Procesamiento digital de documentos manuscritos históricos degradados.

Osvaldo Clúa, María Feldgen

Departamento de Computación/ Facultad de Ingeniería /
Universidad de Buenos Aires
Paseo Colón 850, 4to Piso
{oclua,mfeldgen}ieee.org

Resumen

El objetivo de este proyecto es investigar y desarrollar algoritmos, modelos y metodologías de trabajo con las herramientas asociadas para proveer información generada automáticamente a partir de las imágenes y las otras fuentes de descripción para crear los metadatos descriptivos y un acceso estructurado al conjunto de documentos del acervo histórico perteneciente a la UBA y otras instituciones nacionales depositarias. El resultado serán modelos, métodos y recomendaciones, prototipos de extracción semiautomática de información y visualización de documentos y libros antiguos de los corpus bajo estudio. La información de estas colecciones estará codificada usando estándares tal de permitir el acceso e intercambio entre repositorios sobre un sistema de biblioteca digital propio o estándar (software libre o propietario), usando los mejores métodos y técnicas de investigación que se sugieren en el campo de la preservación digital.

Palabras clave: Digitalización, manuscritos históricos, indexación de contexto, archivo y recuperación de documentos históricos.

Contexto

Los manuscritos atesorados en bibliotecas y museos constituyen una importante fuente de conocimiento e investigación para los historiadores y el público en general. Para preservar su integridad física y proveer acceso a una audiencia mayor, muchos de estos manuscritos están siendo digitalizados. Los catálogos digitales deben proveer descripciones, metadatos e índices de acceso a estos documentos que contengan información de calidad, que permita cruzamiento entre colecciones relacionadas e interoperabilidad.

A partir de un convenio firmado entre la Facultad de Ingeniería y el Instituto de Historia Argentina y Americana Dr Emilio Ravignani se hicieron algunos ensayos en procesamiento y estructuración digital de manuscritos históricos generando prototipos [Feldgen 2002a], [Feldgen 2002b], de sus colecciones de manuscritos (documentos, periódicos y libros). A partir de esta experiencia se presentó y aprobó el PME-62 (“Proyectos de Modernización de Equipamiento de Laboratorios de Investigación”) que permitió mejorar el equipamiento y dio origen al proyecto UBACYT 2008-2010 (I008) “Procesamiento por computadora de imágenes de documentos históricos degradados”, cuyos resultados aparecen resumidos en el apartado de antecedentes.

El proyecto actual forma parte de la Programación científica de la UBA 2012-2015 y ha sido acreditado y financiado por dicha institución.

Introducción

Los proyectos de preservación y acceso al patrimonio cultural en las distintas partes del mundo usan diferentes estándares internacionales para crear un marco normalizado haciendo posible la integración de datos de distintas procedencias y en distintos formatos. Los estándares proporcionan un conjunto de posibilidades de codificación para la catalogación y descripción de los manuscritos e impresos antiguos, sus metadatos, transcripción y

edición de textos, las anotaciones sobre el manuscrito o asociados al mismo y los descriptores de las características de la edición digital. Conversiones entre formatos de codificación y la tecnología WEB hacen posible acceder al patrimonio cultural de las instituciones en todo el mundo sin comprometer su complejidad o diversidad.

Sin embargo, no hay consenso en el uso de un único estándar para la codificación de la información descriptiva y metadatos. Por ejemplo, el proyecto ENRICH determina el uso del estándar TEI-P5. World Digital Library admite colecciones con metadatos en MARC, MODS o Dublin Core. El proyecto Greenstone usa Dublin Core, RFC 1807 y formatos propios de Oceanía (NZGLS (New Zealand Government Locator Service) y AGLS (Australian Government Locator Service) con conversores a MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace. Library of Congress de EEUU usa MARC, MODS y EAD. La Biblioteca Virtual Cervantes usa el estándar europeo MASTER (Manuscript access through Standards for Electronic Records) basado en TEI, etc. Según la European Commission Information Society and Media en su informe sobre los tópicos de investigación en preservación digital para 2020, expresa la necesidad de extender estos modelos de metadatos para crear un único modelo global y garantizar la interoperabilidad. El objetivo es pasar de la preservación de los datos a la preservación del conocimiento. [Billenness].

Para poder integrar colecciones de diferentes procedencias y con codificaciones diversas la modalidad de ingreso de datos descriptivos es por medio de archivos codificados en uno o varios estándares en formato XML, o por medio de formularios en línea que presentan los campos de un estándar. La transcripción de esta información a partir de múltiples fuentes (catálogos antiguos impresos, índices, descripciones varias, etc.) requiere de historiadores y mucho tiempo para cada colección. Debe entrenarse al historiador en codificación XML y en el estándar propiamente dicho.

Uno de los objetivos de los proyectos de este tipo, es agilitar este proceso por medio de una semi automatización de las tareas de codificación e ingreso de datos usando los archivos existentes y los manuscritos para la generación de metadatos y sistemas de búsqueda semántica [Billenness]. La función del historiador sería el ingreso de información descriptiva faltante y de corrección solamente.

Líneas de Investigación, Desarrollo e Innovación

1) Investigar, clasificar y desarrollar filtros para el preprocesamiento de las imágenes para alineación (skew correction) y filtrado de fondo, manchas y translúcidos.

2) Modelar y clasificar características de las diversas colecciones para descubrir automáticamente similitudes que permitan elaborar modelos de referencia y bases de calibración.

3) Investigar, clasificar y desarrollar algoritmos para el segmentado de los documentos en unidades aptas para su reconocimiento automático.

4) Determinar formas de describir la organización espacial de cada documento de la colección con vistas a la elaboración de algoritmos automatizables con la menor intervención de especialistas posible.

5) Investigar, desarrollar y experimentar con el análisis semántico del léxico basado en la teoría de la semántica de marcos para la extracción de texto para los descriptores de información y metadatos de las palabras del manuscrito y sus archivos descriptivos asociados. Investigar y analizar el uso de léxicos electrónicos, con la representación de los resultados del análisis lingüístico en forma de autómatas y en el uso de transductores para sistematizar tanto las tanto características formales del léxico (locuciones verbales) como de la sintaxis (gramáticas para la extracción automática de construcciones de un corpus). Adecuar y completar estos léxicos con los modismos propios de cada colección.

6) Investigar y extender un corpus lexicográfico apropiado para las tareas

descriptas, con las variaciones del uso del español de los siglos XVIII a XIX, en el Rio de la Plata y sus diferentes formas de escritura según el nivel cultural de los funcionarios, muchas basadas en la fonética y no en la gramática (en especial, para apellidos y lugares).

7) Investigar, desarrollar y experimentar con distintos criterios y algoritmos de indexación y sistemas de base de datos para permitir búsquedas semánticas a partir de las estructuras obtenidas en el punto anterior.

8) Transferencia al área educativa, para la formación de profesionales en el área que es una de las prioridades de los planes mundiales en preservación [Billenness]. El análisis de soluciones, el desarrollo de prototipos y herramientas y los análisis de efectividad a efectuar son temas apropiados para la realización de Tesis para los especialistas en Ingeniería en Informática en grado o posgrado. Otros desarrollos resultan ideales para Trabajos Profesionales. Ambas actividades están contempladas en los planes de estudio vigentes en la Facultad de Ingeniería de la Universidad de Buenos Aires. Es un objetivo de este proyecto guiar a los estudiantes interesados en el desarrollo de estos trabajos. Además algunas experiencias pueden incorporarse a los trabajos prácticos de distintas materias como experiencias con alto valor de enseñanza y aprendizaje.

Resultados y Objetivos

Para optimizar el uso de los recursos, se trabaja simultáneamente sobre dos corpus distintos resultados del proyecto anterior (UBACyT 2008-2010 I008). El primer corpus tiene características diversas y catálogos detallados que se estan usando como fuente para la calibración, para aumentar los léxicos existentes y verificar la calidad de los resultados obtenidos. Se trata de un corpus con catálogos y descripciones completas a nivel de ítem, inusual en este tipo de colecciones, e ideal como base de investigación y análisis de resultados. Estos resultados se aplicarán al segundo corpus de los documentos históricos de la UBA. Este corpus tiene una catalogación somera que muestra solamente la estructura física de los archivos y su división en tema y períodos. Los profesionales de historia del Instituto Dr. E. Ravignani de la UBA están realizando su catalogación usando el estándar ISAD(G) con información mínima. Este corpus permitirá verificar el rendimiento y efectividad de los algoritmos y métodos elegidos.

Para obtener una base de calibración para documentos históricos se sigue la metodología proppuesta en [Vamvakas]. Los pasos que propone son:

1. Pre-procesamiento del conjunto de imágenes de documentos.
2. Segmentación para obtener las unidades de reconocimiento.
3. Agrupamiento (clustering) de estas unidades según sus características.
4. Etiquetado de los clusters con las interpretaciones correspondientes.

De estas etapas se estan cumpliendo las primeras dos de ellas

Cada uno de estos pasos debe ajustarse al grupo de documentos en estudio y requiere de distintas técnicas según las características de cada manuscrito.

En el pre-procesamiento las imágenes deben alinearse, un proceso conocido como skew correction. Se deben aplicar filtros para corregir defectos del fondo, manchas y translúcidos que aparezcan en la imagen. Cada documento antiguo tiene diferentes características y por consiguiente será necesario determinar y desarrollar algoritmos adecuados para cada caso. Además se deben limpiar las interferencias frente-dorso de las hojas producto de la penetración de la tinta en el papel y de la acción del tiempo sobre el mismo. También aquí son necesarios algoritmos específicos. En este contexto, se continuará trabajando con el grupo de Dr. Rafael Duere Lins sobre la base de investigación y los algoritmos desarrollados que fueron aplicados con éxito a la colección Nabuco de manuscritos [Nabuco].

Para el análisis de extracción de textos y descripciones se usan técnicas basadas en la teoría de la semántica de marcos están siendo usadas en otros proyectos para la extracción semiautomática de textos de manuscritos e impresos antiguos [Goetz] [Ivanova][Palmer].

En particular y para el idioma español, en el proyecto FrameNet Español [FNE] de la Universidad Autónoma de Barcelona, que está basado en el proyecto FrameNet [FN] del ICSI (International Computer Science Institute) de la Universidad de Berkeley para el idioma Inglés.

Se estan analizando aplicaciones de tratamiento automático de corpus siguiendo los conceptos y métodos aplicados en FrameNet Español [Filmore] [Subirats] [Subirats 2009a], [Subirats 2009b], para su aplicación al idioma español sobre tres periodos de la evolución de la lengua española (siglos XVI y XVII, siglos XVIII y comienzos del XIX y segunda mitad del siglo XIX y siglo XX) [Zamora] [Bia 2001] y sus variantes del Río de la Plata, incluyendo las irregularidades de los apellidos españoles hasta el siglo XIX [Salazar] y sus errores de escritura (escritura de nombres y lugares basados en la fonética) [Zilio].

Se hicieron los ensayos correspondientes generando prototipos [Clúa 2010a], [Clúa 2010b], que fueron probados en el Instituto, de sus colecciones de manuscritos (documentos, periódicos y libros), del Ministerio de Justicia de la Provincia de Buenos Aires (expedientes judiciales), de la Society for Irish Latin American Studies (diario personal de un inmigrante), Archivo Güemes (Salta), Colección de fotos antiguas de la Patagonia para el Museo de La Plata, entre otros. Algunos de los resultados de este trabajo se pueden consultar en http://www.ravignani.filo.uba.ar/. El sitio se encuentra fuera de línea momentáneamente por el proceso de migración para adecuarlo a nuevos estándares de metadatos y por obras internas al edificio. Algunas colecciones se pueden consultar en un sitio alternativo en http://ravignanidigital.com.ar. El diario personal de un inmigrante se se pueden consultar en http://www.irlandeses.org/inicial.html.

Formación de Recursos Humanos

El proyecto cuenta en este momento con tres investigadores formados, tres en formación y cuatro estudiantes del último año.

Están en curso dos tesis de posgrado y cuatro tesis de grado.

Algunas experiencias, sobre todo en el área de indexación, se han incorporado a los trabajos prácticos de distintas materias.

Desarrollos parciales de las técnicas de pre-procesamiento han sido presentadas como Trabajos Profesionales, actividad contempladas en los planes de estudio vigentes en la Facultad de Ingeniería de la Universidad de Buenos Aires como alternativa a la Tesis de Ingeniería en informática.

Referencias

[Bia] Alejandro Bia y Manuel Sánchez-Quero. "Building Spell-Checking Facilities for Ancient Spanish". *The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, 2001 Joint International Conference,* New York City, NYU. 2001.

[Billenness] Clive S. G. Billenness, Report on the Proceedings of the Workshop "The future of the Past – Shaping new visions for EU-Research in digital preservation", *Cultural Heritage and Technology Enhanced Learning, European Commission Information Society and Media Directorate-General*, 2011 (http://cordis.europa.eu/fp7/ict/telearn-digicult/**future-of-the-past**_en.pdf).

[Cervantes] Proyecto de la Biblioteca Virtual Miguel de Cervantes. http://www.cervantesvirtual.com/proyectoES/BIMICESA.shtml

[Clúa 2010a] Osvaldo Clúa y M Feldgen. "Teaching Handwritten Document Restoration Techniques using Statistical Processing". *INTERTECH'2010 – XI International Conference on Engineering and Technology Education.* Ilheus, Bahia, Brasil, Marzo 2010.

[Clúa 2010b] Osvaldo Clúa y M Feldgen. "Image processing (and CATS) as an Introduction to Algorithmic Thinking". *FIE 2010 (40th ASEE/IEEE Frontiers in Education)*, Washington, DC, EE.UU. Octubre 2010.

[Feldgen 2002a] María Feldgen, O Clúa, Fernando Boro, Juan J. Santos. "Building an On Line Manuscript Heritage Digital Library

with XML". *JAIIO (Jornadas Argentinas de Informática e Investigación Operativa) Edición 2002*, Santa Fé, Santa Fé. 2002.

[Feldgen 2002b] María Feldgen, O Clúa, Fernando Boro, Juan J. Santos. "Argentinean Historical Heritage Project", *2002 ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, EE.UU. 2002.

[Filmore] Charles Fillmore. "Corpus linguistics" vs. "Computer-aided armchair linguistics", Directions in Corpus Linguistics. *Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*. Stockholm: M de Gruyter, 1996.

[FN] The ICSI Berkeley FrameNet Project http://framenet.icsi.berkeley.edu/papers/acl98.pdf

[FNE] FrameNet en Español: http://gemini.uab.es:9080/SFNsite

[Ivanova] Ivanova-Sullivan y Tania Dontcheva. "Lexical variation in the Slavonic Thekara Texts: semantic and pragmatic factors in medieval translation praxis". *OhioLINK* / Ohio State University, 2005. http://rave.ohiolink.edu/etdc/view?acc_num

[Nabuco] Mello, C.A.B, Duere Lins, R "Generation of images of historical documents by composition", Document Engineering. *Proceedings of the 2002 ACM symposium on Document Engineering*, 2002, 127 - 133

[Palmer] Marta Palmer, Daniel Gildea y Nianwen Xue, "*Semantic Role Labeling"*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2010.

[Salazar] Jaime de Salazar y Acha, "*Génesis y evolución histórica del apellido en España"*, Real Academia Matritense de Heráldica y Genealogía, 1991 .

[Subirats 2009a] Carlos Subirats. "Spanish FrameNet: A Frame Semantic analysis of the Spanish lexicon". *Multilingual FrameNets in Computational Lexicography Methods and Applications. Hans C. Boas*. Berlin, New York (Mouton de Gruyter), 2009, 135–162.

[Subirats 2009b] Carlos Subirats Rüggeberg. "La función del corpus en FrameNet Español". *Congreso Internacional de Lingüística de Corpus (CILC'09)*. Universidad de Murcia. 2009

[Subirats] Carlos Subirats y Marc Ortega, "Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores". *Estudios de Lingüística del Español* 10, 2000, http://elies.rediris.es/elies10/

[TEI] Text Encoding Initiative. http://www.tei-c.org/index.xml

[Vamvakas] G. Vamvakas, B. Gatos, N. Stamatopoulos y S.J. Perantonis. "A Complete Optical Character Recognition Methodology for Historical Documents." *Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on*, 525-532, 2008. 10.1109/DAS. 2008. 73.

[Zamora] Sergio Zamora, "El origen del Español". *Página del idioma Español* http://www.elcastellano.org/origen.html

[Zilio] Giovanni Meo Zilio, "*Estudios Hispanoamericanos: Temas Lingüísticos* "**,** Bulzoni, 1989.