Efficiently processing and storing library linked data using Apache Spark and Parquet

Kumar Sharma, Ujjal Marjit and Utpal Biswas

Information technology and libraries, ISSN 2163-5226, Vol. 37, n. 3, 2018, p. 29-49

Cada vez más organizaciones y centros de investigación están utilizando tecnologías de la web semántica para presentar datos usando RDF, y entre estas encontramos a las bibliotecas, que tratan de reemplazar los sistemas de catalogación actuales basados en MARC por técnicas de datos vinculados como BIBFRAME. Se ha logrado que la biblioteca forme parte de la web, pero sigue habiendo problemas con los grandes datos de la biblioteca. El término *big data* contiene datos que no se pueden procesar utilizando softwares tradicionales; durante el proceso de conversión de datos de la biblioteca a RDF se pueden encontrar varias dificultades: la detención del proceso por la gran cantidad de datos, capacidad de almacenamiento insuficiente o problemas a la hora de la recuperación, por lo que los bibliotecarios deben conocer programas que les ayuden a resolver estos problemas, ya que las herramientas tradicionales de gestión de datos no tienen la capacidad de gestionar todo el volumen de datos que conocemos como *big data*. Las bibliotecas deben rediseñar la forma en que contribuyen a la web de datos, deben integrar sus datos con la web, pero el estándar MARC es incapaz de expresar relaciones entre registros y campos del registro, por lo que la mayoría de los recursos bibliográficos almacenados en MARC están destinados a la conversión a datos vinculados para aprovechar todo el potencial de la web; los datos de la biblioteca deben transformarse en un formato al que se pueda acceder más allá de la biblioteca utilizando tecnologías como la web semántica y los datos vinculados basados en RDF. Este artículo nos acerca al entorno de los datos enlazados y la web semántica, mostrándonos que hay herramientas como Apache Spark y Parquet que nos pueden ayudar en el proceso de gestión de datos en el entorno *big data*.

Resumen elaborado por Marta Cerrada Rodríguez