MMRepo: Storing qualitative and quantitative data into one big data repository

Ingo Barkow, Catharina Wasner y Fabian Odoni

IASSIST quarterly, ISSN 0739-1137, winter 2016, p. 14-19

Almacenar diferentes tipos de datos de diferentes dominios y enriquecerlos con metadatos es la principal tarea de los centros de búsqueda de datos, archivos de datos y repositorios científicos. Desde la perspectiva de las Tecnologías de la Información la cuestión es si diferentes datos pueden ser almacenados en la misma infraestructura técnica. Los archivos pueden variar en tamaño, documentación y tipo. Los métodos más habituales de tratar esta mezcla son dos: almacenar todos los datos en bases de datos relacionales, lo que permite a los usuarios elegir qué variables del sistema pueden ser exportables. Es un método ventajoso para gran cantidad de datos, pero afecta a la calidad. Los datos de calidad deben ser almacenados como objetos binarios grandes (BLOB), pero estas tecnologías no combinan bien y son frecuentes las incompatibilidades técnicas; el otro método es almacenar los datos como archivos en un servidor. En este caso los metadatos se proporcionarán a través de una base de datos relacional externa, con atributos adjuntos o añadiendo archivos con información de los metadatos. Es un buen método cualitativo, pero afecta a la cantidad y el acceso de los usuarios. Para demostrar cómo los repositorios de datos manejan diferentes tipos de datos, dos organizaciones de Alemania fueron seleccionadas como ejemplos: el German Institute for International Educational Research (DIPF) y el Leibniz Institute for Social Sciences (GESIS). DIPF tiene activos tres repositorios, para datos cualitativos, cuestionarios y respuestas y documentación de datos. Tecnológicamente son totalmente diferentes y están optimizados para sus respectivos contenidos: cantidad, calidad y metadatos. GESIS tiene una aproximación similar, solo que usa desarrollos independientes para cada departamento. La separación de almacenamiento de datos es válida y explicable porque las bases de datos utilizadas llevan años e incluso décadas en uso. Pero queda pendiente la cuestión de si aproximaciones más modernas podrán lograr la unificación. Un candidato es la tecnología de inteligencia e datos, que permitiría el uso de sistemas de archivo basados en clúster, el acceso a búsquedas semánticas y la aplicación de minería de textos. Un prototipo que usa esta tecnología de inteligencia de datos es MMRepo, que experimenta con metadatos, datos cualitativos y datos cuantitativos en una única infraestructura. De momento (otoño de 2016) es un proyecto en elaboración. Se ha sometido a diversos test, con un resultado conceptualmente positivo pero con ciertas limitaciones en su aplicación.

Resumen elaborado por Antonio Rodríguez Vela