# Carga de pubmed en local

## PubMed

Pubmed (<https://www.nlm.nih.gov/bsd/pubmed.html>

) es un repositorio público de citas sobre artículos científicos relativos a biomedicina, promovido por la Biblioteca Nacional de Medicina de Estados Unidos a partir de las citas y los resúmenes elaborados en la base de datos denominada Medline (<https://www.nlm.nih.gov/bsd/medline.html>). Las citas son recogidas mediante un registro que se clasifica siguiendo un vocabulario medico (MESH: <https://www.nlm.nih.gov/mesh/meshhome.html>), y que en función del propio artículo es enlazado con PubMedCentral si el artículo es de libre acceso o con el proveedor adecuado. Una vez constituido el registro, este pasa a estar accesible en varios formatos, tanto para usuarios finales, como para sistemas integrados. Esta integración se puede hacer mediante servicios Web, pero para estudios sobre todo el conjunto, también proporciona un servicio de transferencias de ficheros, en el que se proporciona una línea base anual y los incrementos de diarios de cada año. Los ficheros se encuentran comprimidos y con un código MD5 que permite garantizar su descarga correcta.

# Formato de los ficheros

Los ficheros de PubMed siguen un esquema XML basado en DTD y/o en XSD que se actualizan con cada línea base. La DTD más actual se encuentra en <https://www.nlm.nih.gov/databases/dtd/>. Al ser fruto de sistemas de gestión de datos complejos, el mecanismo que ha resultado más fácil para poder tratar los ficheros en un entorno de programación Java – JAXB ha sido la DTD. Mediante un proceso basado en la herramienta XJC del estándar de mapeado de clases Java y Xml denominado JAXB (Java Architecture for XML Binding) se pueden obtener una serie de clases Java que mapean directamente el contenido XML basado en una DTD.

# ¿Porque descargar las citas?

Como se ha comentado, las citas son accesibles mediante interface REST por ejemplo. Este mecanismo para accesos puntuales o para mostrar toda la información disponible es suficiente ara muchos tipos de proyectos, pero en el caso que nos ocupa, el aprendizaje automático a partir de todo el banco de citas hace inviable el acceso a los más de 28 millones de registros. De hecho, el propio sistema otorga límites de acceso para evitar que accesos legítimos puedan ser identificados como ataques de DoS.

# Para facilitar mecanismos de entrenamiento de estas herramientas, y, e partida, estudiar el propio conjunto de datos se han desarrollado a lo largo del tiempo varias iniciativas que permiten contar con una base de datos relacional que incorpora el contenido de PubMed, siendo la más parecida a la que se enmarca en este proyecto PubMedPortable: A Framework for Supporting the Development of Text Mining Applications).

# Aportaciones de la descarga local de este proyecto

Las iniciativas comentadas se basan en la descarga del repositorio para poder hacer posteriormente los estudios y trabajos principales. En general reproducen en un modelo relacional el documento XML, pero determinadas entidades como las revistas de origen o los autores solo se vinculan como atributos de la cita, en vez de generarse una entidad que permita la obtención de información propia de un modelo relacional. Por otro lado, su orientación en la mera disposición de los datos, en general con una capacidad de búsqueda limitada a la identificación de la cita y poco más.

En el caso de este proyecto, se aportan tres herramientas adicionales:

1. Modelo JAXB para Java que permite tratar con toda la información de una cita de PubMed
2. Modelo relacional de datos revisado que genera entidades no solo para las citas, sino para las revistas, los autores, etc.
3. Se amplía la capacidad de búsqueda insertando en una base de datos documental la información de las citas. Este apartado permite obtener resultados de citas en términos de “scores” en el contexto de la búsqueda o del repositorio, citas “cercanas”, que se pueden utilizar desde las herramientas software, permitiendo arropar la información de una cita con datos de su posición en el repositorio respecto a una búsqueda.