Text data collections enable the deployment of artificial intelligence algorithms for novel tasks. Such collections often contain miscellaneous personal data and other sensitive information that complicates sharing and further processing due to the personal data protection requirements. Searching for personal data is often carried out by sequential passes through the complete text. The objective of this thesis is to create a tool that helps the annotators decrease the risk of data leaks from the text collections. The tool utilizes pseudonymization (replacing a word with a different word, based on a set of rules). During the annotation process, the tool tags the words as "public", "private" and "candidate". The task of the annotator is to determine the role of the candidate words and detect any other untagged private information. The private words then become the subject of the pseudonymization process. The auto-tagging tool utilizes a named entity recognizer and a database of rules. The database is automatically improved based on the decisions of the annotator. Different named entity recognizers were compared for the purpose of personal data search on the collection of the ELITR project. During the comparison, a method was found which increased the sensitivity of the named entities detection which also increased the sensitivity of the "candidate" words detection. The work also introduces the reader to the legal issues of personal data protection during the creation of text collections.