

CommonCrawlDataDumper.java

dump method

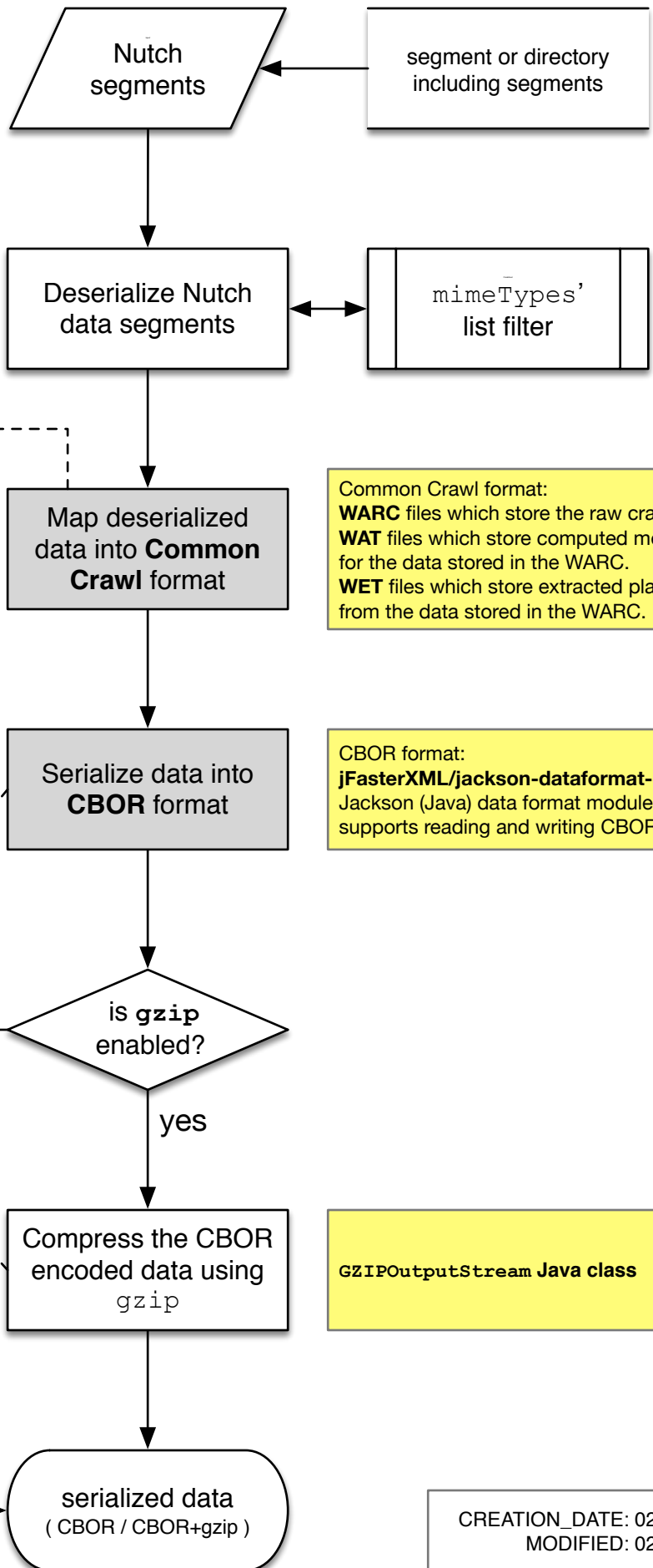
javadoc example: JSON-based structure

```
{
  "url": "...",
  "timestamp": "1411623696000",
  "request": {
    "method": "GET",
    "client": {
      "hostname": "...",
      "...": "..."
    },
    "headers": {
      "Accept": "...",
      "...": "..."
    },
    "body": null
  },
  "response": {
    "status": "200",
    "server": {
      "hostname": "somepage.com",
      "address": "55.33.51.19",
    },
    "headers": {
      "Content-Encoding": "gzip",
      "...": "..."
    },
    "body": "...",
  },
  "key": "...",
  "imported": "1411623698000"
}
```

Build one single CBOR-serialized/CBOR+GZIP-compressed file

LEGEND

- current specification
- task to be implemented
- (semi)implemented task
- official documentation



Common Crawl format:
WARC files which store the raw crawl data.
WAT files which store computed metadata for the data stored in the WARC.
WET files which store extracted plaintext from the data stored in the WARC.

CBOR format:
jFasterXML/jackson-dataformat-cbor
Jackson (Java) data format module that supports reading and writing CBOR.

GZIPOutputStream Java class

CREATION_DATE: 02-24-2015
MODIFIED: 02-25-2015