# 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 producido a partir de la clasificación y resúme elaborados por dicha Biblioteca, y que constituyen 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>), que en función del propio artículo es enlazado con PubMedCentral si el artículo es de libre acceso o el proveedor lo permite. Una vez constituido el registro, este pasa a estar accesible en varios formatos, tanto para usuarios finales, como para sistemas informáticos. Esta integración se puede hacer mediante servicios Web, pero para estudios sobre todo el conjunto, se proporciona un servicio de transferencias de ficheros, en el que se pueden descargar ficheros a partir de una línea base anual y de los incrementos diarios. Los ficheros se encuentran comprimidos y con un código MD5 que permite garantizar la descarga correcta.

## Formato de los ficheros

Los ficheros de PubMed siguen un esquema XML basado en DTD y/o en XSD que se actualizan anualmente 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 es Java – JAXB. Mediante un proceso basado en la herramienta XJC del estándar de mapeado de clases Java y Xml JAXB (Java Architecture for XML Binding), se pueden obtener una serie de clases Java que mapean directamente el contenido XML basado en una DTD.

## ¿Como se ha procesdo el XML?

El principal inconveniente de este mecanismo es que cada fichero incorpora en torno a 30000 citas con una ocupación de 400 megabytes, que se debe subir a memoria. Si el objetivo fuese el hacer un tratamiento sobre algunos de los datos de cada uno de estos ficheros (analizar los abstracts de las citas por ejemplo), se hubiese utilizado la tecnología de stream STAX, preparada para procesar ficheros XML pasando por la información por un buffer limitado en tamaño, pero donde la programación se hace muye específica para cada problema en concreto.

En contrapartida el modelo de clases proporcionado por JAXB, facilita infinitamente la programación debido a que cada uno de los items (citas) que componen el fichero es de gran complejidad ya que el detalle de información con la que se cuenta es muy amplio.

Como uno de los objetivos del proyecto es contar con un DataSet para la investigación es necesario recuperar el máximo de información posible para su almacenamiento en el entorno local.

## ¿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, de 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 en entornos locales que incorporan 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).

# Estas herramientas incorporan una parte de la información muy concreta, la relativa a PubMed, pero no integran diferentes modelos que complementan a PubMed como PubMed Central (Artículos a texto completo), MESH (Tesauro de información médica para la clasificación temática de artículos u otros modelos resultantes de diferentes aproximaciones y estudios.

Las iniciativas comentadas se basan en la descarga completa del repositorio, su inclusión en la Base de Datos a partir de la que poder hacer posteriormente los estudios y trabajos requeridos. 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 integrarse generando entidades que permita explotar las relaciones entre los datos propia de los modelos relacionales. En definitiva, su orientación es tener los ficheros XML serializados en tablas SQL, utilizando las relaciones para mantener la jerarquía del fichero XML, sin integración de datos y por lo tanto con una capacidad de búsqueda y de información limitada.

# Aportaciones de la descarga local de este proyecto

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

1. Modelo JAXB para Java que permite tratar con toda la información de una cita de PubMed.
2. Stream de citas que permite el consumo de ficheros XML mediante programación funcional, con un modelo de beans simplificado, ya que el modelo obtenido de la DTD incluye muchos pasos intermedios y nomenclaturas diferentes para campos semánticos idénticos (abtract, otherabstract para citas, abstract para artículos en libros, etc son campos diferentes en la DTD pero semánticamente son lo mismo).
3. Modelo relacional de datos revisado que genera entidades no solo para las citas, sino para las revistas, los autores, etc. Este modelo va a permitir interrogar la base con mecanismos propios de Business Intelligence y poder obtener **datos de recuento y estadísticas del** propio **repositorio**.
4. Modelo documental de citas que amplía la capacidad de búsqueda. Este apartado permite relacionar los resultados de las búsquedas en términos de “scores”, “cercanía”, porporcinando **datos** asociados a la búsqueda.

Pubtator BIOC

https://academic.oup.com/database/article/doi/10.1093/database/bau073/2634750

Conceptos de partida DataSets

tmChem

Dnorm

tmVar

SR4GN

GenNorm

https://github.com/jimmyskull/tmVar/blob/master/tmBioC.key