).

Fig. 7. Verificación de la estructura del documento. A la derecha vemos un bloque de texto: arriba su trascripción por parte del OCR y abajo la imagen original para poder contrastar y corregir, si procede, las palabras que el programa ha marcado como dudosas

7. Generación del fichero XML. Una vez seguidos todos los pasos, ya solo resta la generación automática del código XML para todas las páginas de la obra completa digitalizada. Este texto se divide en dos tipos de ficheros XML, un único fichero METS [7] y una lista de ficheros ALTO (Analyzed Layout and Text Object).

El fichero METS (ver Fig. 8) contiene los datos generales y la estructura de la obra, junto con referencias a los ficheros ALTO referentes a todas las páginas de la misma, en los que se encuentran las posibles imágenes que se hubieran encontrado.

```
<?xml version="1.0" encoding="UTF-8" ?>
 <mets xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
   xmlns="http://www.loc.gov/METS/"xsi:schemaLocation="http:
   //www.loc.gov/standards/mets/ http://www.loc.gov/standards/mets/mets.xsd"
   xmlns:dc="http://purl.org/dc/elements/1.0/"   xmlns:dig35="http://www.digitalimaging.o
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:xlink="http://www.w3.org/TR/xlink"TYPE="METAe_Monograph">
 <metsHdr CREATEDATE="2002-10-07T11:28:10"LASTMODDATE="2002-10-07T11:28:10" />


 <rdf:RDF>
    <rdf:Description>
       <dc:title>SUR LA GÉOGRAPHIE DES PLANTES.
     ESSAI</dc:title>
       <dc:language>FR</dc:language>
    </rdf:Description>
 </rdf:RDF>
```

Fig. 8. Fragmento de un ejemplo de fichero METS

Algunos de los datos generales que se almacenan en el fichero METS son, como vemos en el ejemplo, el título de la obra, el idioma, la fecha de creación, el tipo de obra, etc. Pero no solo este tipo de datos, sino también otros como características del escaneado de las obras como el modelo de escáner, y la orientación, tamaño y resolución de las páginas.

Por otra parte, cada fichero ALTO contiene la distribución de elementos dentro una página y marca, para cada elemento, su posición horizontal y vertical (HPOS, VPOS) en la página y sus dimensiones (WIDTH, HEIGHT), así como atributos propios de cada elemento, como su contenido. Los elementos que nos encontramos son, por ejemplo, márgenes, imágenes, bloques de texto, líneas de texto, cadenas de texto (cada una de las palabras).

```
<Layout>
  <PageID="PAGE1"PHYSICAL_IMG_NR="1"HEIGHT="3508"WIDTH="2592">
   <TopMargin ID="P1_TM00001"HPOS="0"VPOS="0"WIDTH="2592"HEIGHT="381" />
    <InnerMargin ID="P1_IM00001"HPOS="2180"VPOS="381"WIDTH="412"HEIGHT="2516" />
    <OuterMargin ID="P1_OM00001"HPOS="0"VPOS="381"WIDTH="241"HEIGHT="2516" />
    <BottomMargin ID="P1_BM00001"HPOS="0"VPOS="2897"WIDTH="2592"HEIGHT="611" />
    <PrintSpaceID="P1_PS00001"HPOS="241"VPOS="381"WIDTH="1939"HEIGHT="2516">
    <TextBlockID="P1_TB00001"HPOS="873"VPOS="381"WIDTH="683"HEIGHT="140"STYLEREFS="PAR_LEFT">
      <TextLineID="P1_TL00001"HPOS="738"VPOS="321"WIDTH="576"HEIGHT="119">
        <String ID="P1_ST00001"HPOS="738"VPOS="321"WIDTH="576"HEIGHT="119"CONTENT="POEMA" />
      </TextLine>
    </TextBlock>
</Layout>
```

Fig. 9. Fragmento de un ejemplo de fichero ALTO

En los ficheros ALTO (ver Fig. 9) se describen los bloques de texto con los metadatos extraídos por el programa. La estructura arborestente del XML permite representar claramente qué elementos se encuentran dentro de otros como los bloques de texto, que se componen de líneas de texto que, a su vez, están formadas por cadenas de texto. En la

figura 9 vemos la información generada para representar adecuadamente la palabra “poema”, que ha sido encontrada en solitario en un bloque de texto.

4 Discusión y conclusiones

El proyecto METAe ha producido una herramienta capaz de producir colecciones digitales. Diversas bibliotecas importantes del consorcio europeo[1] que ha dado como fruto esta herramienta han dedicado tiempo y esfuerzo a obtener una herramienta multilingüe que solucione sus necesidades en este campo.

La automatización del proceso de digitalización de una obra permite reducir enormemente el coste del proceso. En la actualidad, la versión en español del software METAe está siendo evaluada en la biblioteca virtual Miguel de Cervantes [8]. Fruto de esta evaluación, en el contexto de la experiencia adquirida durante estos años de funcionamiento de la biblioteca, se ha derivado una serie de consideraciones que es preciso tener en cuenta, como las que se enuncian a continuación.

El programa METAe ofrece a las bibliotecas la posibilidad de crear colecciones digitales de una forma eficiente en costes y recursos. La clave es que con este software se simplifica mucho el proceso de digitalización y extracción de metadatos. Las entradas y salidas se basan en estándares comúnmente aceptados.

La supervisión por expertos es aún capaz de asegurar una mayor calidad en la trascripción de un texto que la reproducción automática. Por ello, en los casos en los que dicha calidad es esencial, no podremos prescindir totalmente de la corrección manual. Por contra, en aquellos documentos donde la preservación del aspecto sea lo más importante (por ejemplo, porque contienen ilustraciones), METAe permite aumentar la capacidad de preservación.

No toda la metainformación que contiene un texto es sencilla de extraer automáticamente. Por ejemplo, aunque cualquier lector es capaz de identificar los personajes que intervienen en un fragmento de una obra, los sistemas informáticos no son capaces todavía de utilizar información contextual. Por tanto, el operador no se puede evitar, pues asume un papel de control de calidad y corrección de las posibles inexactitudes que se generen.

Dadas estas limitaciones, conviene perseverar en la investigación y el desarrollo de sistemas de ayuda a los procesos de trascripción y de marcado.

Por otra parte, la integración de METAe en un proceso de producción establece un sistema mínimo de control de la producción. Esto es importante en aquellos proyectos que no han diseñado o adoptado uno propio. En el caso de proyectos con un sistema propio deben valorarse las ventajas e inconvenientes del cambio o de la integración de METAe mediante la adecuada reingeniería de procesos.

[1] Los integrantes del consorcio METAe del V Programa Marco de la Unión Europea son: University Innsbruck (coordinator), Austria; University of Linz, Department for Applied Informatics, Austria; Mitcom (Abbyy Europe) Neue Medien GmbH, Germany; CCS Compact Computer Systeme, Germany; Universidad de Alicante, Spain; Friedrich-Ebert Foundation, Germany; Cornell University Library. Department of Preservation and Conservation, USA; Bibliothhque Nationale de France; The National Library of Norway, Rana division, Norway; Biblioteca Statale A. Baldini, Italy; Dipartimento di Sistemi e Informatica, University of Florence, Italy; University Graz Library, Austria; Scuola Normale Superiore, Centro di Ricerche Informatiche per i Beni Culturali, Italy; Higher Education digitisation Service HEDS, UK.

Agradecimientos

Este trabajo ha sido financiado por el proyecto METAe del V Programa Marco de la Unión Europea, referencia: IST-1999-20021.

Referencias

1. G. Mühlberger, "Automated digitization of printed material for everyone: the metadata engine project". *RLG Diginews,* vol. 6**,** núm. 3, junio 15, 2002.
2. XML recommendation. http://www.w3.org/XML
3. Buscador para texto estructurado. http://cervantesvirtual.com/herramientas/textos
4. Dublin Core. http://dublincore.org/
5. Text Encoding Initiative. http://www.tei-c.org
6. B. Stehno y G. Retti, "Modelling the logical structure of books and journals using augmented transition network grammars". *Jorunal of Documentation,* vol. 59**,** núm. 1, pp. 69-83.
7. METS. Metadata Encoding & Transmission Standard. http://www.loc.gov/standards/mets
8. Biblioteca virtual Miguel de Cervantes. http://cervantesvirtual.com