

Can indexing be automated? The example of the Deutsche Nationalbibliothek

Ulrike Junger

Cataloging & Classification Quarterly, ISSN 1544-4554, Vol. 52, n. 1, 2014, p. 102-109

La Deutsche Nationalbibliothek (DNB, Biblioteca Nacional Alemana) ha estado trabajando en el desarrollo y la implementación de procedimientos para la asignación automatizada de encabezamientos de materia para las publicaciones en línea. En este artículo se describe este proyecto, sus resultados y problemas. El aumento de las publicaciones en línea hizo evidente que la DNB no tenía capacidad para catalogar e indizar todo este material de forma tradicional (es decir, intelectual y manualmente). Por ese motivo, se decidió en 2009 que se iniciaría un gran proyecto para desarrollar métodos para la automatización del proceso de publicaciones monográficas en línea. Un objetivo importante del proyecto fue automatizar la asignación de encabezamientos de materia (Gemeinsame Normdatei, GND). Este proyecto llamado *Petrus* se realizó de 2009 a 2011, con continuaciones en 2012 y 2013. Una de las condiciones del proyecto fue que debería ser un sistema o software disponible en el mercado. Los primeros dos años se dedicaron al análisis del mercado y a probar exhaustivamente varios sistemas. Al final, se decidió por Averbis Extraction Platform, un sistema desarrollado por la compañía Averbis. El proceso que el sistema Averbis realiza, comienza con un análisis textual de las publicaciones en línea sobre la base de diversos métodos lingüísticos con el fin de extraer términos, seleccionando contenidos de las partes textuales y de los títulos de las publicaciones. A continuación ordena los términos extraídos de acuerdo a su significado e importancia. Los términos extraídos se cotejan con el vocabulario controlado del GND. Después de la adquisición del sistema Averbis se llevaron a cabo una serie de pruebas para explorar qué configuraciones traerían mejores resultados. Uno de los problemas que se planteó fue cómo hacer la evaluación y medición de la calidad de los encabezamientos de materia generados automáticamente. Se decidió controlar intelectualmente los resultados utilizando una muestra de títulos. Después de la prueba se consideró que la precisión era deficiente, y los resultados generales obtenidos en la indización automatizada insatisfactorios. Entre los problemas detectados estaban la discriminación entre los nombres de persona, geográficos y términos temáticos, o los términos generales que necesitan otro encabezamiento para darles un significado concreto. Otro de los temas es la necesidad de seguir trabajando en los valores de confianza y, hasta ahora, el valor informativo del valor de confianza no ha sido suficiente. Aunque quedó claro que no es fácil desarrollar y poner en práctica un procedimiento que utiliza un vocabulario controlado universal para la indización automática de una colección universal, la DNB decidió a principios de 2012 continuar con el proyecto dos años más.

Resumen elaborado por Natividad Escavias Extremera