Szövegbányászati rendszer fejlesztése a Magyar Elektronikus Könyvtár számára

A tanulmány az Infopark Alapítvány pályázati támogatásával megvalósuló, a Magyar Elektronikus Könyvtár (MEK) számára készített szövegbányászati rendszer három aspektusát mutatja be. Ismertetésre kerül a rendszer architekturális felépítése, amely egy távoli eljáráshívással megszólítható skálázható és elosztott kliens-szerver megoldás. Részletesen bemutatásra kerül az alkalmazott dokumentum reprezentációs módszer, amelynek lényege, hogy veszteségmentesen képes kompakt módon reprezentálni HTML vagy XML módon formázott dokumentumokat olymódon, hogy a dokumentum reprezentációban a szövegbányászati műveletek nagy hatékonysággal elvégezhetőek. Lényege egy speciális adatbázis struktúra, amely felett az indexelést egy végesállapotú automata segítségével végezzük el. Végül bemutatásra kreülnek az alapvető szövegbányászati funkciók, amelyek képesek kielégíteni mind a MEK, mind pedig üzleti szereplők igényeit.

Kulcsszavak: szövegbányászat, dokumentum reprezentáció, architektúra, Magyar Elektronikus Könyvtár, MEK