David RAPIN, 2012-11-30

WEB ARCHIVING

AT ina

(1/3) OVERVIEW

CONTINUOUS CRAWLS

(CRAWLING NEVER STOPPED SINCE 2009)

10 000 WEBSITES

(CRAWLED AT DIFFERENT DEPTHS AND FREQUENCIES)

16 MILLION RESOURCES COLLECTED DAILY

(PAGES: 25%, IMAGES: 60%, OTHER: 15%)

1.6 TB OF DATA CRAWLED DAILY

(BEFORE DEDUPLICATION)

IN ADDITION

65+ GB OF AUDIO/VIDEO CRAWLED DAILY

BY EMBEDDED-MEDIA EXTRACTORS

(YOUTUBE, DAILYMOTION, VIMEO, SOUNDCLOUD, ETC.)

(2/3) STORAGE

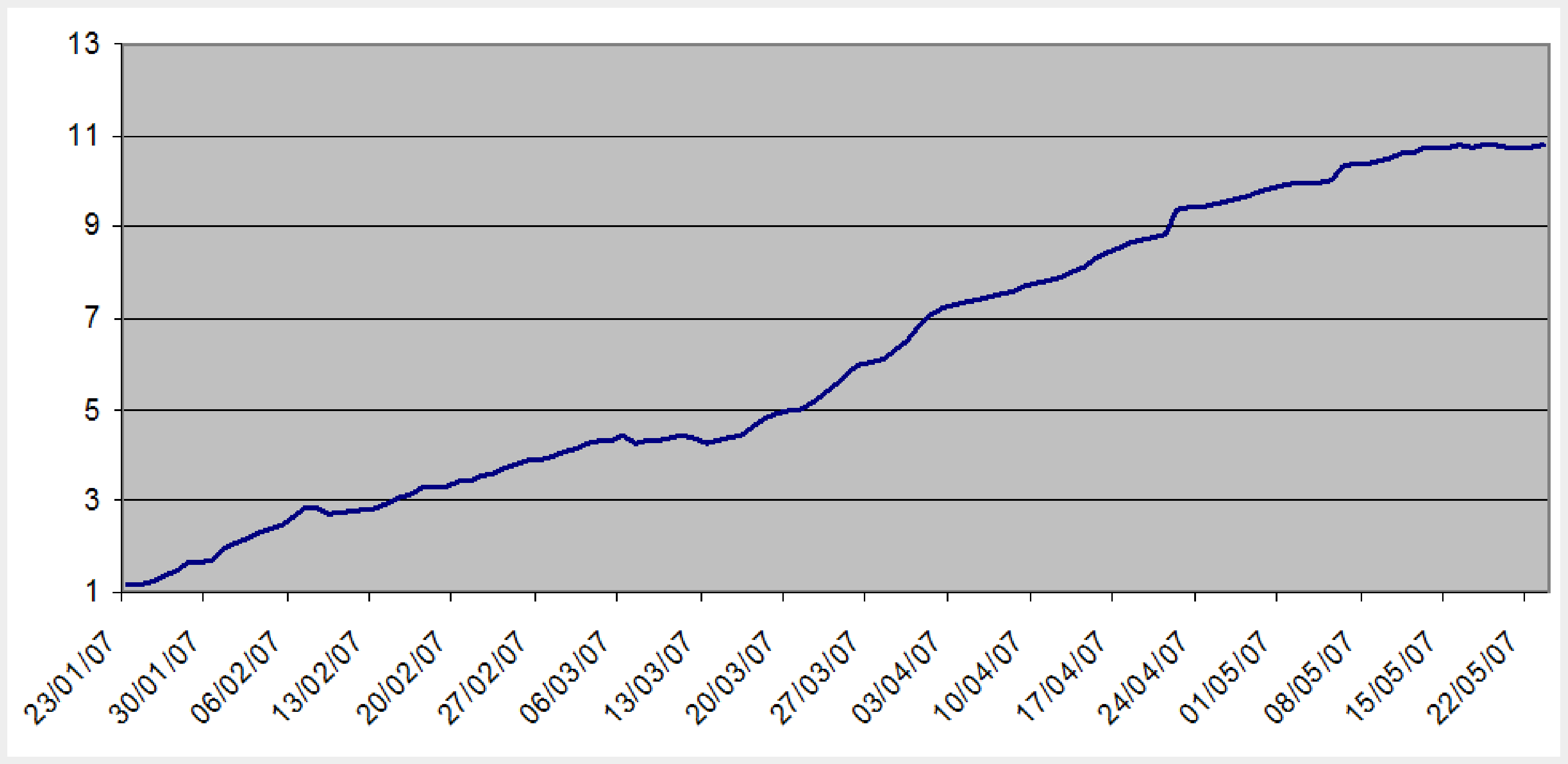
DIGITAL ARCHIVE FILE FORMAT

(SIMPLE, BUILT-IN INTEGRITY CHECK)

CONTENT DEDUPLICATION METHODOLOGY

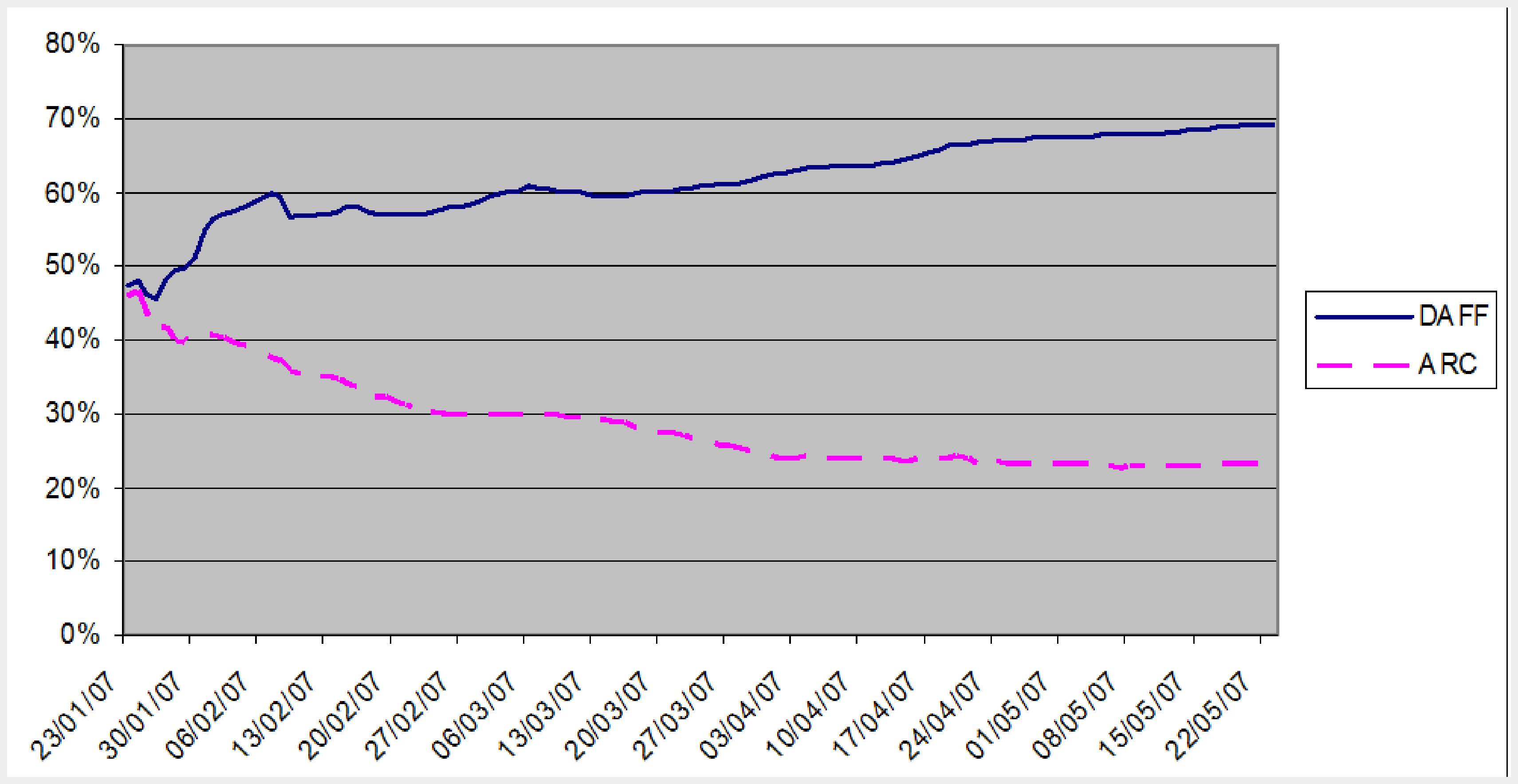
- Data and metadata stored separately
- Data indexed by content signature (SHA256)

DEDUPLICATION RESULTS



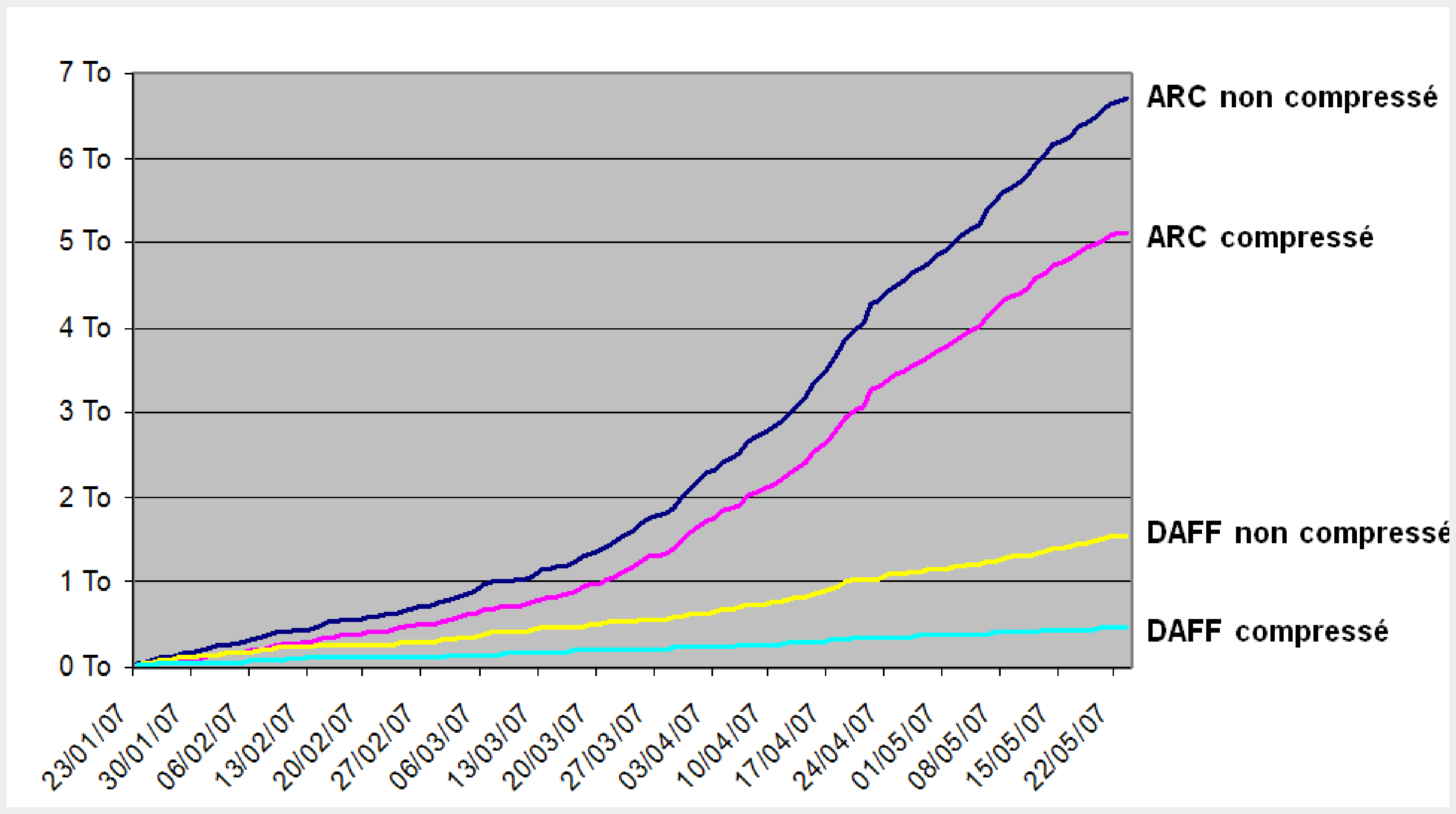
- Observed compression factor: **10**
- Depends on crawl frequencies

DEDUPLICATION EFFECTS ON COMPRESSION



- Higher HTML files proportion
- Better compression factor (**70%**)

STORAGE NEEDS EVOLUTION



(3/3) CRAWLING MODERN WEBSITES

MULTIPLE CRAWLERS FOR DIFFERENT NEEDS

PHAGOSITE

- Historical crawler
- Fast, basic JavaScript link extraction

CROCKET

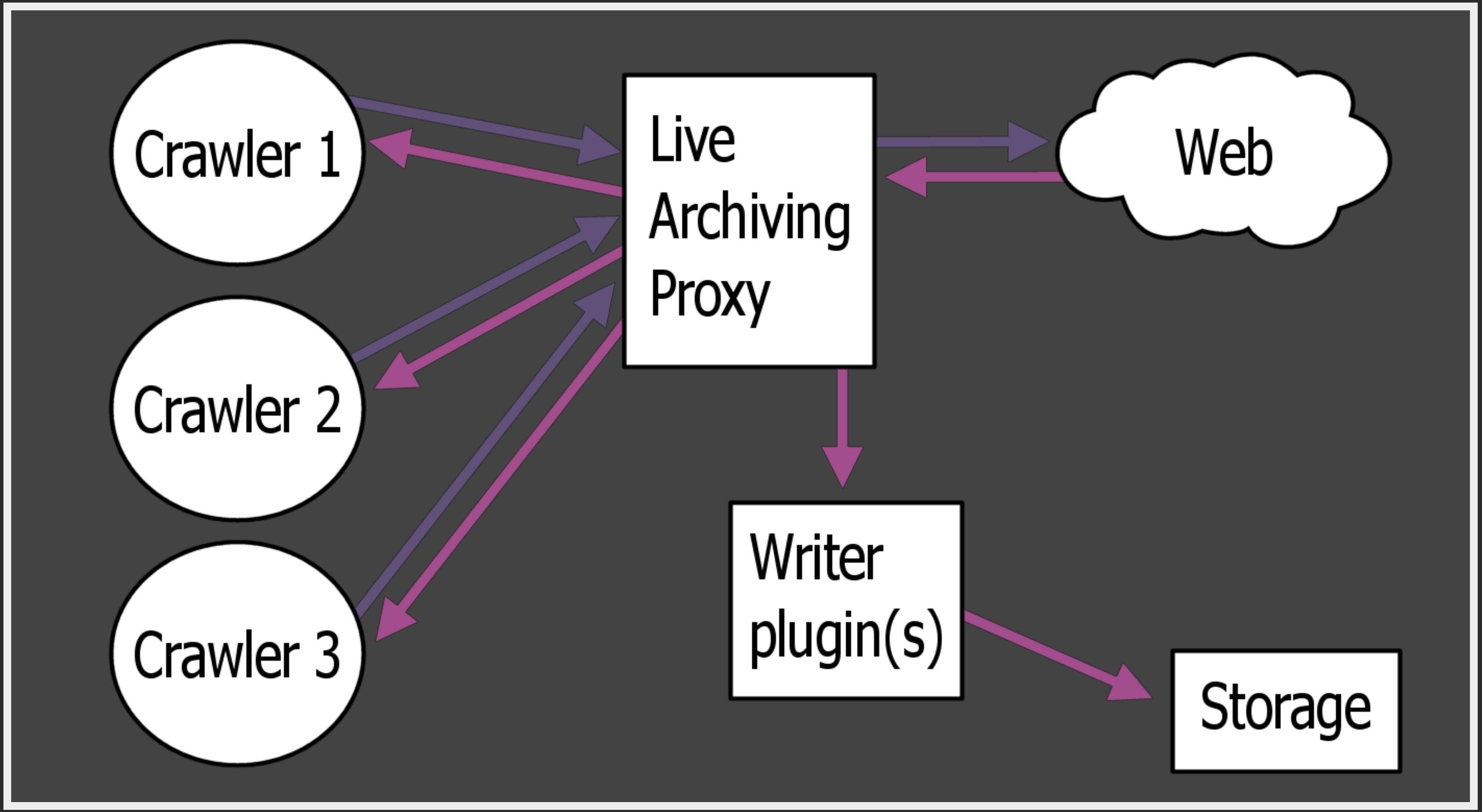
- Firefox based
- Very slow, fully JavaScript enabled

FANTOMAS

- PhantomJS based
- Slow, fully JavaScript enabled

AN INFRASTRUCTURE FOR MULTI-CRAWLER ARCHIVING

LIVE ARCHIVING PROXY



LIVE ARCHIVING PROXY

- One storage backend
- Plug-and-play 'writer' plugins
- No storage on crawl servers
- Multiple writers provide fault tolerance

LIVE ARCHIVING PROXY

AVAILABLE APRIL, 2013

THANK YOU !