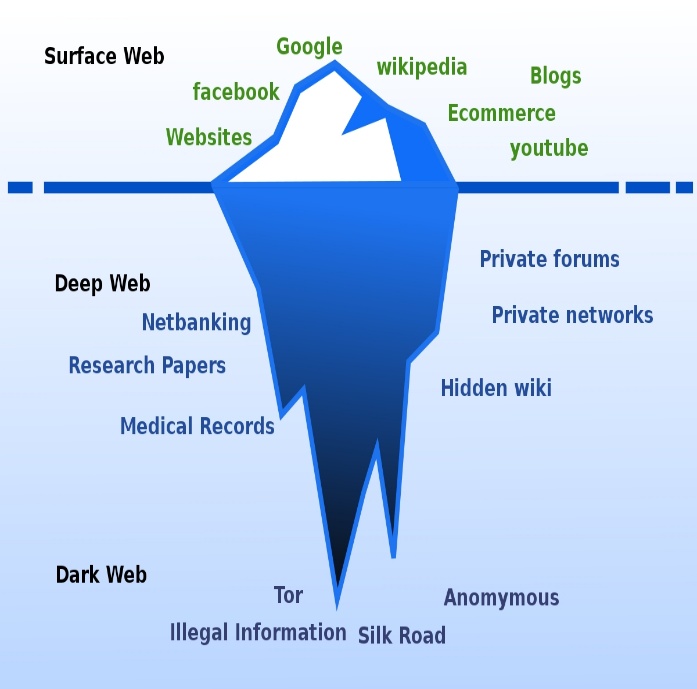
**CronBatch © in 45 minutes**

**Overview**

CronBatch is a programmable web application written in php, javascript, html, sql, c, c++ to explore the world wide web areas: surface, deep and dark. The CronBatch main purpose is to provide up-to-date info on the diversity of the two web areas through sampling and scenarios built out of the analytics. It is a web application ideal for cloud enabled environments and integrates well in enterprise systems.

CronBatch can be programmed to harvest the world wide web for information (seeds) over a specific timeframe, specific area of interest. Surface areas are explored with immediacy and results can be obtained istantaneously whilst the deep areas, being interconnected with many resources are explored offline by cron applications.

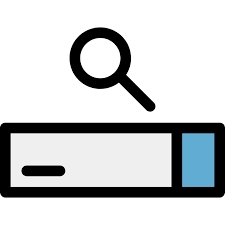


CronBatch can harvest seeds when programmed to gather documents all over the internet via search engines, private forums remote and local files, databases or simply by input of the web addresses to target as potential storage of info. CronBatch carries out internet scraping through querying search engines at surface and deep level. Remote or local files can also contain dark information, CronBatch can scrape such files for linking with higher levels. The module /search/search.php contains the class Search with methods for scraping the internet, databases and many file types.

****

**Seeds harvesting and first crawling frontier**

LP

******** LO mo mo

DB

Queries

/

Data dumping

**Google, Internet Archive, Wikipedia, Wolframalpha**

**DPLA, arXiv, IEEE, Elsevier, Springer, Core, Unpaywall**

Scraped data sources tracking and seeds filtering

Remote repository

Local files

Internet query

Scrape files and databases

**Cron jobs**

****

First

Crawling

Frontier

Files

scraping

L

First

Crawling

Frontier

Internet

scraping

/

Search engine queries

Module

frontier.php

frontier.php

Module

search.php

The first phase of seeds search is fundamentally scraping of available resources (Internet, local/remote databases, files, directories). Subsequently queries for fine tuning and filtering of the data can be activated so that the resulting records of the first crawling frontier will be more respondent to user needs.

**Crawling frontier**

First

Crawling

Frontier

First

Crawling

Frontier

First

Crawling

Frontier

**P**

Module

jsonquery.php

frontier.php

Module

frontier.php

frontier.php

At the end of the first frontier different seeds from different sources needs filtering in order to avoid unnecessary or inconclusive subsequent searches. The filtering can be done increasing queries targets with more keywords accuracy. Test results can show whether or not more or different keywords a to better results. Result lists can also be cut at a point fixing a number so to limit unnecessary tails or excessive long lists with insignificant results. Other checks can be carried out on each single scraped seed with user defined routines before they reach the end of the frontier. The crawling frontier is thus defined with filtered seeds so that the crawler can target documents, files, and web pages on the world wide web. Crawling frontier seeds are JSON formatted following the standard RFC 8529.

**Seed format**

{ "issearch" :

{

“total\_records” : “123456789”,

“search\_date” : “00/00/00”,

“search\_time” : “hh:mm:ss",

“search\_source” : “search engine / db / file”,

“search\_scheme” : “search scheme”,

“search\_host” : “search host”,

“search\_port” : “search port”,

“search\_user” : “search username”,

“search\_pass” : “search password”,

“search\_path” : “search path”,

“search\_query” : “search query”,

“execution\_time” : “search execution time”

}, "records": [{

"title" : "title",

"id" : "url",

"summary" : "snippet / summary / abstract",

"author" : "main author",

"doi" : "doi"

}, {

"title" : "title",

"id" : "url",

"summary" : "snippet / summary / abstract",

"author" : "main author",

"doi" : "doi"

}]}

**Crawling schema and scenario design**

**Remote and local batch programming**

**Cron jobs**

Crawl urls by keywords and document structure

Define crawling frontier

Remote repository

Local files

BN

Build network with web site maps

Select relevant urls

Remote repository

Local repository

Select relevant urls

Search the network for keywords

**Web**

**App**

**Cron jobs**

Cron jobs are scheduled for crawling the web once CronBatch has been programmed and seeds have been filtered at the frontier.