Web Archive Data Processing

Karl Stringer (MirrorWeb) &
Grace Bicho (LC)

MirrorWeb

MirrorWeb has been providing web and social media archival for digital heritage and regulatory compliance to governments, banks, regulated firms, and brands since 2017.

We are a Cloud-First company, embracing the AWS stack using Serverless, Containerization, Scalable Databases, Object Storage, Glacier, and more.

Offices in Austin TX, and Manchester UK.

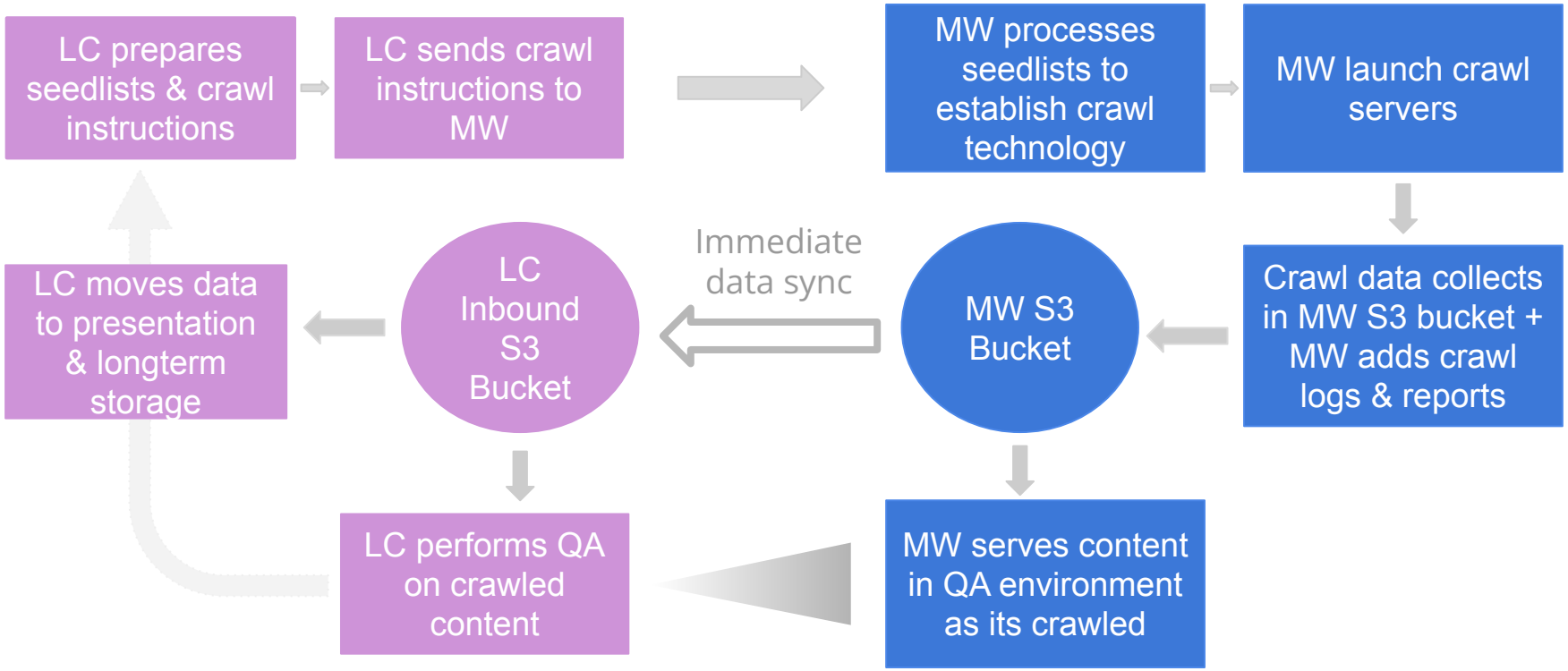
MirrorWeb & LC together since 2021

4 Petabytes collected

Entering Year 3 (TO3)

19.5 Billion URIs

The MW-LC ecosystem (abridged)



Embracing Object Storage

- **Lifecycle management**
  Automatically transition data to different storage classes, and delete after a predetermined time

- **Daily manifests**
  Receive automatic CSVs of bucket contents daily

- **Built-in resilience**
  99.99% availability without managing storage systems

- **Scalability**
  Never worry about finite storage on physical drives

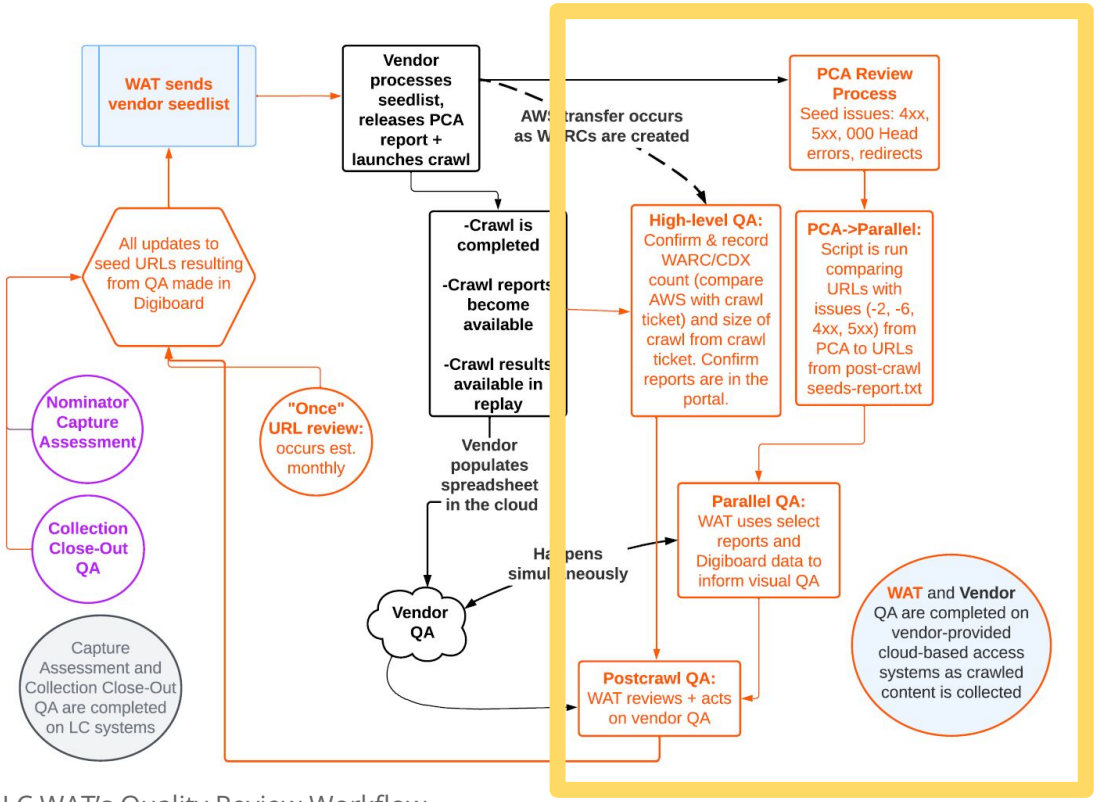
Immediate possession of the data

- **Immediate**
  Cloud delivery - no waiting for transfers
  Use the data the moment it’s collected

- **Collaborative**
  Integrity and quality management is a shared transparent endeavour.
  Spot issues immediately, not at the end of a contract task order (end of TOs are a breeze because the content is already there!)

- **Trust**
  Relationships based on trust and transparency; no withholding data, and no masking of issues

Systematic quality review



LC WAT's Quality Review Workflow

- WAT's quality review workflows based on immediacy of crawled data availability
- Timely reactions to crawl issues provide a feedback loop between ongoing crawls
- Ensures we are only collecting what was selected for the collection

Streamlining the process

- **Client-controlled data**
  Contracting party's object storage as Primary provides a single source of truth. Simplifies contracts.

- **Automation**
  Enhancing automated pipelines for event-based triggers and workflows

- **Third-party integrations**
  Direct integration into visualization services
  Immediate surfacing of data outside of the MW-LC service umbrella

Karl Stringer

Chief Data Officer, MirrorWeb

karl@mirrorweb.com

Grace Bicho

Web Archiving Team, Library of Congress

grth@loc.gov