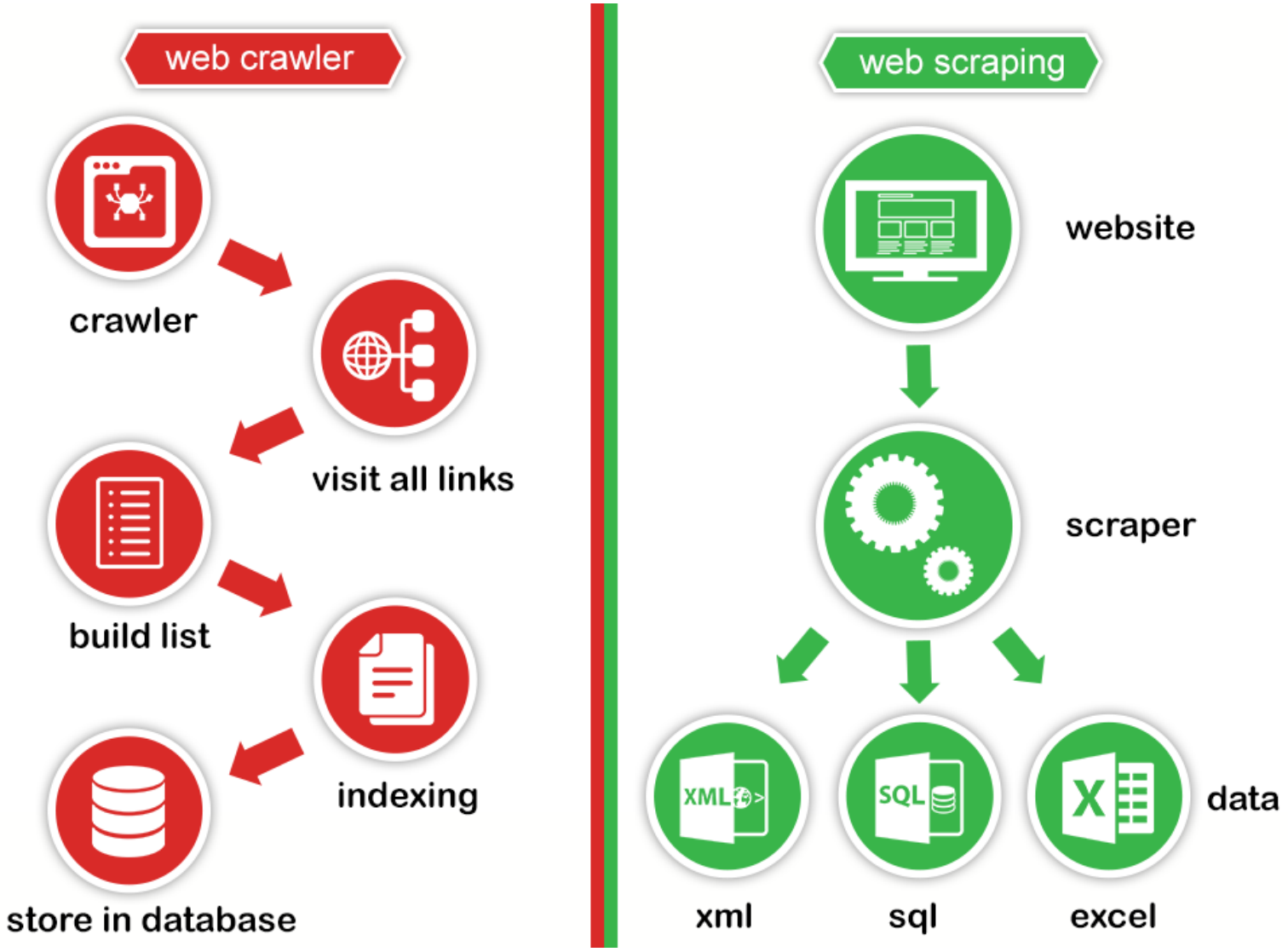
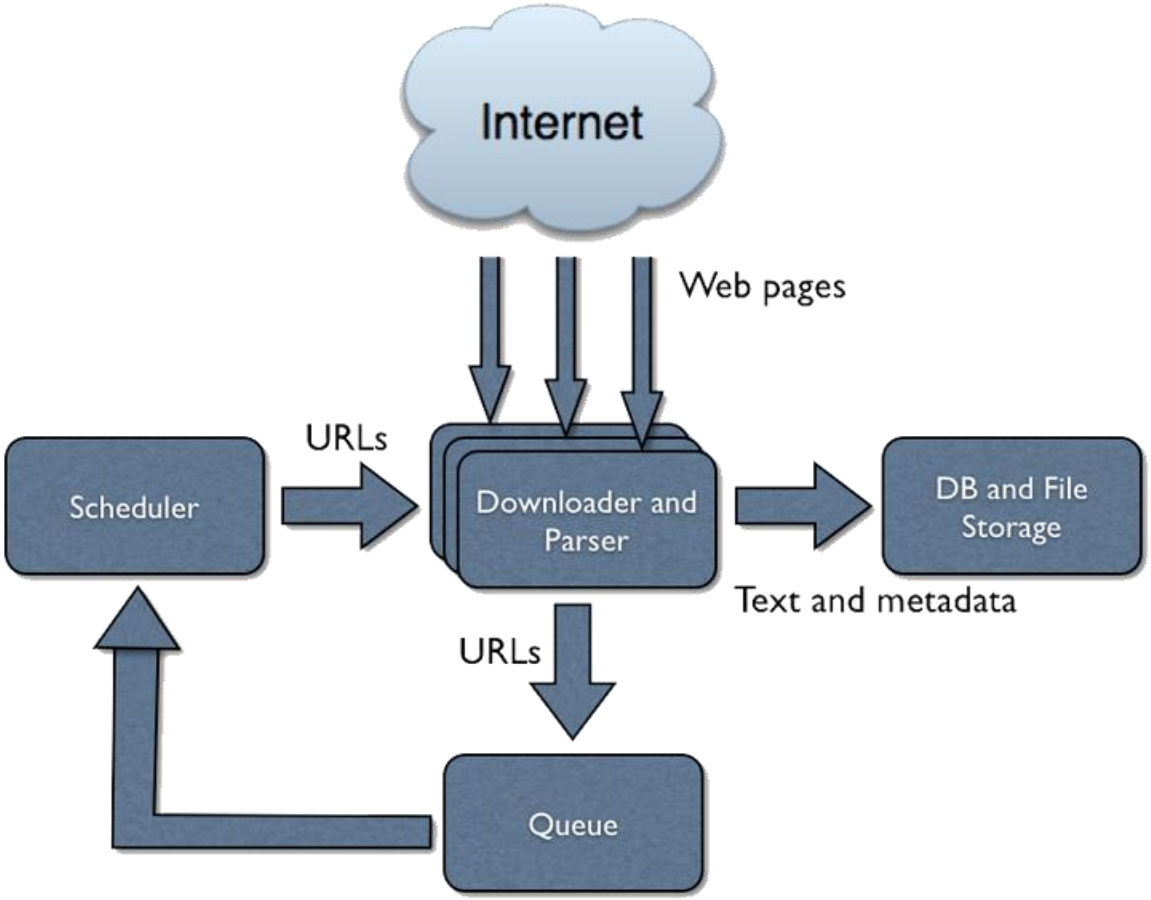
Scraping

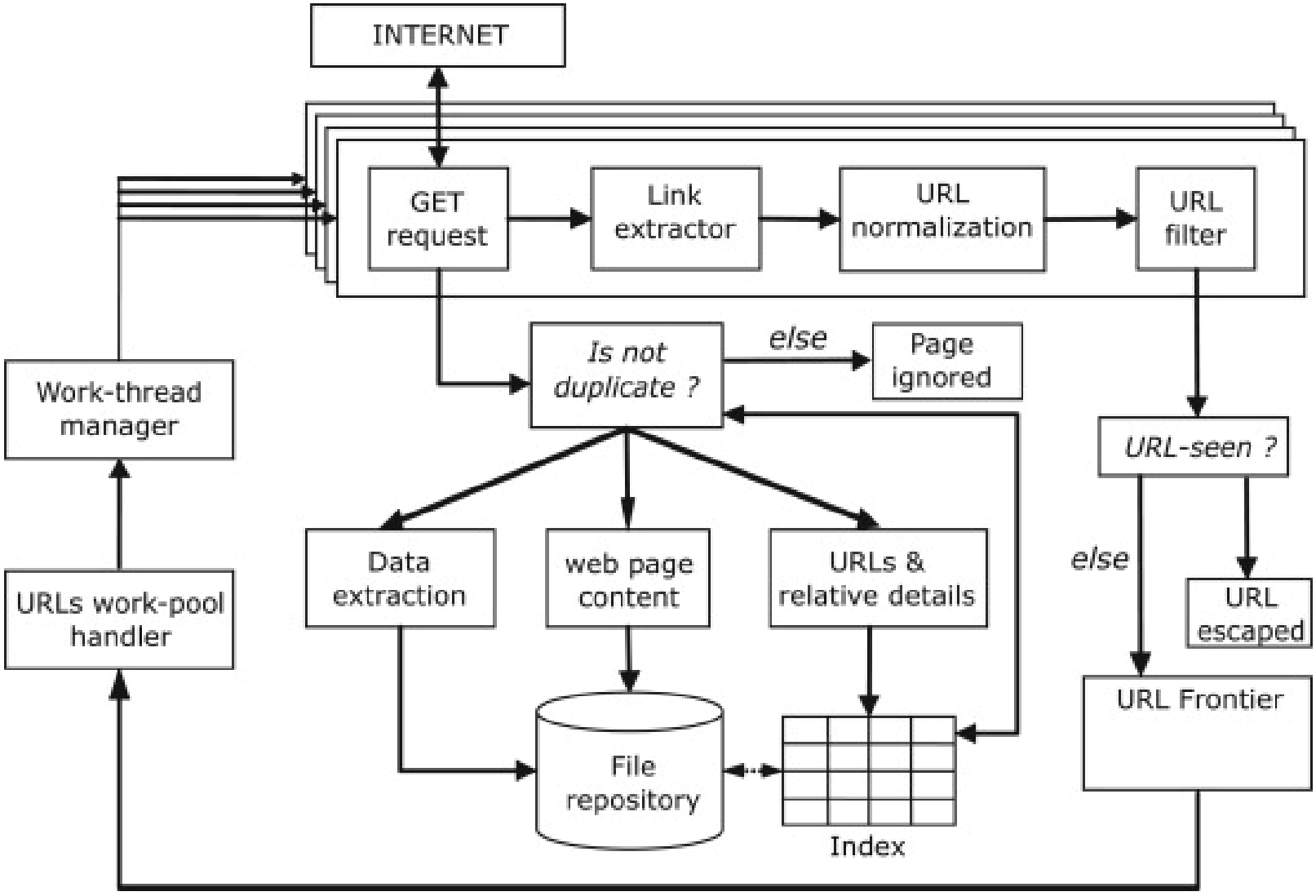
Data scraping	Data Crawling
Involves extracting data from various sources including web	Refers to downloading pages from the web
Can be done at any scale	Mostly done at a large scale
Deduplication is not necessarily a part	Deduplication is an essential part
Needs crawl agent and parser	Needs only crawl agent


web crawler
crawler
visit all links
build list
indexing
store in database
web scraping
website
scraper
XML
SQL
X
data
xml
sql
excel

Architecture

Process of **automatically** requesting a web document and **collecting** information from it




INTERNET
GET request
Link extractor
URL normalization
URL filter
Work-thread manager
Is not duplicate ?
else
Page ignored
URL-seen ?
Data extraction
web page content
URLs & relative details
else
URL escaped
URLs work-pool handler
File repository
Index
URL Frontier

exercises

https://validator.w3.org/unicorn/?ucn_lang=ko