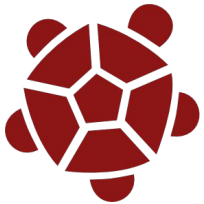
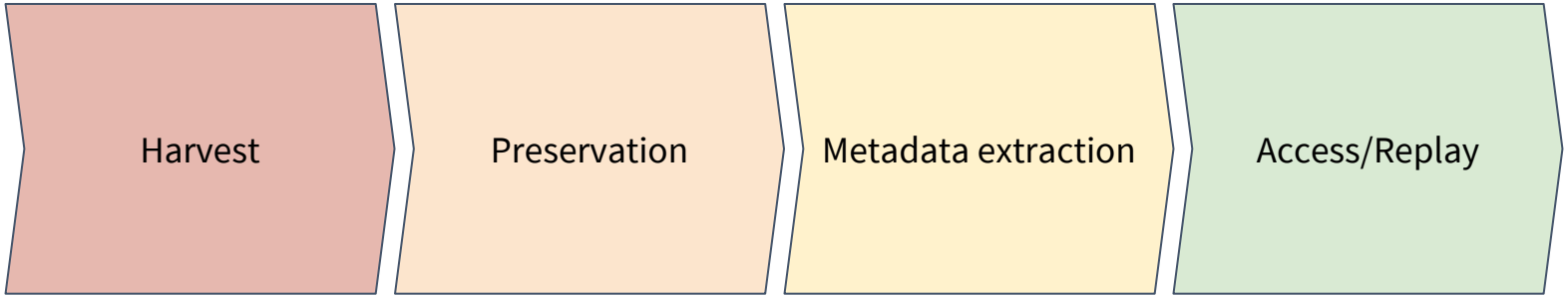


LOCKSS

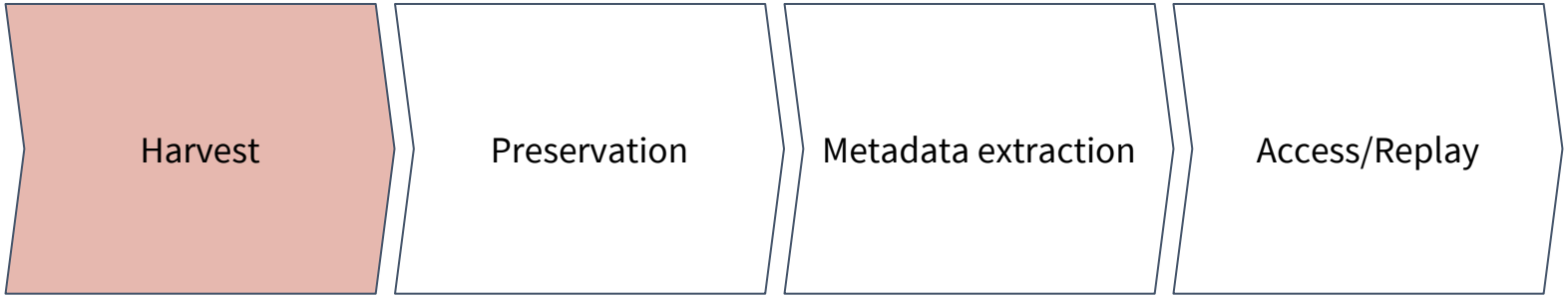
Sifting Needles out of (Well-Formed) Haystacks: Using LOCKSS Plugins for Web Archive Metadata Extraction

Thib Guicherd-Callin – Technical Manager, LOCKSS Program
thib@cs.stanford.edu – github.com/thibgc

A Day in the Life



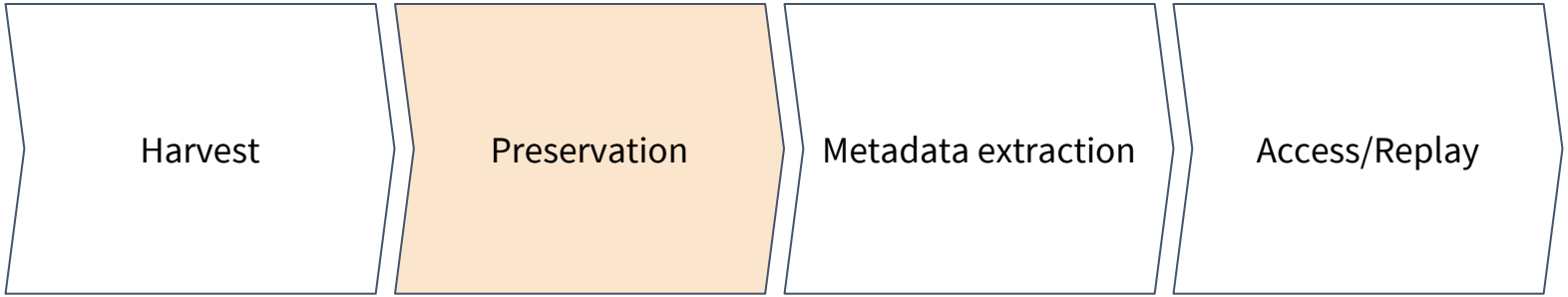
Harvest



- Permission URLs
- Start URLs
- Crawl rules
- HTTP response handlers
- Content validators
- Crawl filters

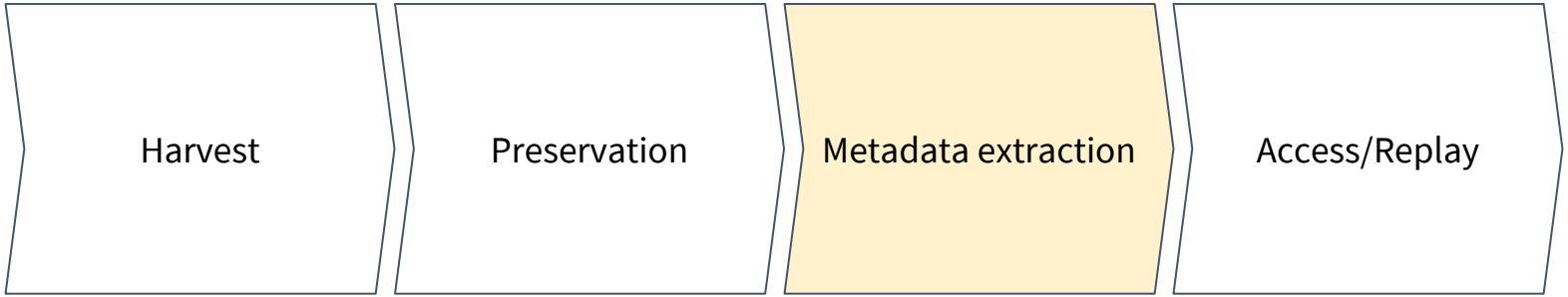
- Link extractors
- URL normalizers
- URL consumers
- Permission checkers
- Substance checkers
- Crawl windows
- Fetch interval controls

Preservation



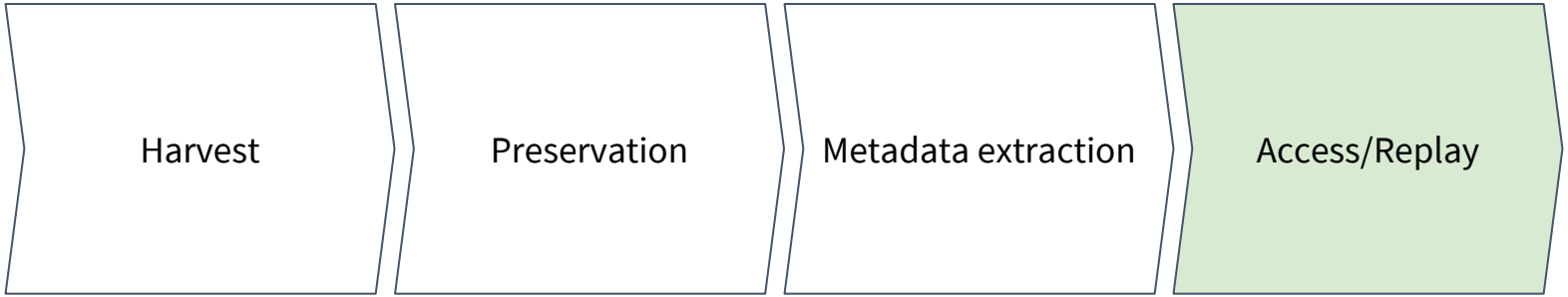
- Hash filters
- URL weighting

Metadata Extraction



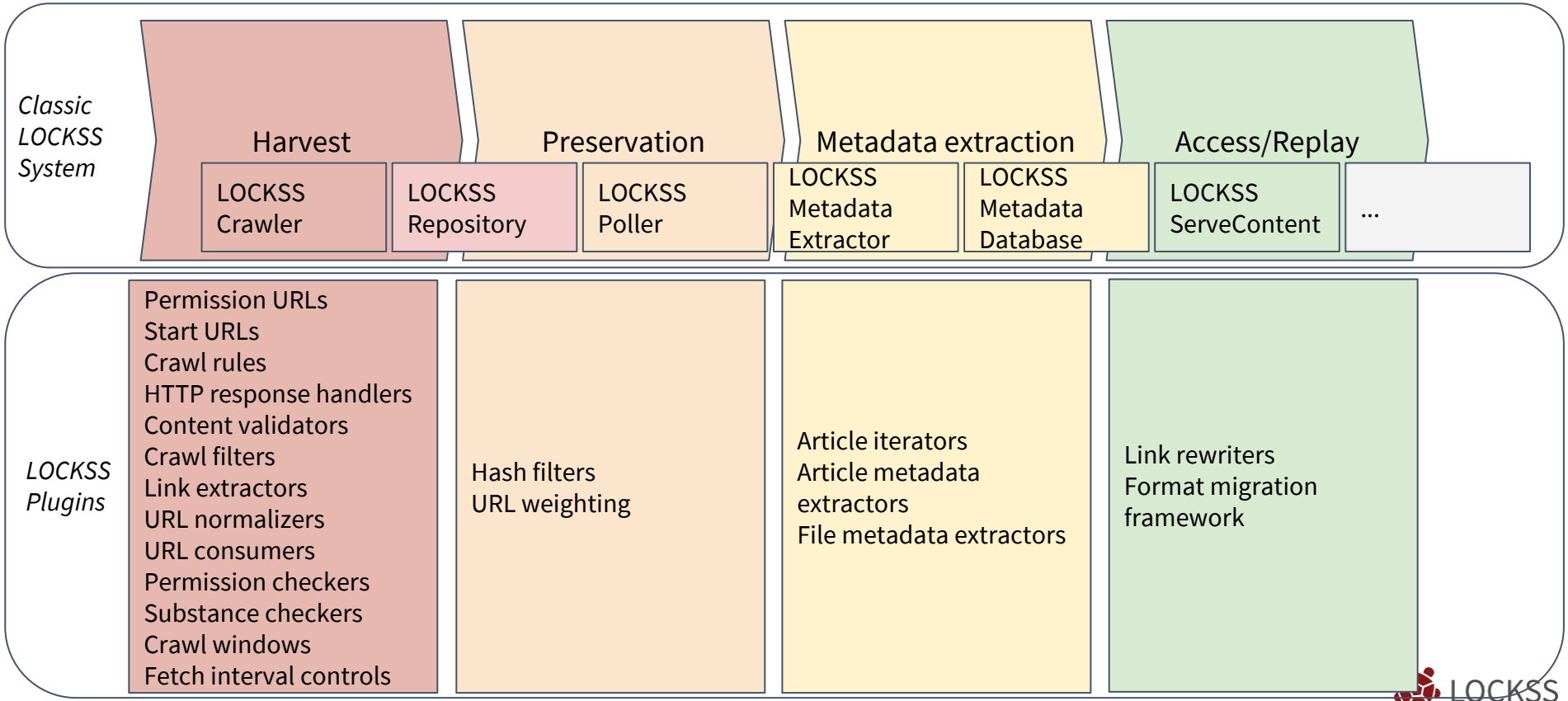
- Article iterators
- Article metadata extractors
- File metadata extractors

Access/Replay

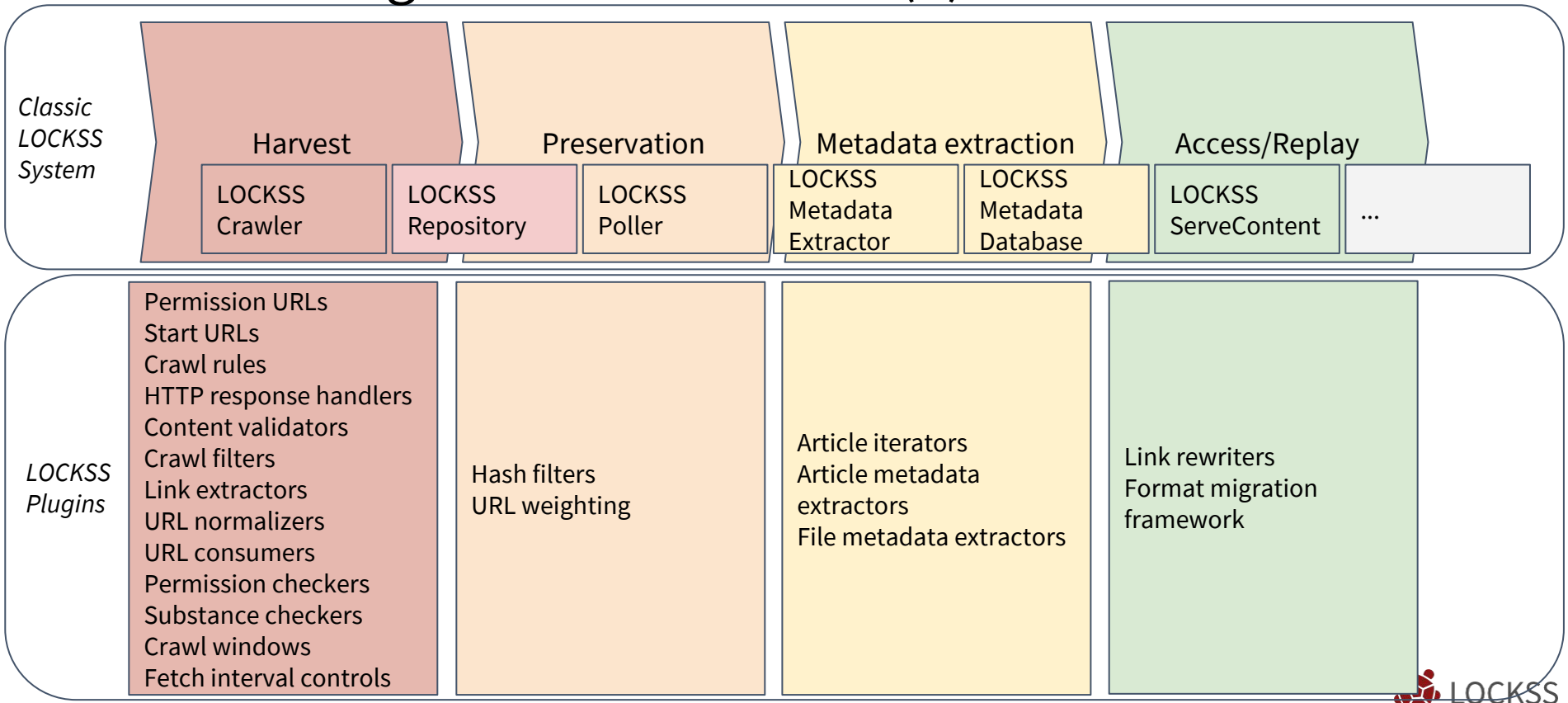


- Link rewriters
- Format migration framework

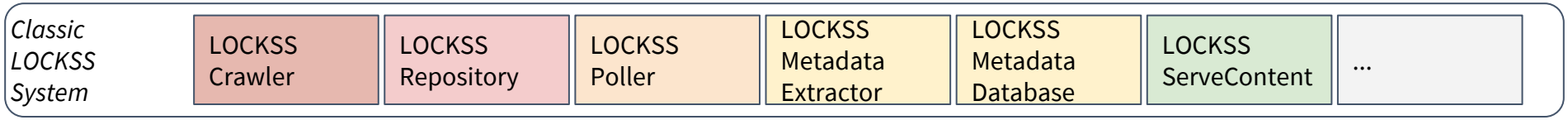
Classic LOCKSS Architecture



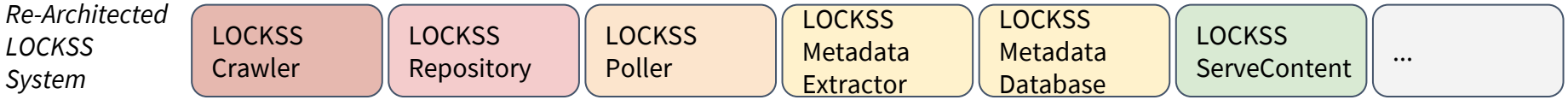
What's Wrong with This Picture? (1)



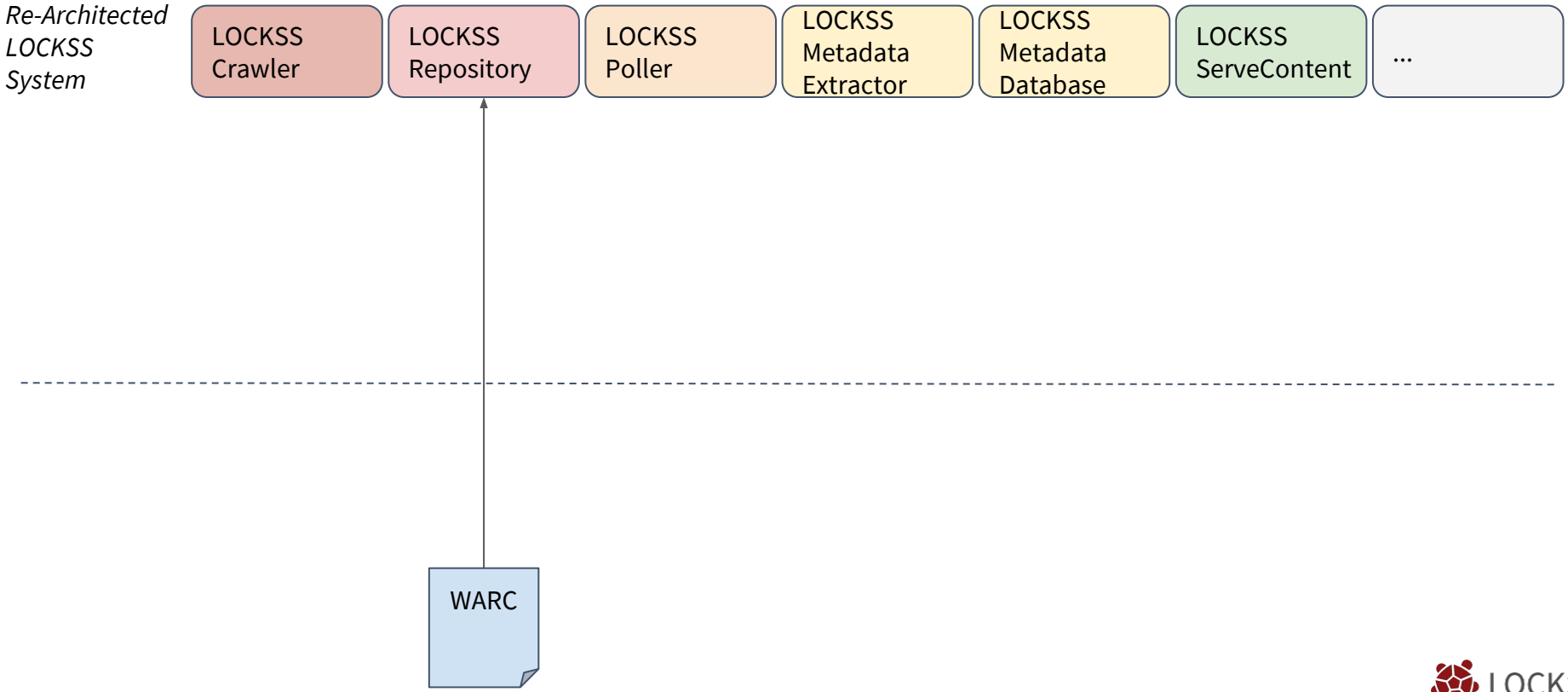
LAAWS Re-Architecture



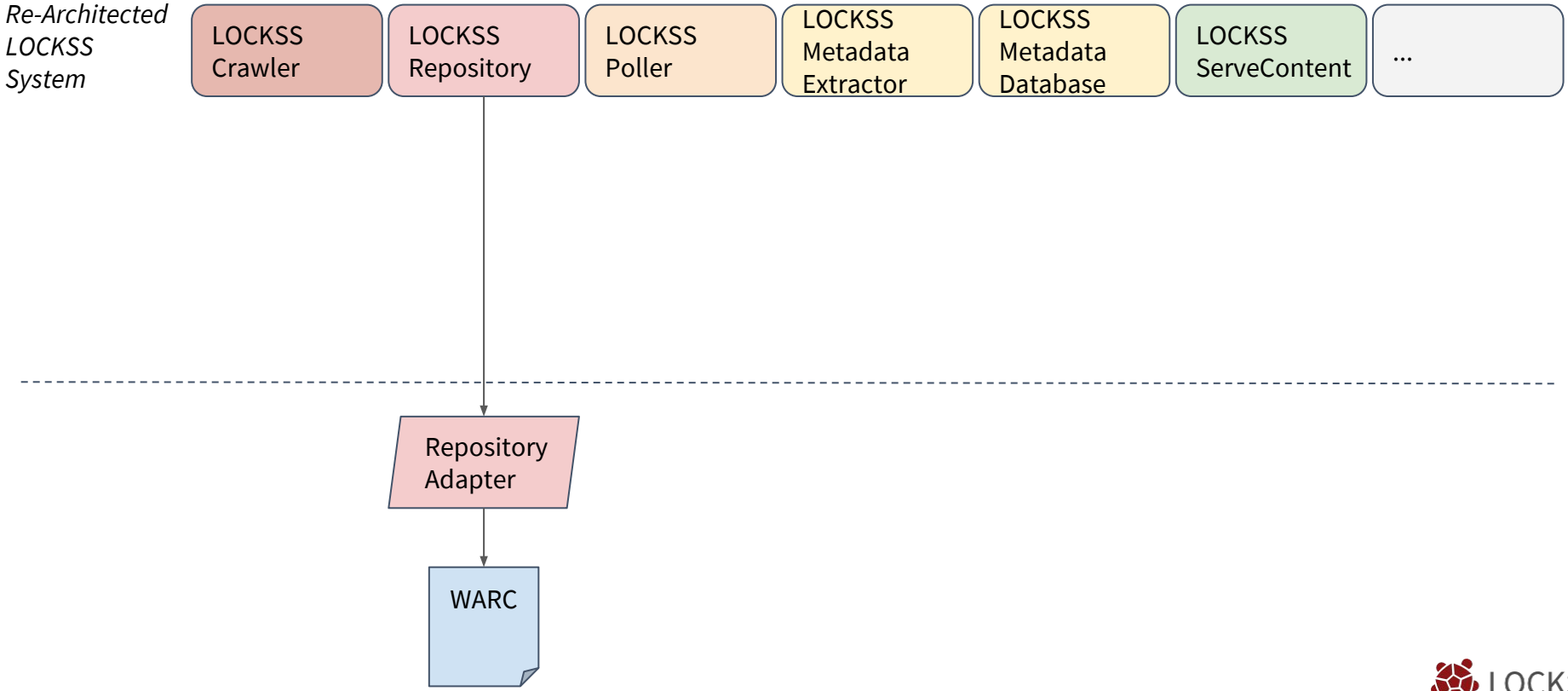
LAAWS Re-Architecture



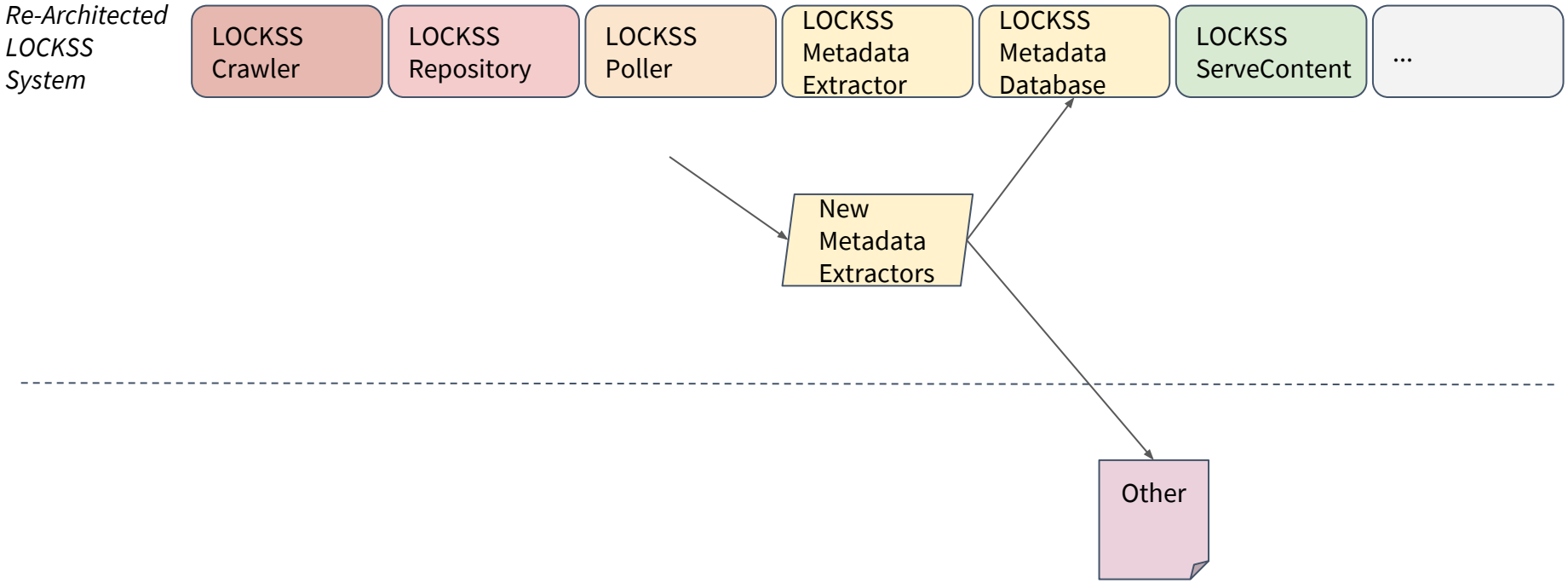
Repository Interoperability



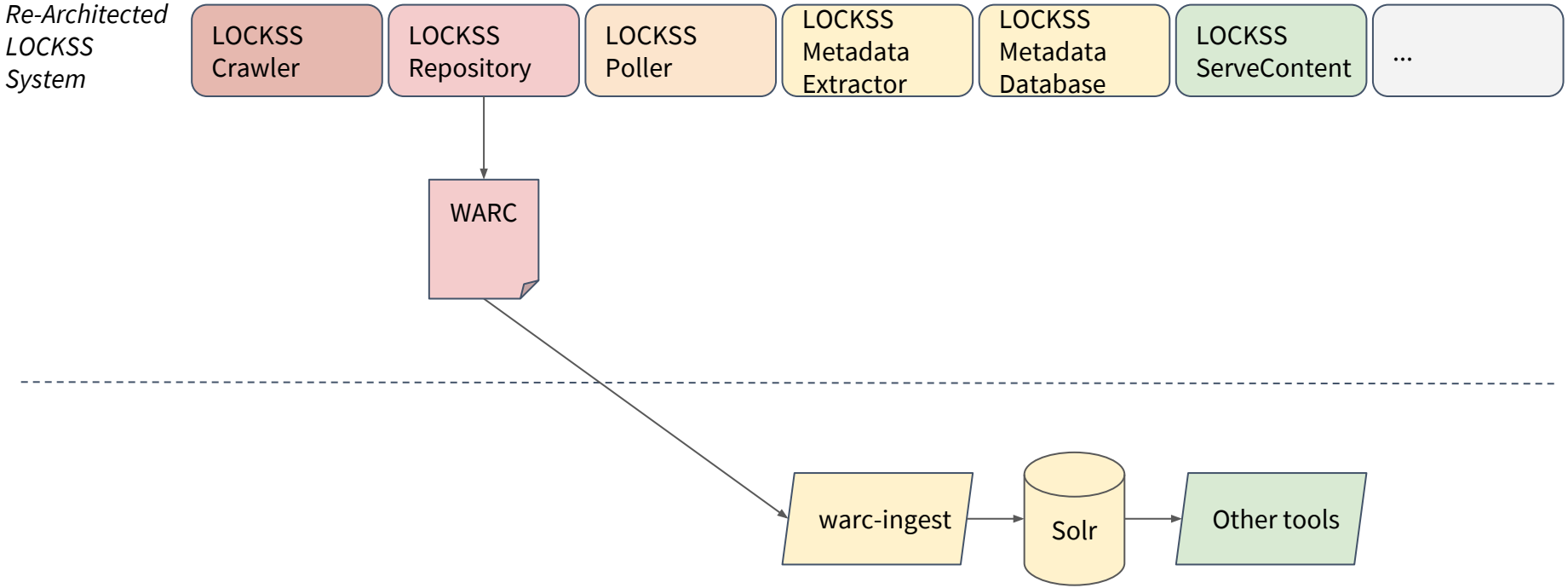
WARC Ingest


Re-Architected LOCKSS System
LOCKSS Crawler
LOCKSS Repository
LOCKSS Poller
LOCKSS Metadata Extractor
LOCKSS Metadata Database
LOCKSS ServeContent
...
Repository Adapter
WARC

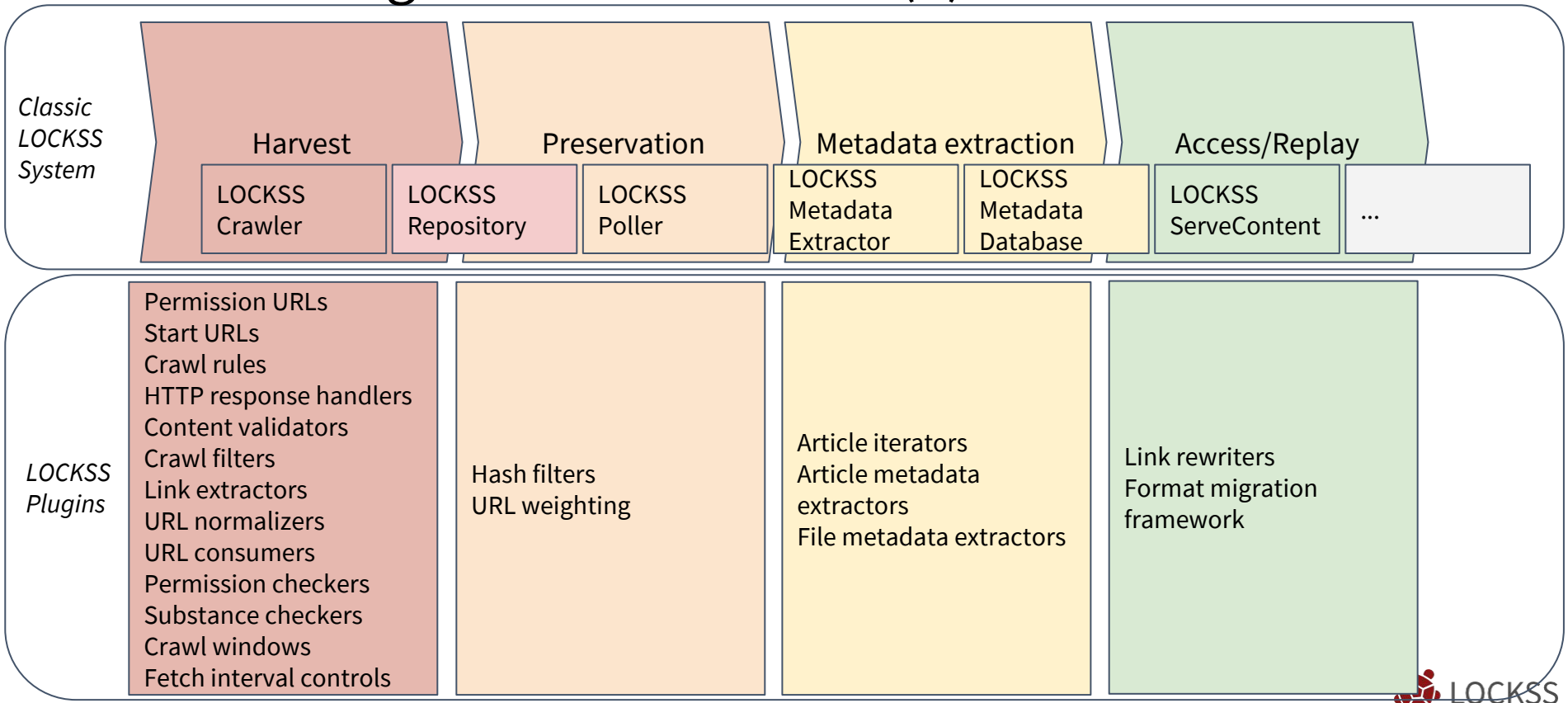
WARC Ingest


Re-Architected LOCKSS System
LOCKSS Crawler
LOCKSS Repository
LOCKSS Poller
LOCKSS Metadata Extractor
LOCKSS Metadata Database
LOCKSS ServeContent
...
New Metadata Extractors
Other

WARC Ingest


Re-Architected LOCKSS System
LOCKSS Crawler
LOCKSS Repository
LOCKSS Poller
LOCKSS Metadata Extractor
LOCKSS Metadata Database
LOCKSS ServeContent
...
WARC
warc-ingest
Solr
Other tools

What's Wrong with This Picture? (2)



Java Tooling from LOCKSS Plugins

- Running metadata extractors externally
- HTML transforms
- PDF transforms
- RIS parser
- Running other things externally:
  - Link extractors
  - Filters
  - Content validators
  - URL normalizers
  - etc.

Thank You

- Nascent community tools:
  - LOCKSS Documentation Portal: https://lockss.github.io/
  - LOCKSS Slack: https://tinyurl.com/slackjoinlockss
- Q&A