Diplomová práce je zaměřena na návrh a implementaci systému pro sběr veřejně dostupných dokumentů z rodiny XML na Internetu. Záměrem je zautomatizovat a zjednodušit proces sběru dat a dosáhnout stažení kompletních struktur dokumentů z rodiny XML.
Na začátku práce byla provedena analýza čtyř systémů pro sběr dokumentů z Internetu, aby jeden z nich mohl být vybrán jako základ pro řešení diplomové práce. Jako nejvhodnější se ukázal open source webový crawler Apache Nutch. Nově byly navrženy a implementovány úpravy tohoto crawleru tak, aby byl efektivní při sběru XML dokumentů.
Pro zpracování stažených dokumentů byla využita aplikace Analyzer, která byla na základě testu na reálných datech upravena tak, aby zpracování těchto dat umožnila.
Hlavním přínosem diplomové práce je reálně využitelný systém pro sběr dokumentů z rodiny XML z Internetu. Díky rozšíření a úpravám crawleru Apache Nutch se podařilo podstatně eliminovat stahování a ukládání zbytečných dokumentů a zlepšit skladbu stažených dokumentů ve prospěch XML dat.