

## Package structure

v1.0

Crawling results are converted into packages before being transmitted or stored. The package is a TLV stream. TLV is an abbreviation from Type Length Value.

The first four bytes of TLV (**Type**) contain transmitted data type. There are assumed following data types:

- 1 – Hostname
- 2 – Website Sitemaps
- 3 – Website Parse Template (WPT)
- 4 – URL
- 5 – HTML Scripts
- 6 – Web Page Internal & External Links
- 7 – Web Page Extracted Content
- 8 – Web Page Parse Result
- 9 – External Hostnames

The next four bytes (**Length**) contain size of transmitted data included in the remained part of TLV (**Value**).

Each package consists of following sequence of TLVs:

TLV1 (type=1)	TLV2 (type=2)	TLV3 (type=3)	TLV4 (type=4)	TLV5 (type=7)	TLV6 (type=5)	TLV7 (type=8)	TLV8 (type=6)	TLV9 (type=9)
---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------	---------------

Sequence	Data Type	Description
TLV 1	Hostname	Hostname the crawling has been performed for.
TLV 2	Website Sitemaps	Sitemap of corresponding host generated by OMFICA Crawler.
TLV 3	Website Parse Template (WPT)	WPT file describing specific web page.
TLV 4	URL	URL of specific web page.
TLV 5	Web Page Extracted Content	Actual content of specific web page detected by OMFICA Crawler.
TLV 6	HTML Scripts	HTML file of specific web page.
TLV 7	Web Page Parse Result	Structured data of specific web page extracted by OMFICA Crawler.
TLV 8	Web Page Internal & External Links	Internal and external links of specific web page.
TLV 9	External Hostnames	Available external hosts.