Webrecorder

Preserving Scalar Based Works

Ilya Kreymer

Web Server Preservation + Web Archives

- **Preservable Web Server (Containerized in Docker, no external dependencies)**
  - Can have full fidelity of server software, but only from one web server
- **Web Archive (WARC files + replay system)**
  - Can crawl any website, but may not get everything
- **Combining for ‘best of both worlds’!**
  - Custom routing merges preserved server and web archive

Web Server + Web Archive Routing

Web Archive + Server System Routing System (simplified):

- **http://webarchive.example.com/replay/http://myscalar.example.com/**
  - Routed to Server (in Docker container)
- **http://webarchive.example.com/replay/http://youtube.com/**
  - Routed to web archive replay

Scalar - (Digital Publishing Platform, Open Source)

- **Great Import/Export capability, allows importing projects into new server**

Also: Preservable Browsers (oldweb.today system)

How it works: Scalar + Web Archive Capture

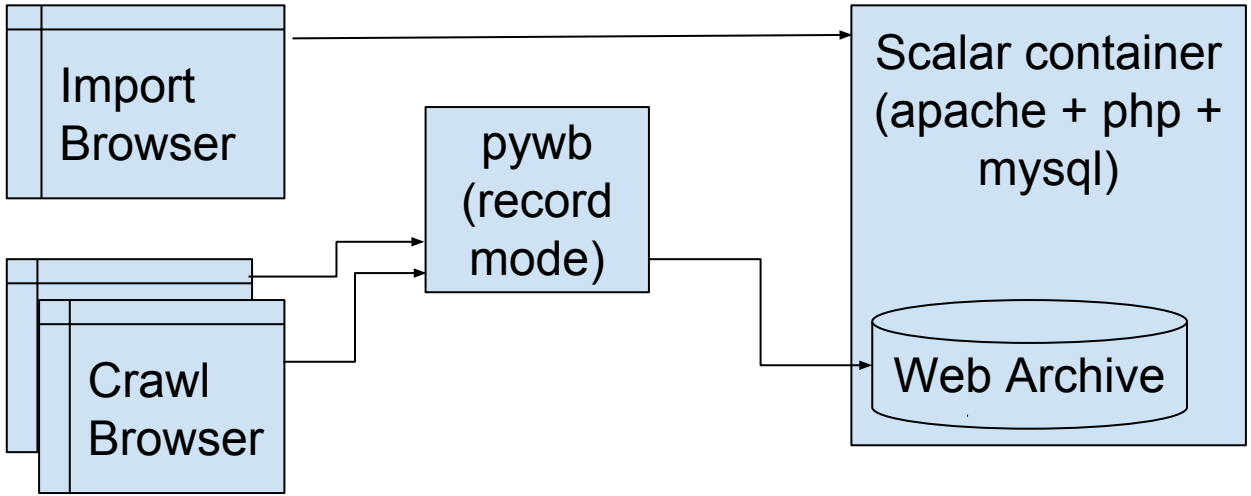
1. **Start new Docker container with empty Scalar install**
2. **Automate browser to use Scalar import plugin to import external Scalar project into containerized Scalar.**
3. **Use Scalar API to determine what pages contain external embeds**
4. **Start 4 automated browsers to load pages with external content through a web archiving proxy, creating WARC files**
5. **Copy WARC files into original the Scalar docker instance**
6. **Commit container to make new Docker image**

How it works

Video of archiving using https://scalar.webrecorder.net

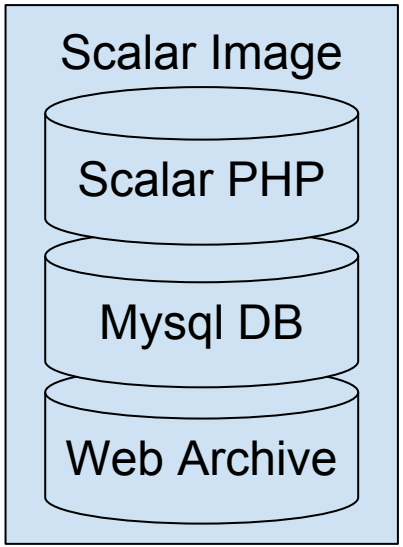
Capture Process

To capture an instance, 4+ Docker Containers are needed:



Preserved Image

- **The web archive data (WARCs, other metadata) and Scalar mysql database are stored on a single Docker container**
- **A new image is creating, preserving the WARCs + Scalar data + mysql db in one image.**



Access Process

To access the preserved system, 3 Docker images are launched from the single Scalar image: