The Diploma Thesis is targeted to design and implement the system for collecting XML-family data from the Internet. The aim of the task is to automate the data collection process and download full structures of XML documents.
A comparison of four existing data collection systems took place at the beginning to choose one of the systems as a base of the solution. The open source web crawler Apache Nutch was identified as the most suitable. Then necessary extensions and modifications of the crawler were designed and implemented in order to make the crawler efficient in downloading XML-family documents.
Downloaded XML-family data were analyzed and evaluated using the Analyzer application, which was enhanced within this Diploma Thesis in order to process the data.
The main outcome of Diploma Thesis is an exploitable system collecting the XML-family documents from the Internet. Implemented modification and extensions of the system lead to elimination of „useless“ documents download, improving the ratio targeted XML-family documents.