NLM Repository Migration Considerations

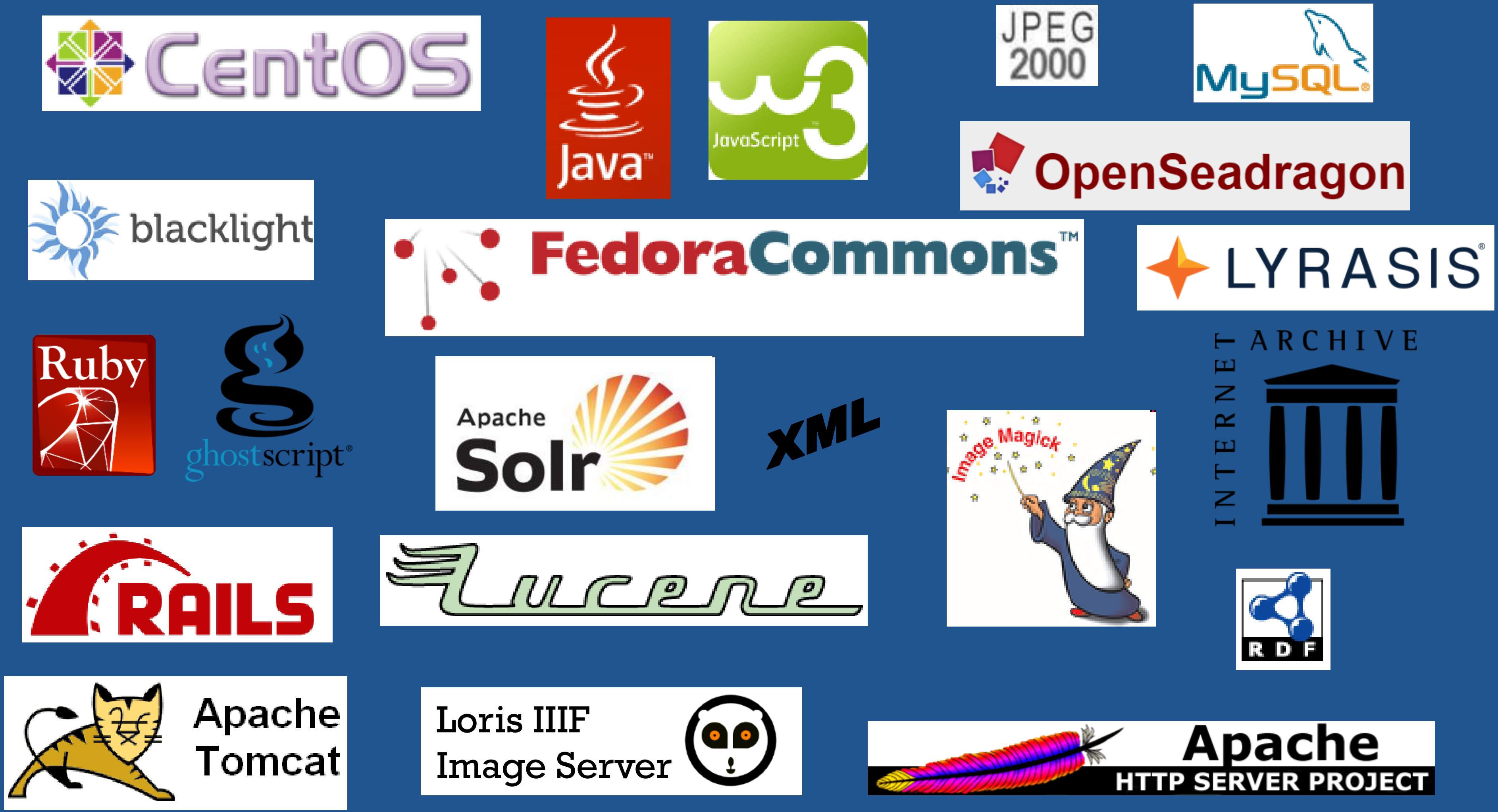
Doron Shalvi
Office of Computer and Communication Systems
National Library of Medicine
National Institutes of Health
U.S. Department of Health and Human Services

August 2020

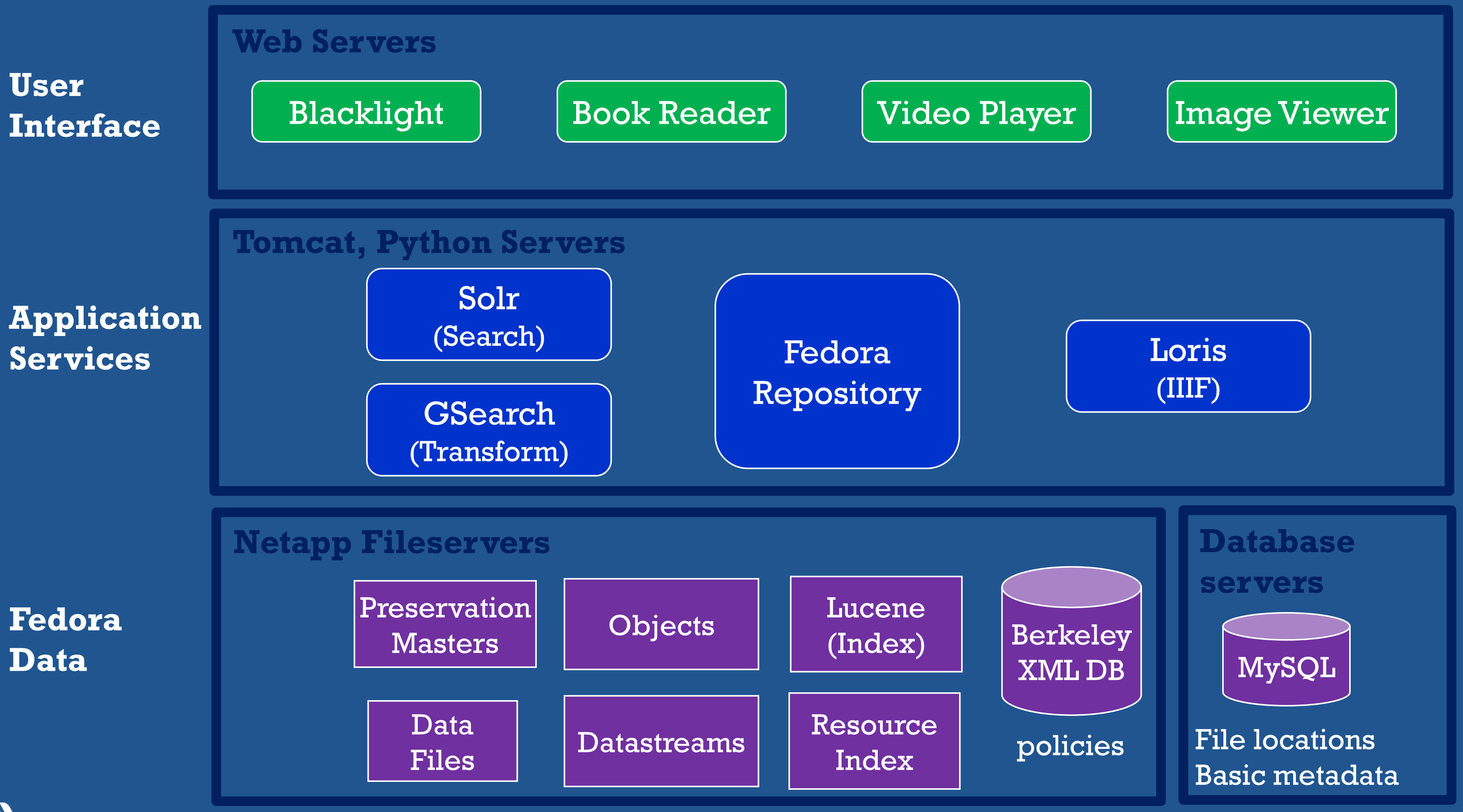
NLM Digital Repository

- 9M objects, 70 TB
- Historical and present-day books, images, film, video, audio, manuscripts, archival material, born-digital, citations, court cases, software, maps
- Fedora 3.8.1, Solr, GSearch, MySQL, CentOS
- Blacklight, IIIF Loris, content-specific viewers
- Java, JavaScript, Ruby, bash

Open Source Tools



Logical System Components



User Interface

Web Servers

Blacklight

Book Reader

Video Player

Image Viewer

Application Services

Tomcat, Python Servers

Solr
(Search)

GSearch
(Transform)

Fedora
Repository

Loris
(IIIF)

Fedora Data

Netapp Fileservers

Preservation
Masters

Objects

Lucene
(Index)

Berkeley
XML DB

Data
Files

Datastreams

Resource
Index

policies

Database servers

MySQL

File locations
Basic metadata

Fedora Migration

- Fedora 6 pilot partner
- Initial migration tests from Feb. 2020
  - 4.6M objects (books, pages) migrated in 7 days from F3 legacy to F6 pairtree
  - 3.8M objects (XML citations) migrated in 5 days from F3 akubra to F6 truncated
  - Initial issues resolved, ongoing issues with large inodes (current 500M limit), lengthy migration times
- Current work – deploy realistic target architecture
  - CentOS 7, Tomcat, MySQL 8
- Future: JMeter load tests, ingest, access

Target Architecture

- Public web layer in cloud
- AWS/GCP, Kubernetes, Docker. Cloud agnostic
- Digitization hardware on-site, CentOS VMs
- Back-end applications can be in cloud or on-premises
  - Fedora / OCFL on-premises or in-cloud? TBD
- Identify workflow step where content is moved to cloud
- 3 copy LOCKSS including cloud, on-premises

Design Considerations

- External Content
  - Leave files, migrate links in F6
  - Migrate to OCFL separately from F6 OCFL
  - Migrate to F6 managed OCFL
- Metadata in RDF.
- RDF/LDP concerns. Use our own URIs as subjects.
- Typically ingest using side-loading for performance in moving large files, prebuilt FOXML
- Clustering for availability
- Revisit Preservation requirements

Thank You!

NLM Digital Collections: https://collections.nlm.nih.gov/

Acknowledgments: Jennifer Gilbert, Calvin Xu, Steve Liu, National Library of Medicine

doron.shalvi@nih.gov