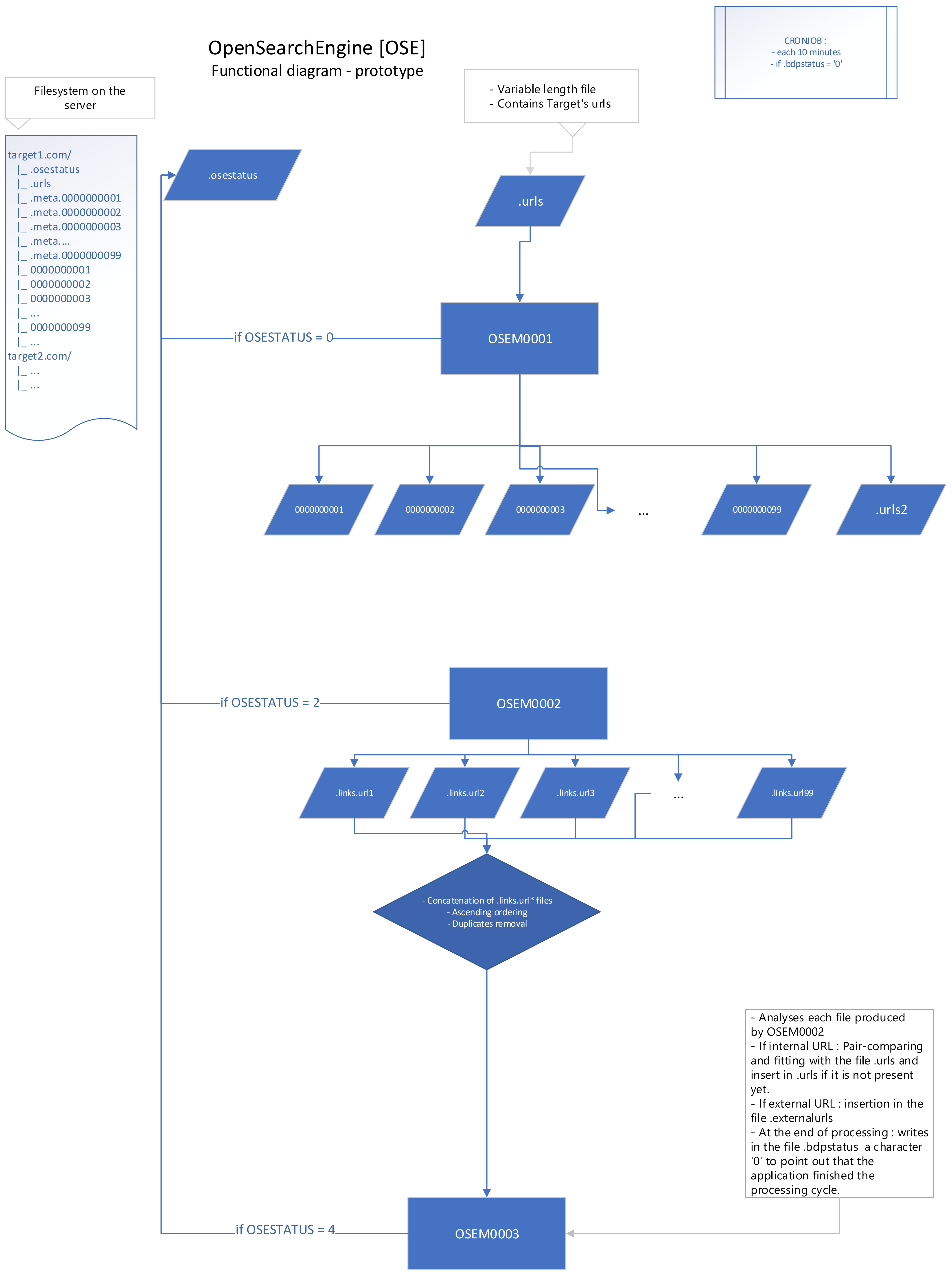
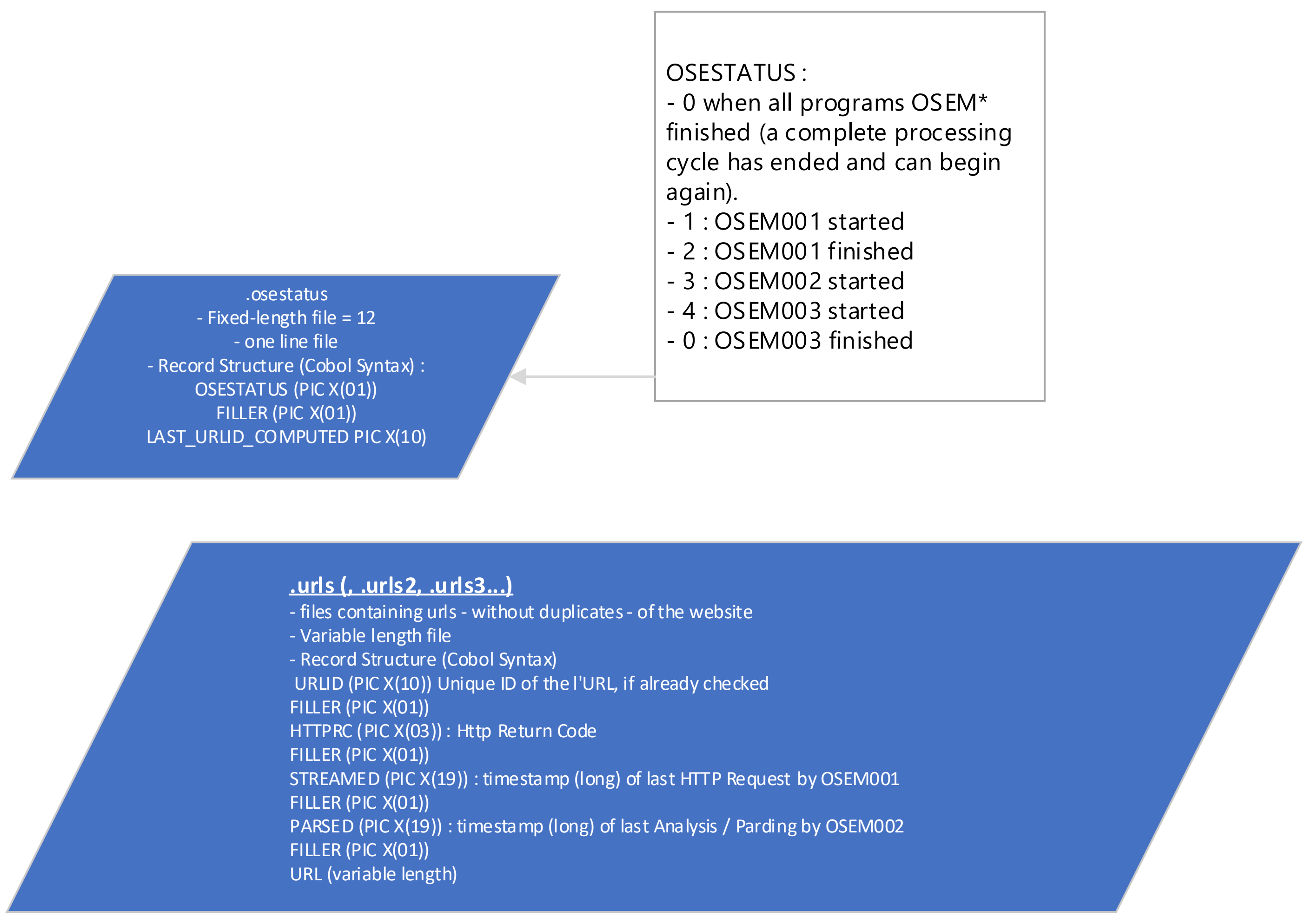
OpenSearchEngine [OSE]

Functional diagram - prototype



```
target1.com/
   |_ .osestatus
   |_ .urls
   |_ .meta.0000000001
   |_ .meta.0000000002
   |_ .meta.0000000003
   |_ .meta....
   |_ .meta.0000000099
   |_ 0000000001
   |_ 0000000002
   |_ 0000000003
   |_ ...
   |_ 0000000099
   |_ ...
target2.com/
   |_ ...
   |_ ...
```

- Analyses each file produced by OSEM0002
- If internal URL : Pair-comparing and fitting with the file .urls and insert in .urls if it is not present yet.
- If external URL : insertion in the file .externalurls
- At the end of processing : writes in the file .bdpstatus a character '0' to point out that the application finished the processing cycle.



.osestatus
- Fixed-length file = 12
- one line file
- Record Structure (Cobol Syntax) :
OSESTATUS (PIC X(01))
FILLER (PIC X(01))
LAST_URLID_COMPUTED PIC X(10)

OSESTATUS :
- 0 when all programs OSEM* finished (a complete processing cycle has ended and can begin again).
- 1 : OSEM001 started
- 2 : OSEM001 finished
- 3 : OSEM002 started
- 4 : OSEM003 started
- 0 : OSEM003 finished

.urls (, .urls2, .urls3...)
- files containing urls - without duplicates - of the website
- Variable length file
- Record Structure (Cobol Syntax)
 URLID (PIC X(10)) Unique ID of the l'URL, if already checked
FILLER (PIC X(01))
HTTPRC (PIC X(03)) : Http Return Code
FILLER (PIC X(01))
STREAMED (PIC X(19)) : timestamp (long) of last HTTP Request by OSEM001
FILLER (PIC X(01))
PARSED (PIC X(19)) : timestamp (long) of last Analysis / Parding by OSEM002
FILLER (PIC X(01))
URL (variable length)

```
- open .osestatus (read-only), reads its content, stores the content in the program memory, and closes
the file
- If OSESTATUS = 0 {
    - get current timestamp
    - open .bdpstatus (write mode), writes the character '1' = job is currently running
    - opens .urls and opens a file .urls2 in write mode
    - For every url in .urls,
      if the Url has not been checked yet (No UrlId in .urls) {
        - wait a random time period between 10 and 30 seconds
        - a unique identifier with 10 characters is computed, for examples 0000000001, with +1
incrementing from the last UrlId read in .osestatus.
        - http GET request on every URL
        - in the file .urls2 :
            - records the UrlId
            - records the HTTP Return Code
            - records Streamed, the current timestamp
            - records the URL
        - in a file 000000000*
             - records the returned data (header, body)
        - The Unique Identifier URLID 000000000* is recorded into .bdpstatus
     }
     we get the current timestamp, we compute the elapsed time (execution time of the program), and log
the result.
}
- writes BDPSTATUS = 2
- End of loop : we close .bdpstatus, .urls and .urls2, and we move .urls2 to .urls
- opens .bdpstatus (read-only), reads its content, records it in the program memory, and then closes
the file
- If BDPSTATUS = 2 {
    - gets the current timestamp
    - opens .bdpstatus (writing mode), writes the character '3'' = the job is currently running
    - opens .urls2 and opens a file .urls3 (writing mode)
    - for every url in .urls2,
      if the url has not been parsed yet (no PARSED in .urls2) and if Url has allready been STREAMED {
        - opens the corresponding 000*** file (read-only mode)
       - opens the file .links.000*** (writing mode)
       - opens the file .links.malformed.000*** (writing mode)
       - extracts every link of the page corresponding to the pattern '<a(.*)href=["'](.*)\["']'
       - for every link found {
           - checks the validity of the link : [http|https]://*.*.*/+(.*)
                - if valid : writes the link in .links.000***
                - else : writes the link in .links.malformed.000***
           }
        - in the file .urls3 :
            - records PARSED, the current timestamp
     }
     - gets the current timestamp, computes the elapsed time (running time of the program), and logs
the result
     - writes BDPSTATUS = 4
     - end of loop : closes .bdpstatus, .urls2 and .urls3, and moves .urls3 vers .urls
}
```