Re-Architecture and Collaboration Opportunities: What Unlocking LOCKSS Looks Like

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

Re-Architecture and Collaboration Opportunities: What Unlocking LOCKSS Looks Like

1. LOCKSS At 20:
   Looking in the Mirror
2. Re-Architecture Initiative:
   LOCKSS Inside Out
3. Unlocking LOCKSS:
   The Payoff of Openness

Credits: Linda A. Cicero / Stanford News Service

Credits: Undisclosed to protect the innocent

Re-Architecture and Collaboration Opportunities: What Unlocking LOCKSS Looks Like

1. LOCKSS At 20:
   Looking in the Mirror
2. Re-Architecture Initiative:
   LOCKSS Inside Out
3. Unlocking LOCKSS:
   The Payoff of Openness

The Multiple Facets of LOCKSS

LOCKSS = "Lots Of Copies Keep Stuff Safe"

- A widely-accepted, research-based methodology for long-term digital preservation
- An open source software stack for peer-to-peer digital preservation
- A Stanford University Libraries program focused on digital libraries
- An international community of institutions and digital preservation networks

Credits: Nicholas Taylor

20 Years of Accomplishments

- Used to power successful digital preservation initiatives
  - International: Global LOCKSS Network (GLN), CLOCKSS Archive, MetaArchive Cooperative, SAFE Archive FEderation (SAFE PLN), Public Knowledge Project (PKP) Network, Perma.cc...
  - National: US Docs PLN, Canadian Government Information (CGI) Network, Rede Cariniana...
  - Regional: Alabama Digital Preservation Network (ADPN), WestVault…
- Enduring software platform
- Self-sustaining program
- Unique research-based approaches to digital preservation threats
- Validated principles
- Digital preservation thought leadership

20 Years of Change

- Scope
  - Preservation targets
  - Type of institutions
- Information technology (IT)
  - Software
  - Storage
  - Infrastructure
- Libraries
  - Born-digital assets, institutional repositories
  - Data sets, "big data"
  - Text mining, data mining
  - Machine learning (ML), artificial intelligence (AI)

20 Years of Challenges

Web preservation has been getting exponentially more difficult

- Beyond the dynamic Web
  - From the document Web to the prefabricated Web to the canvas Web
  - One-time access corridors
  - Rich content visualizers
- Web harvest
  - Resource externalization
  - Link obfuscation

← *bit.do was down a lot this week*

- Web replay

20 Years of Challenges

Custodians are increasingly dissociated from IT

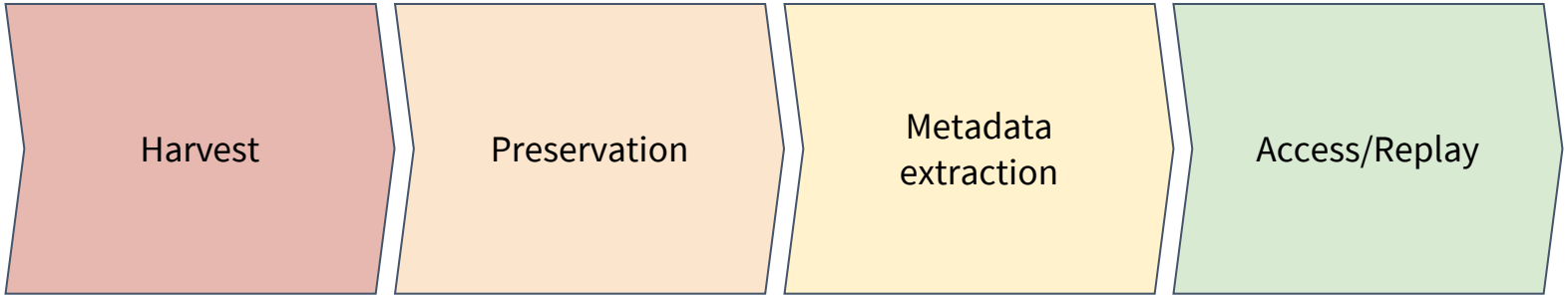
- Library/department IT → institutional IT → outsourced IT
- Appliance → data center → virtualized infrastructure
- Implications
  - Local custody
  - Geographical diversity
  - Hardware and software diversity
  - Operator error
  - External attack
  - Internal attack
  - Economic failure
  - Organizational failure

Reference: *Requirements for Digital Preservation Systems: A Bottom-Up Approach* by Rosenthal et al. (DOI: 10.1045/november2005-rosenthal)

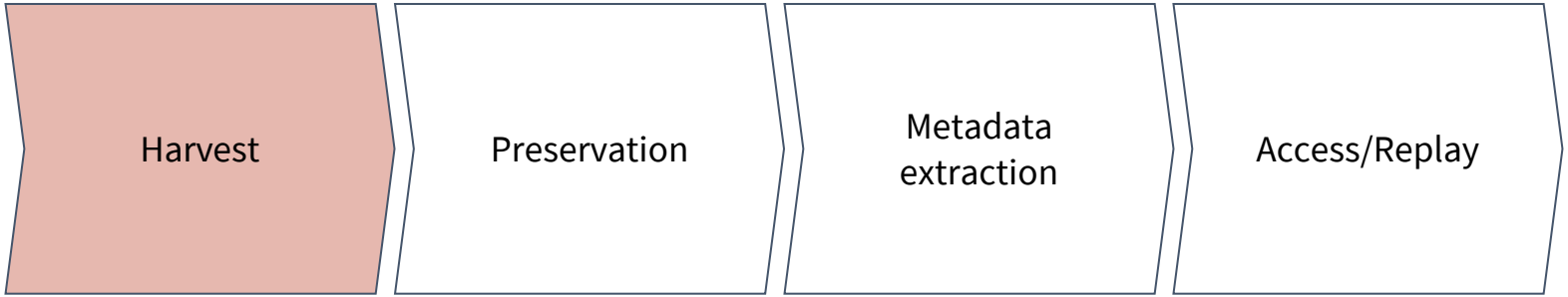
Re-Architecture and Collaboration Opportunities: What Unlocking LOCKSS Looks Like

1. LOCKSS At 20: Looking in the Mirror
2. Re-Architecture Initiative: LOCKSS Inside Out
3. Unlocking LOCKSS: The Payoff of Openness

A Day in the Life



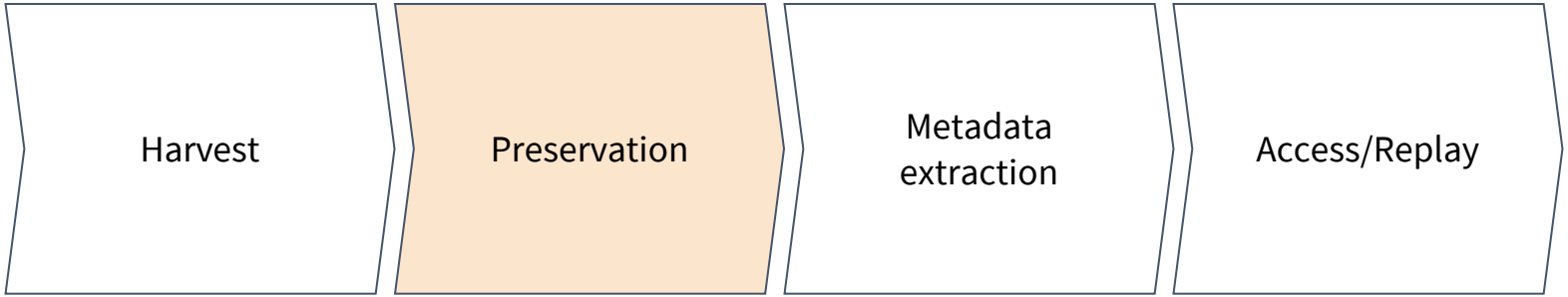
Harvest



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

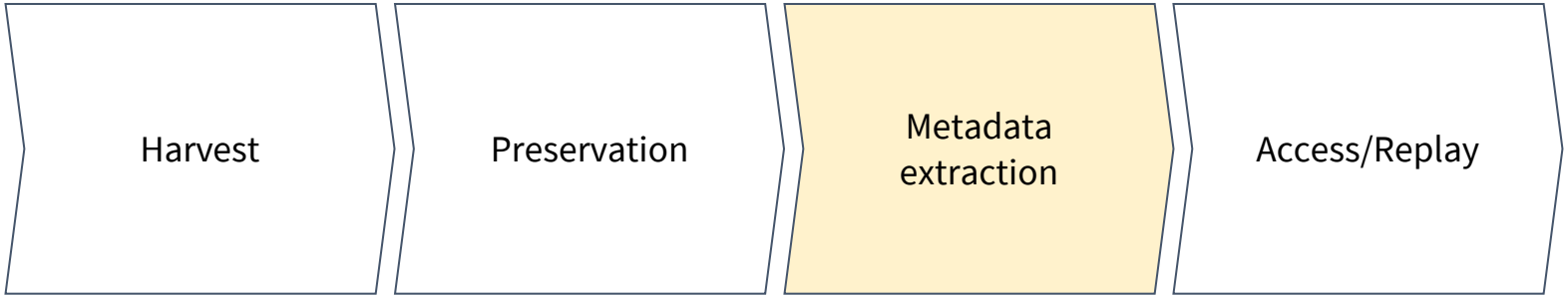
- Link extractors
- URL normalizers
- URL consumers
- Permission checkers
- Substance checkers
- Crawl windows
- Fetch interval controls

Preservation



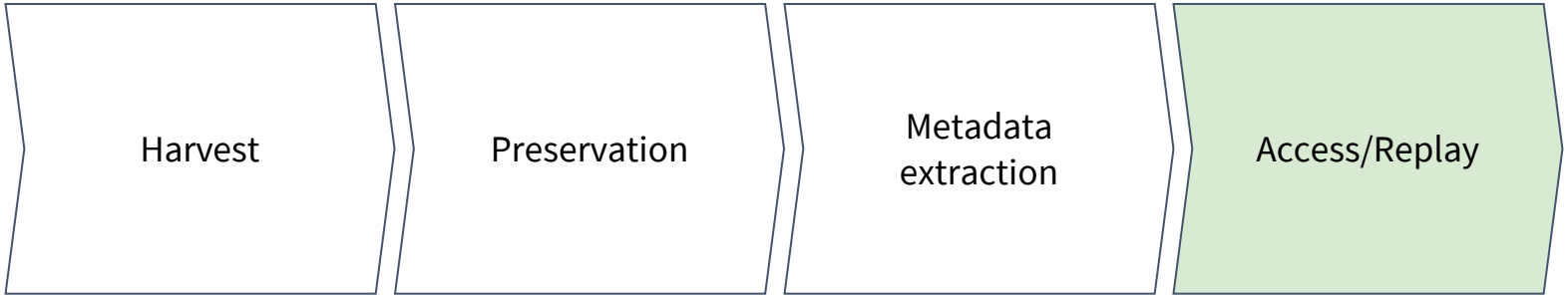
- Hash filters
- URL weighting

Metadata Extraction



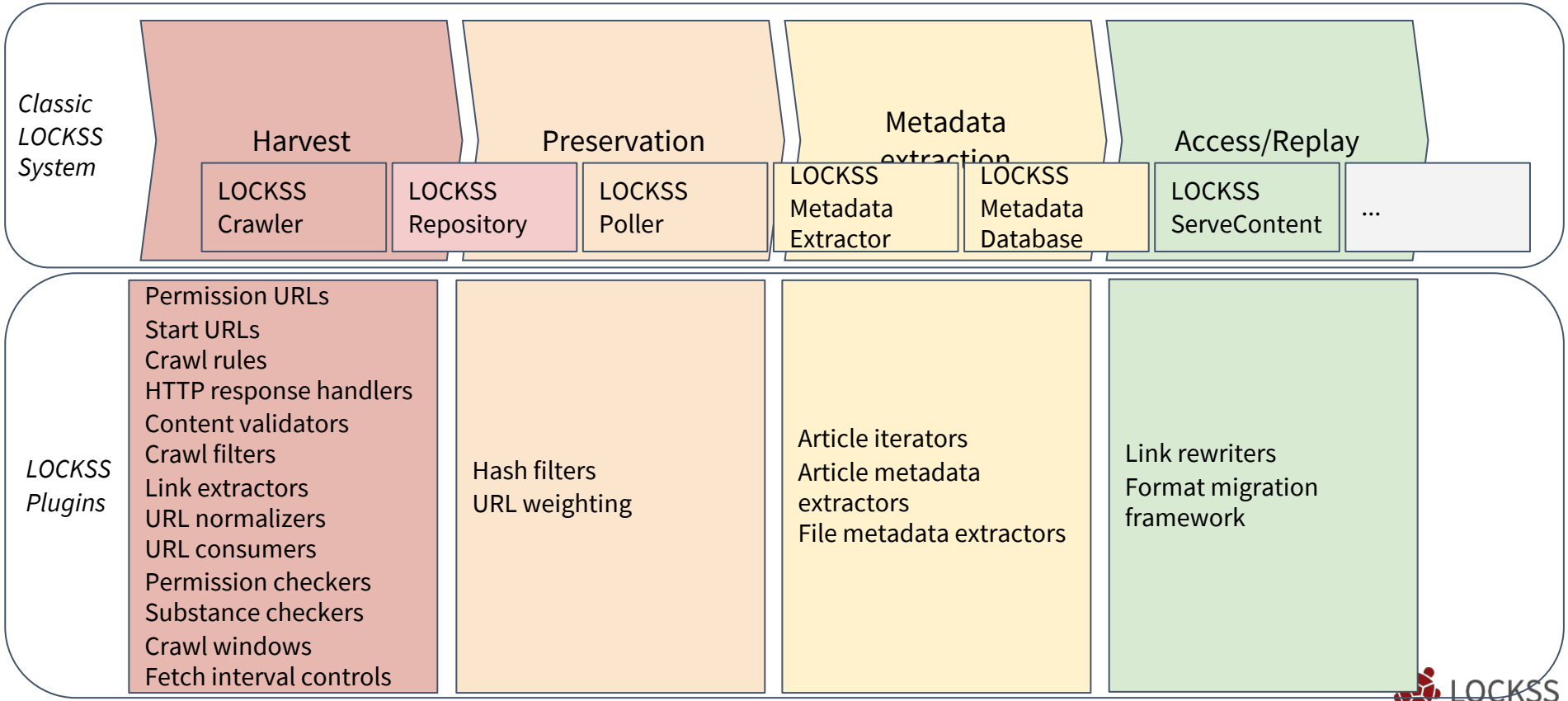
- Article iterators
- Article metadata extractors
- File metadata extractors

Access/Replay

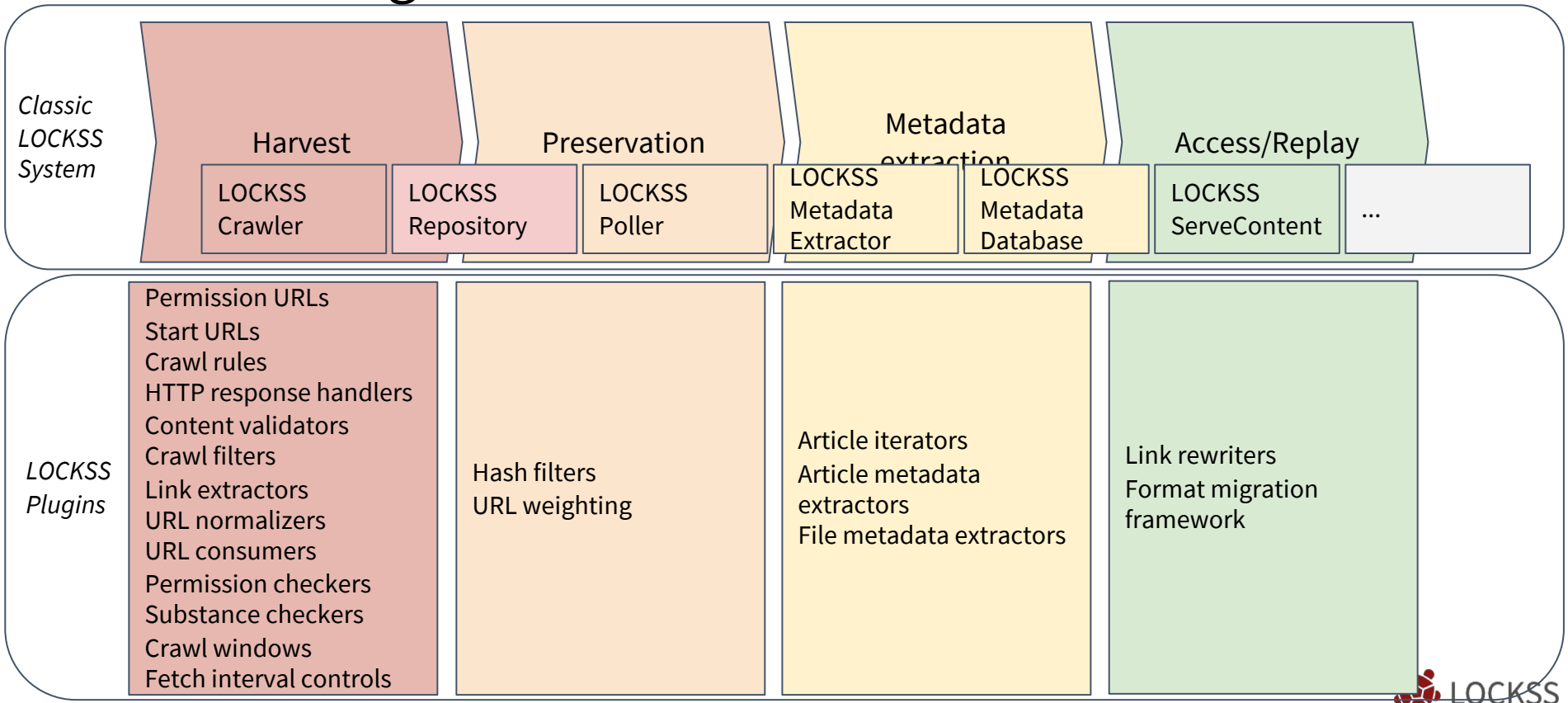


- Link rewriters
- Format migration framework

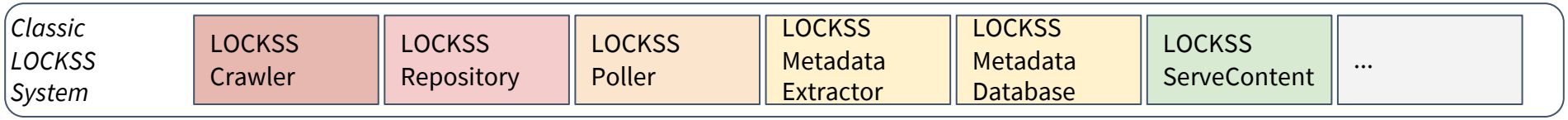
Classic LOCKSS Architecture



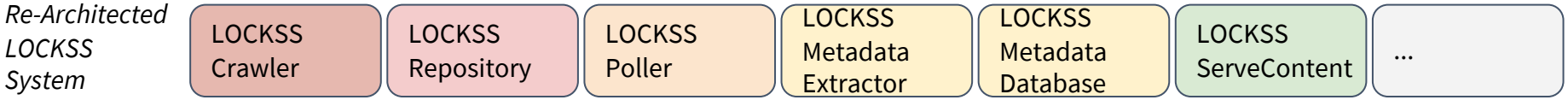
What's Wrong with This Picture?



LAAWS Re-Architecture



LAAWS Re-Architecture



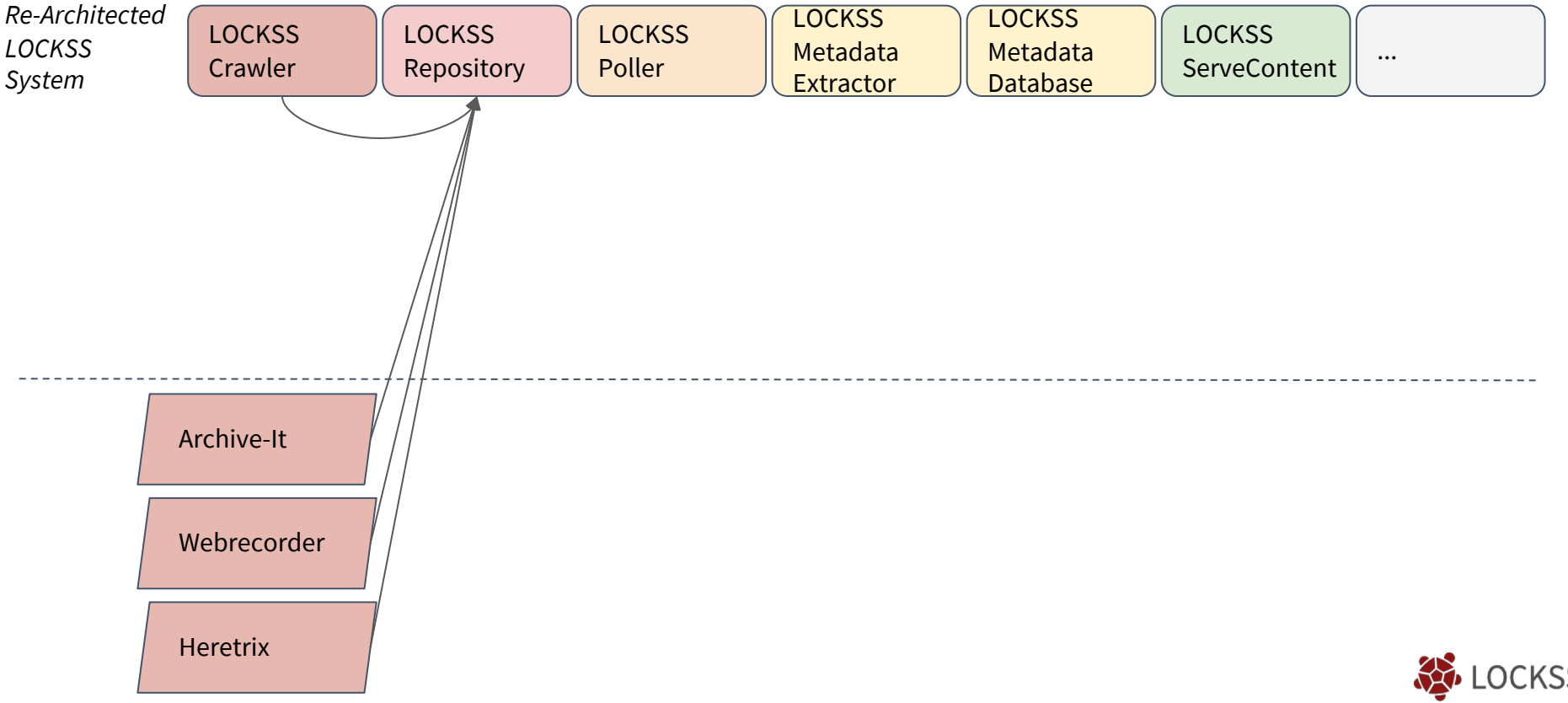
LAAWS Grant and Re-Architecture Initiative

- Mellon Foundation grant, codename LAAWS (LOCKSS Architected As Web Services)
- Software refresh
- Infrastructure modernization

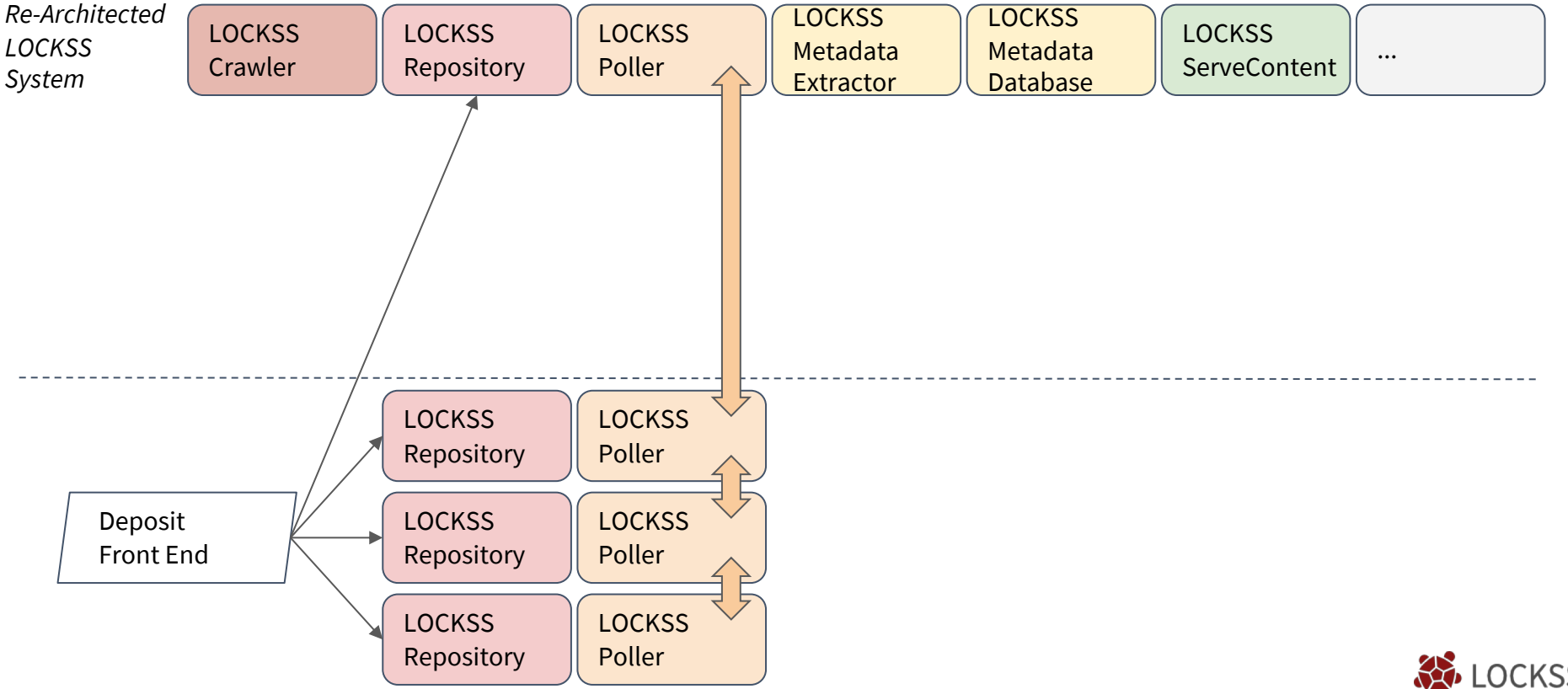
Re-Architecture and Collaboration Opportunities: What Unlocking LOCKSS Looks Like

1. LOCKSS At 20: Looking in the Mirror
2. Re-Architecture Initiative: LOCKSS Inside Out
3. Unlocking LOCKSS: The Payoff of Openness

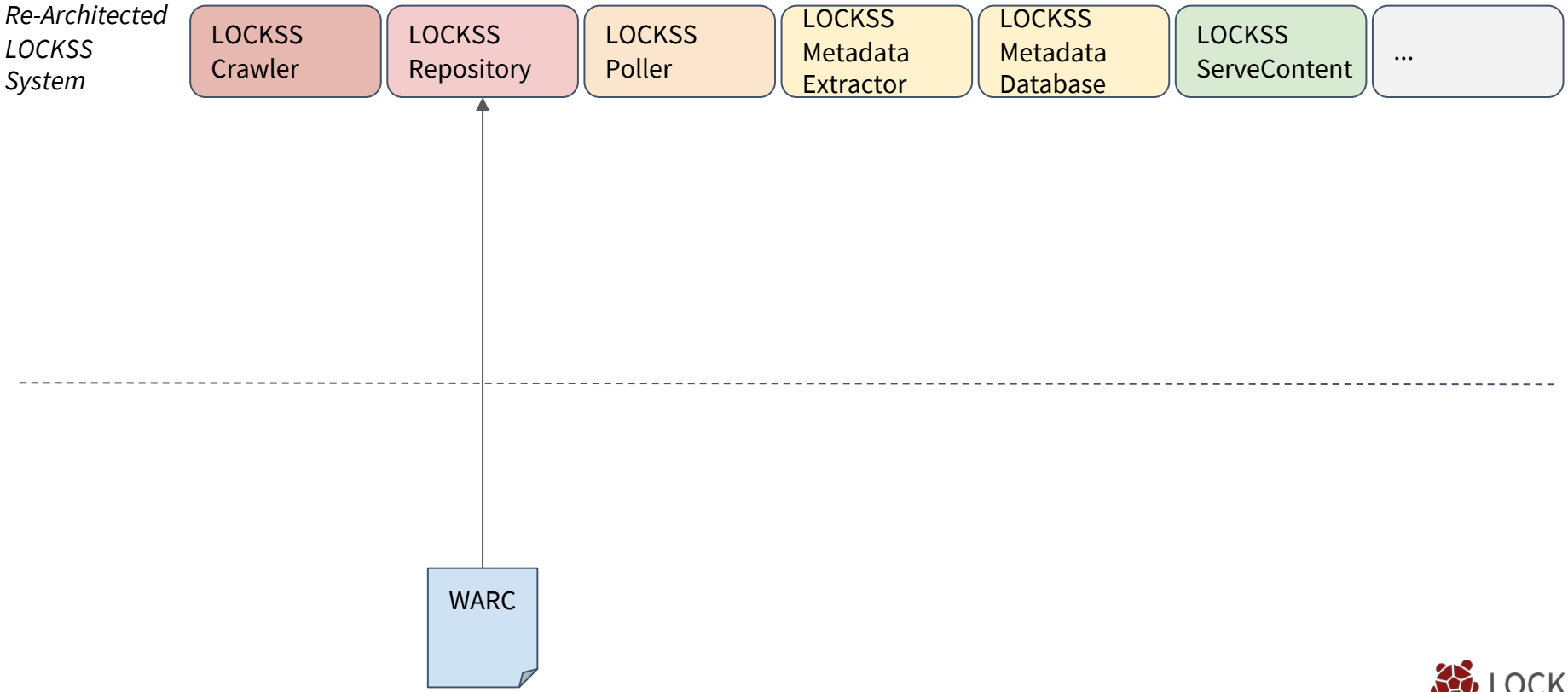
Web Crawling Engines



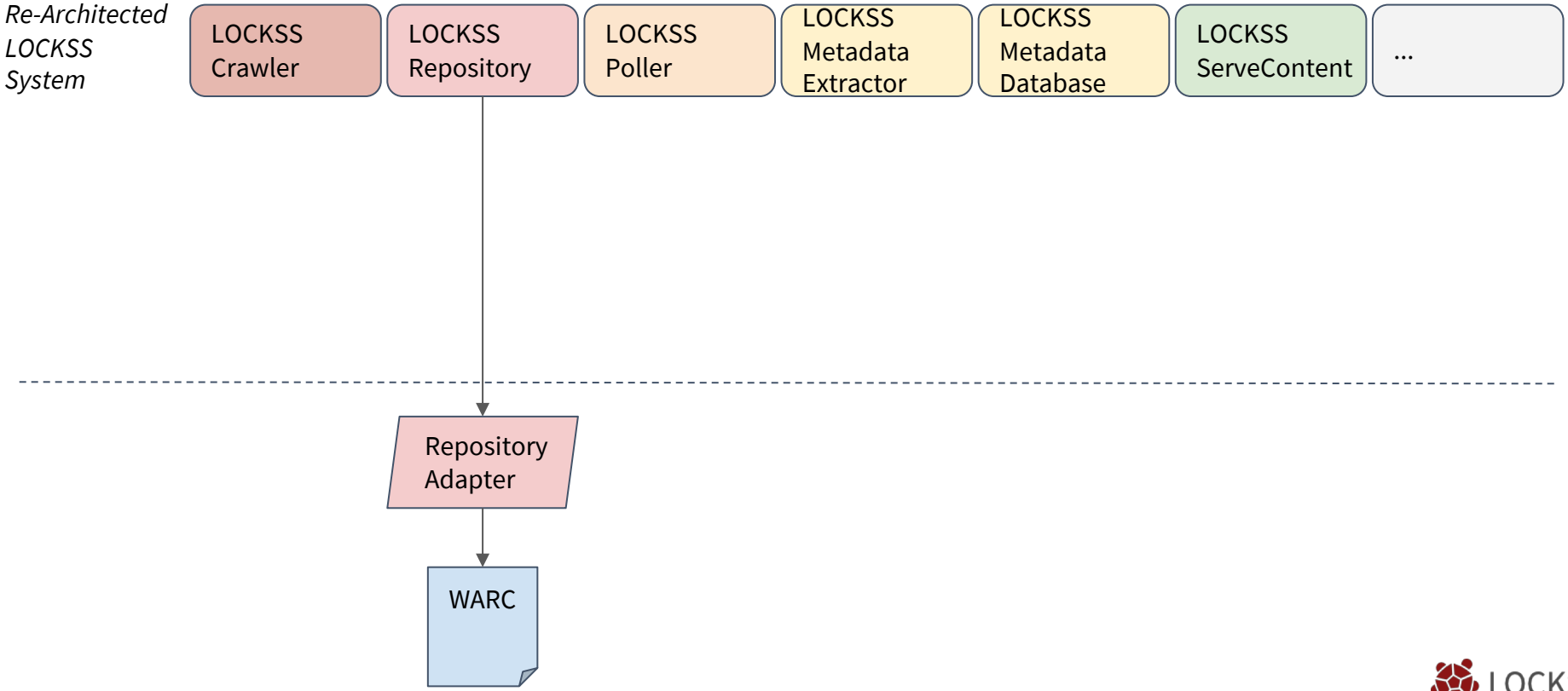
Centralized Deposit



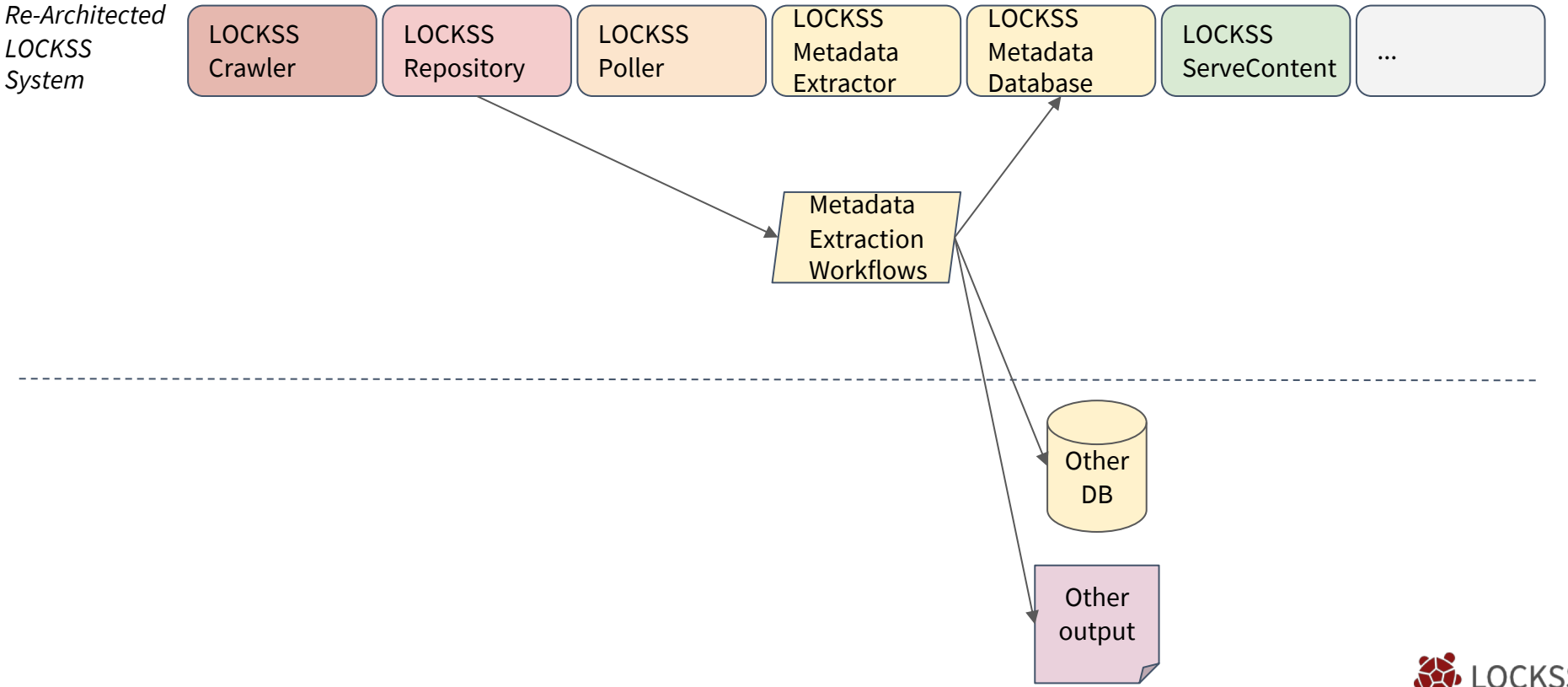
Repository Interoperability



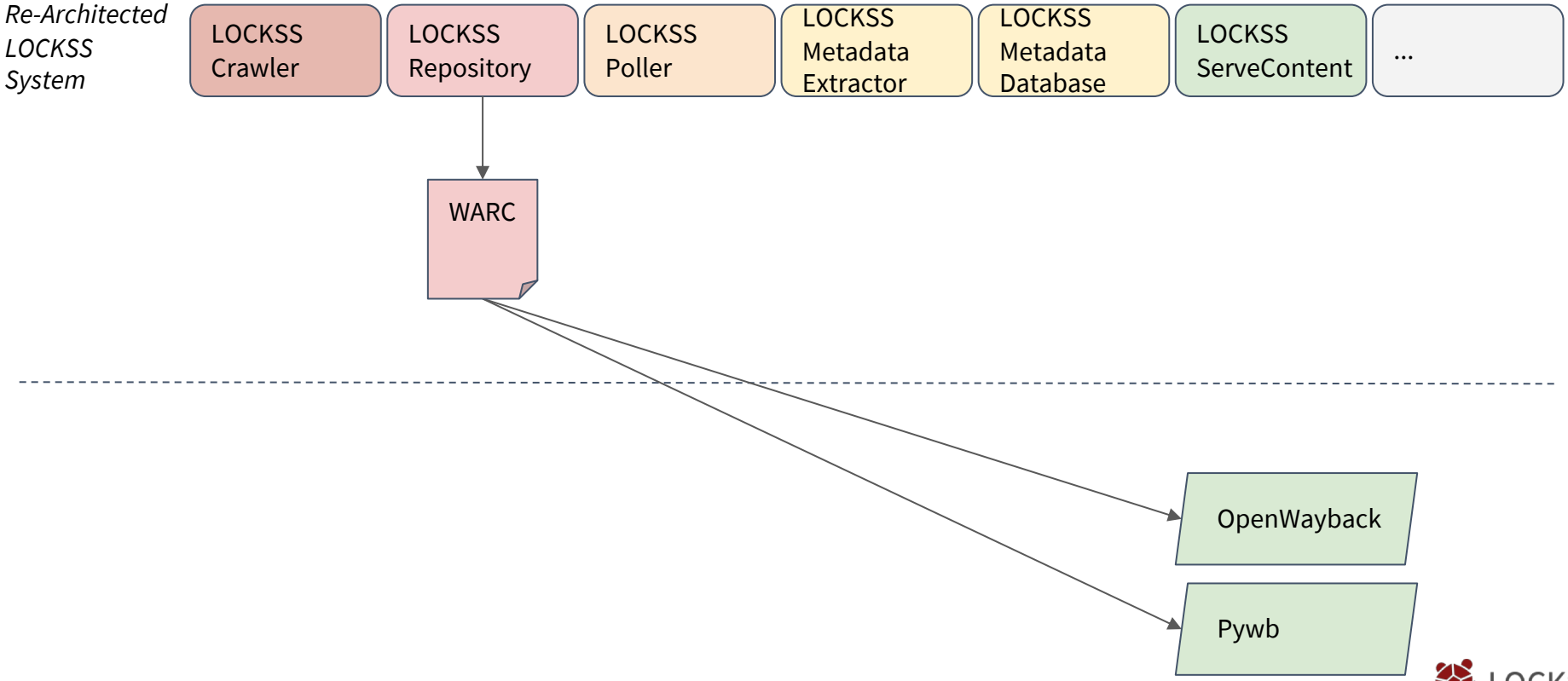
WARC Ingest


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

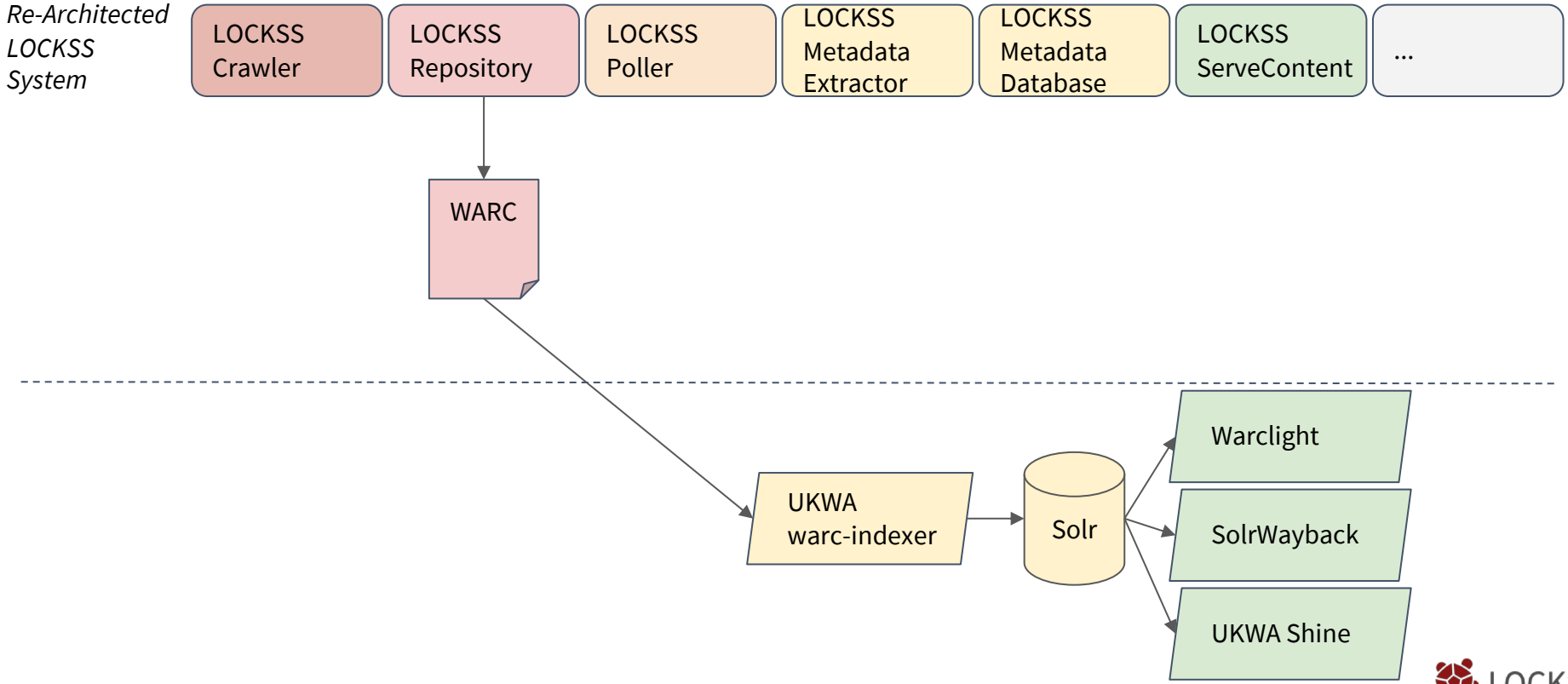
Metadata Extraction Workflows



Web Replay Engines



WARC Ingest


Re-Architected LOCKSS System
LOCKSS Crawler
LOCKSS Repository
LOCKSS Poller
LOCKSS Metadata Extractor
LOCKSS Metadata Database
LOCKSS ServeContent
...
WARC
UKWA warc-indexer
Solr
Warclight
SolrWayback
UKWA Shine

But Foremost

"*Plus ça change et plus c'est la même chose*" -- Jean-Baptiste Alphonse Karr
("The more things change, the more they stay the same")
("*Cuanto más cambia algo, más se parece a lo mismo*")

- University systems
- Related government agencies
- Regional or national consortia
- Neighboring states or countries
- Isolated institutions in the same space

Thank You

- Resources
  - LOCKSS Web site: lockss.org
  - LOCKSS Documentation Portal: lockss.github.io
- Software
  - LOCKSS at GitHub: github.com/lockss
  - LOCKSS at Maven Central: group ID org.lockss
  - LOCKSS at Docker Hub: hub.docker.com/u/lockss
- Communication
  - Twitter: twitter.com/lockss
  - Slack: tinyurl.com/slackjoinlockss
- Q&A
  - Please use the microphone so your question can be translated