ETL Specification

Date: 1 November 2018
Author: Ulrich Karstoft Have
Project: Project: P002 – mHealth in Denmark: Findings from the web archive

Data to be delivered

The project needs three data sets for each of the search queries: *metadata*, *text content* and *links*.

The search engine used was NetSearch (Blacklight). The fields in the section “Data Description” correspond to the fields in the Solr index of Netarkivet.

The output format of the ETL process should be a text file that uses a tab to separate values. The file can include a header line on the first line with the field names but this is not required if the order of the fields is the same as in the table below.

Search Queries

The project has defined 3 queries.

1. Dansksprogede sider m. mHealth

Pages in Danish containing "mHealth" excluding content from Twitter or Youtube.

Field: Text

Search words:

```
mHealth content_language:"da" -domain:"twitter.com" -domain:"youtube.com"
```

Total No. of documents found: 54,131

2. Engelsksprogede sider m. mHealth

Pages in English containing "mHealth", excluding content from Twitter or Youtube.

Field: Text

Search words:

```
mHealth content_language:"en" -domain:"twitter.com" -domain:"youtube.com"
```

Total No. of documents found: 104,856

3. Mobile sundhedsteknologier (og varianter af denne søgning)

Field: Text

Search words:

```
mobil* sundhedsteknologi* content_language:"da" -domain:"twitter.com" -domain:"youtube.com"
```

Total No. of documents found: 478,631

Data Description

Metadata

The fields to be extracted:

Field No.	Field Name
1	id
2	arc_harvest
3	arc_job
4	crawl_date
5	wayback_date
6	hash
7	url
8	title
9	content_type_norm

Text

The fields to be extracted:

Field No.	Field Name
1	id
2	crawl_date
3	hash
4	url
5	content

Links

The fields to be extracted:

Field No.	Field Name
1	id
2	crawl_date
3	hash
4	url
5	links_domains

Any technical questions can be directed to Ulrich Karstoft Have (ukh@kb.dk).