

# Automatizando la extracción semántica de metadatos

Desde mayo de 2005, el Instituto de Información y Tecnología Humanística Avanzada (HATII) de la Universidad de Glasgow, ha emprendido una iniciativa para automatizar la extracción semántica de metadatos de objetos digitales, motivada por un estudio previo en la automatización o semiautomatización de la ingesta y los procesos de preservación (p. e. [2]). La construcción de diversos metadatos que describan el contenido, la información bibliográfica, la procedencia y las técnicas y requerimientos administrativos de un objeto digital, es un elemento crucial para la gestión y el mantenimiento de los repositorios, bibliotecas y archivos digitales ([8], [9]). La creación manual de estos metadatos es un proceso laborioso e intensivo, y con el ratio exponencial con el que se producen nuevos objetos digitales hacen que sea imposible confiar en que se lleven a cabo manualmente. El objetivo de esta iniciativa es avanzar en la automatización de metadatos antes que el volumen de objetos digitales sea demasiado grande y se necesite con urgencia.

## Ámbito

Formato del documento (PDF): Seleccionando un formato específico podemos reducir el espacio del problema a uno más manejable. Más específicamente, se espera que una herramienta que pueda manejar el formato PDF (cuyo uso es adoptado ampliamente en todos los repositorios, bibliotecas y archivos digitales, cómo también en el sector comercial y el uso particular) se use inmediatamente en una amplia gama de comunidades.

Los esfuerzos iniciales están centrados en los documentos de texto: Los métodos del Procesamiento del Lenguaje Natural (NLP) han demostrado su eficacia en

recuperación, extracción y clasificación de los documentos. Este artículo presenta cómo los métodos NLP y otras técnicas de aprendizaje de máquinas son un candidato perfecto para la aplicación de las primeras etapas de la extracción de metadatos. También se espera que el desarrollo de herramientas para la extracción de metadatos en documentos de texto tenga consecuencias en otros tipos de objetos, como

disponer de varios procesos de extracción para otros tipos de media (por ejemplo, imagen o material audiovisual) que dependerían de la minería asociada a los de texto.

## Progreso

En HATII, la Clasificación de Género Automatizado ha sido identificado como un paso fundamental para la Extracción Semántica de Metadatos Automatizada

El Género es una clasificación estructural y funcional de un documento que refleja una o más de las siguientes características:

- la intención del creador (p. e.: para informar, argumentar, instruir),
- la interpretación de la comunidad de usuarios (p. e.: como una colección de hechos, como una expresión de opinión, como un parte de una investigación),
- la prescripción de un proceso (p. e.: un artículo de una revista de publicación, descripción de las características de contratación, actas de una reunión), y,
- el tipo de estructura de datos (p.e: tabla, gráfico, tabla, lista).

Por qué clasificación de género?

1. Identificar el género limitará la estructura del documento a partir de la cual se podrán extraer más metadatos:

- El espacio de búsqueda para extraer metadatos se reducirá. Dentro de un mismo género se puede esperar que los metadatos como autor, las palabras clave y números de identificación o referencias aparezcan en un estilo y región similar.

2. Identificar el género va a crear una herramienta global que se va a relacionar con el trabajo específico de género:

- Existen trabajos independientes ([1], [3], [4], [10], [11]) para la extracción de metadatos dentro de un género específico que puede combinarse con un clasificador general de géneros para la extracción de metadatos en una gran amplitud de dominios.

## Referencias

- [1] Bekkerman, R., McCallum, A., Huang, G. (2004) Automatic Categorization of Email into Folders. Benchmark Experiments on Enron and SRI Corpora', CIIR Technical Report, IR-418.
- [2] ERPANET: Packaged Object Ingest Project. [http://www.erpanet.org/events/2003/rome/presentations/ross\\_rusbridge\\_pres.pdf](http://www.erpanet.org/events/2003/rome/presentations/ross_rusbridge_pres.pdf)
- [3] Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. A. (2000) Automatic Document Metadata Extraction using Support Vector Machines. Proceedings 3rd ACM/IEEECS Conference on Digital Libraries, 37-48.
- [4] Giurida, G., Shek, E. Yang, J. (2000) Knowledge-based Metadata Extraction from PostScript File. Proceedings 5th ACM International Conference on Digital Libraries, 77-84.
- [5] Kim, Y. and Ross, S. (2007) Detecting family resemblance: Automated genre classification. CODATA Data Science Journal, Volume 6, S172-S183, ISSN:1683-1470.
- [6] Kim, Y. and Ross, S. (2006) Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, Proceedings European Conference on advanced technology and research in Digital Libraries (ECDL), volume 4172 of Lecture Notes in Computer Science, pages 63-74. Springer.
- [7] Kim Y. and Ross, S. (2006) "The Naming of Cats": Automated genre classification. To appear International Journal of Digital Curation, preprint available at <http://eprints.erpanet.org/123>
- [8] PREMIS (PREservation Metadata: ImplementationStrategy) Working Group: <http://www.oclc.org/research/projects/pmwg/>
- [9] Ross S and Hedstrom M. (2005) Preservation Research and Sustainable Digital Libraries. International Journal of Digital Libraries (Springer) DOI: 10.1007/s00799-004-0099-3.  
Formatted: Bullets and Numbering

- Los recursos disponibles para la extracción de metadatos son diferentes para cada género, por ejemplo, los artículos de investigación, a diferencia de los de periódicos, tienen una lista de referencias a otros estrechamente relacionados, que permiten obtener una mejor clasificación del tema.

3. Alcanzar nuevos géneros que aparentemente no pertenecen al contexto convencional de las bibliotecas, pero que son necesarios para mantener las prácticas de gestión de objetos digitales hasta la fecha.

4. Diferentes instituciones podrían centrarse en objetos digitales de diferentes géneros. La clasificación en géneros soportará la identificación, selección i adquisición de objetos en conformidad a las directrices locales.

## La herramienta de clasificación

La herramienta de clasificación en género está destinada a evaluar los documentos estadísticamente basándose en sus características visuales (p.e.: la cantidad de espacio en blanco en el documento), los elementos estilísticos (p.e.: la frecuencia y el uso de determinados artículos), modelo lingüístico (los términos importantes que caracterizan el documento), los patrones semánticos (p.e.: el uso de frases subjetivas), y los recursos externos disponibles (p.e.: la URL de origen), para determinar su género. Los resultados de esta investigación se han publicado en varios artículos ([5], [6], [7]).

## Flujo de trabajo de la Extracción de metadatos

El estudio en este proyecto estará atado a una arquitectura general de extracción automatizada de metadatos y el proceso de ingestión, usando las siguientes flujo de trabajo:

- Recibir objeto digital,
- determinar el género o la clase estructural del objeto,
- Decidir la mejor herramienta de extracción de metadatos para el género identificados, o la mejor opción basada en la estructura,
- Si no existe tal herramienta, podrá solicitar su creación,
- Una vez que la herramienta ha sido seleccionada o creada, extraer los metadatos necesarios,
- ingerir el objeto y los metadatos en el repositorio o en el archivo.

## Conclusiones

La automatización de ingesta, la preservación, y los procesos de selección dentro de un repositorio digital han dejado de ser una conveniencia para convertirse en una necesidad. El almacenamiento, el compartir, y el uso cruzado institucional de la información se ha convertido en una realidad, i el control manual de estos procesos es lento, inefectivo e ineficiente. Los desafíos de adaptarse a la información se han de encontrar con innovadores procesos de extracción, autenticación i evaluación. La extracción semántica automática de metadatos aún está en sus inicios. Cosa que hace necesario, que se sigan dedicando esfuerzos para tirarlo adelante y refinar las herramientas de extracción, para así integrarlo con otros procesos en las bibliotecas, archivos y comunidades relacionadas.