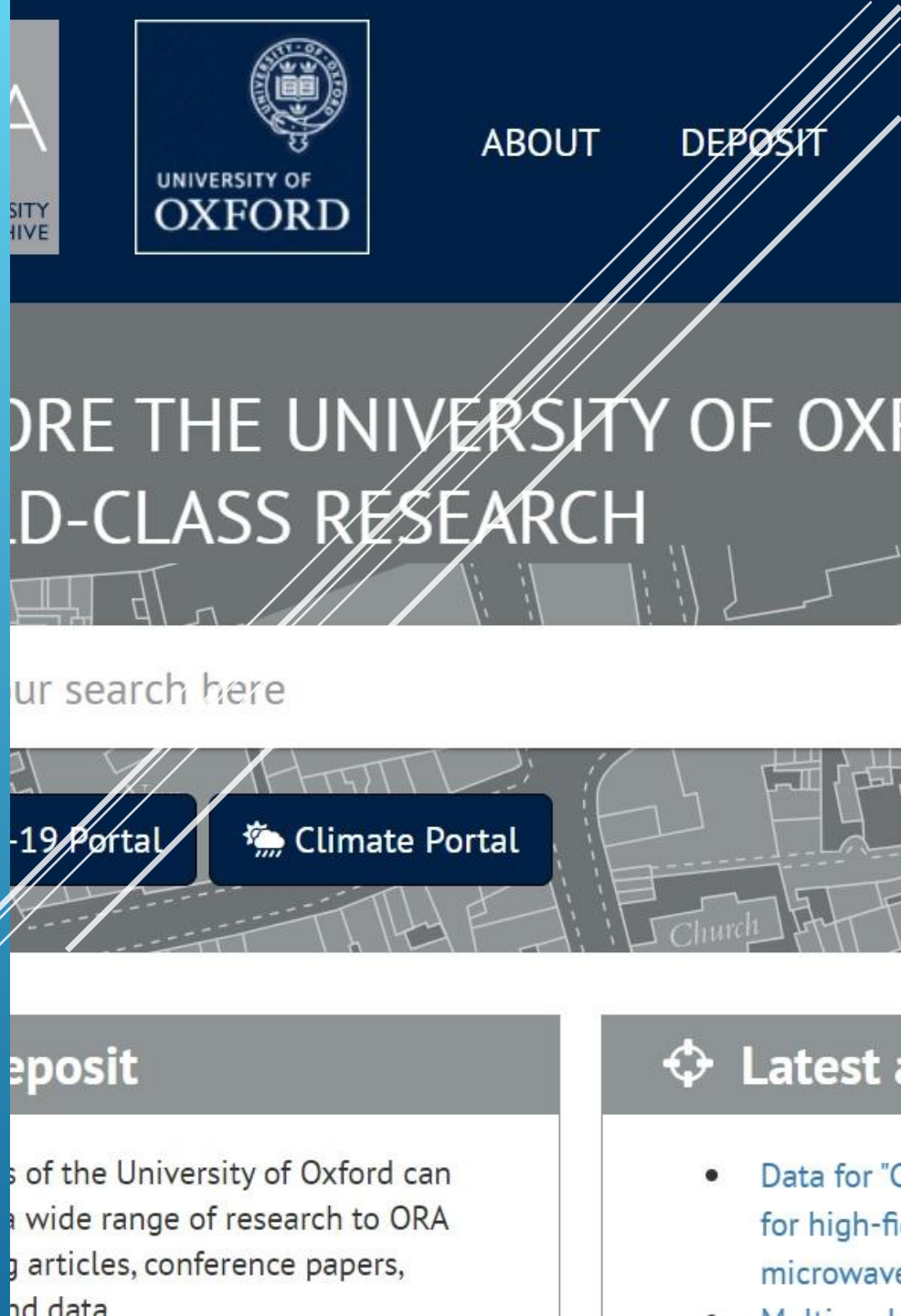
THE ORA DATA PRESERVATION SERVICE

A LIGHTWEIGHT, OPEN-SOURCE, DIGITAL REPOSITORY SOLUTION

Tom Wrobel
Oxford Research Archive (ORA), Bodleian Libraries, University of Oxford

Open Repositories 2024
Developer Track Session 3, 2024-06-06



ORA

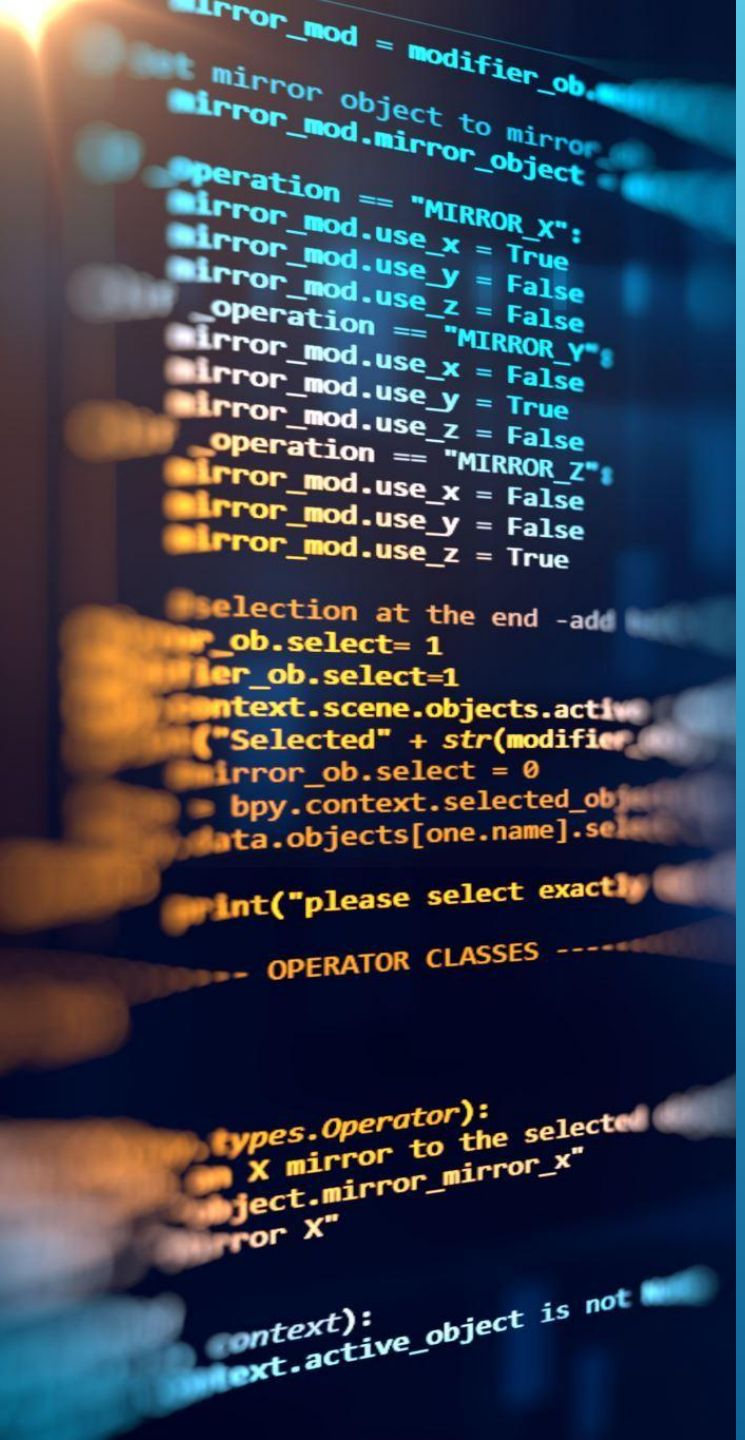
(OXFORD RESEARCH ARCHIVE)

https://ora.ox.ac.uk

ORA is the institutional repository for the University of Oxford, founded in 2006

300,000 digital research objects incl. journal articles, theses, small-medium size datasets, and working papers

Built from Samvera Hyrax 2.5, using Fedora 4.7 for storage



Independent of operational repository

Not dependent on single supplier/software

Co-locate object metadata and binary files

Store versioned digital objects

Data storage in an application agnostic format

Portable and re-usable

ORA DATA PRESERVATION SERVICE (DPS) REQUIREMENTS

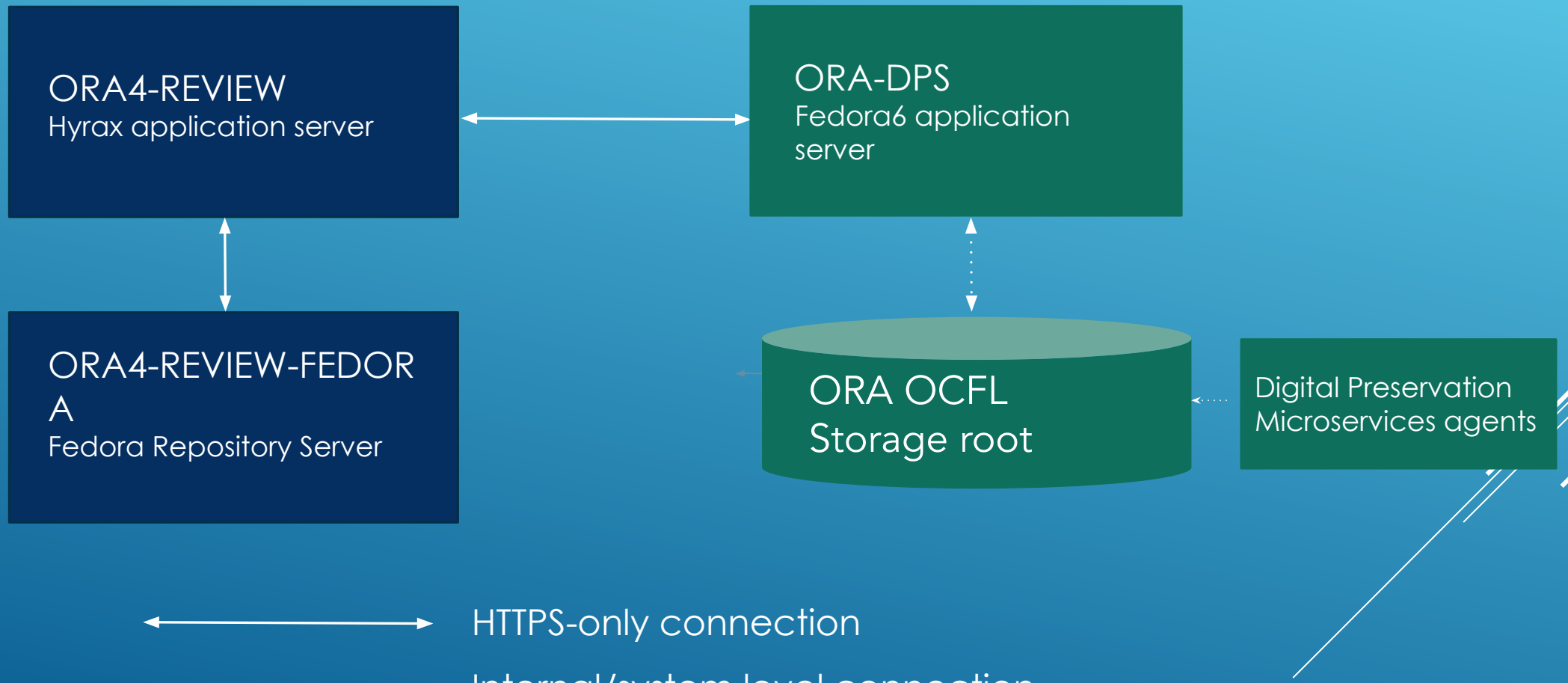
THE ORA DPS IMPLEMENTATION

THE ORA DPS USES A FEDORA6 (V6.5) REPOSITORY WITH ARCHIVAL GROUPS AS AN API LAYER TO MANAGE AN OCFL FILE SYSTEM

THE DPS IS THE OCFL FILE SYSTEM ON DISC, NOT THE FEDORA6 LAYER

ALL CODE RELEASED UNDER MIT LICENSE, INCLUDING FEDORA6 TESTING CODE

ORA REVIEW AND ORA DPS SERVICES

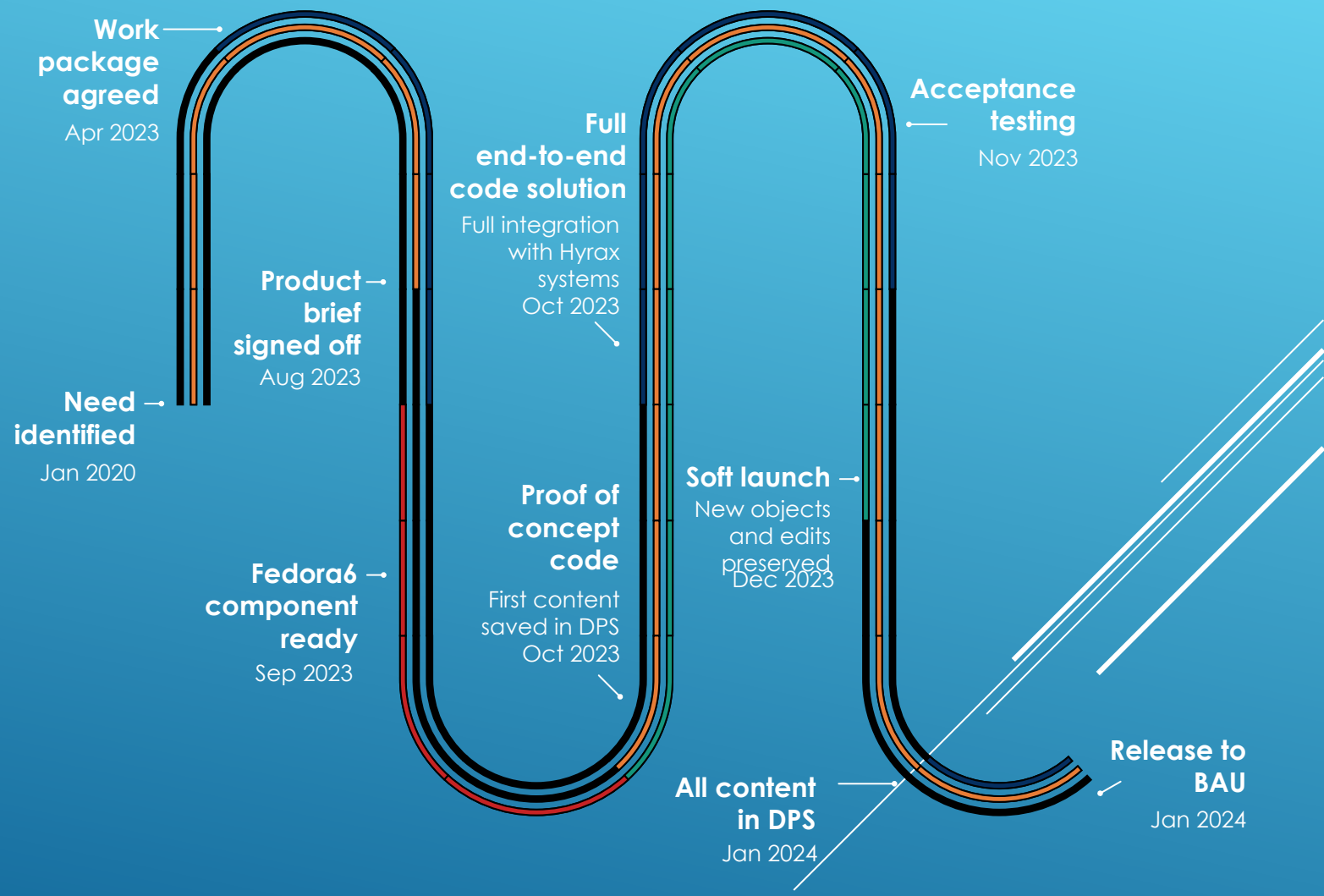


ORA DPS TIMELINE

4 months Fedora6 performance testing and configuration

6 months Contractor time - including performance testing

18 months Product scoping and requirements gathering



Preserving ORA objects in the ORA DPS via the UI, a walkthrough

Tom Wrobel, for Open Repositories 2024

SERVICE BENEFITS

beyond initial requirements

digital preservation

centralised, content agnostic , traceable

The ORA DPS exposes the content within to external digital preservation and system monitoring tools, such as virus checking, while retaining object and file structures. Increased backup coverage and performance.

versioned digital objects

metadata and binary content

Metadata is a first-class citizen in digital preservation, enhancing data security. Point in time versions can be downloaded

migration-proof data

future-proofing built in

The application-agnostic nature of data stored in the ORA DPS means that whatever the next generation of ORA will be, this will be its DPS, and we can populate any new system from the objects in the DPS without the need to migrate

POPULATING THE ORA DPS

ORA REPOSITORY IN DEC 2023

OBJECTS
295,000

OBJECTS W. BINARY FILES
120,000

DISTINCT BINARY FILES
250,000

REPOSITORY SIZE ON DISK
2.6TB

12 hours, in three batches

1,650,000 distinct files in the filesystem, including Fedora metadata files

832,000 ORA metadata and binary files

250,000 ORA binary files

2,000,000 API calls

2.6TB of data

RUNNING THE ORA DPS
EACH MONTH

ORA REPOSITORY IN MAY 2024

OBJECTS
300,000

OBJECTS W. BINARY FILES
125,000

DISTINCT BINARY FILES
280,000

REPOSITORY SIZE ON DISK
2.7TB

3,500 object pushes
1,200 new objects

17,000 API calls per month

40GB of data

SPECIFICATION AND COSTS

build	RHEL 7, 12 CPUs, 32GB RAM, 8TB Dell Compellent Storage
cost	On premises hosted with no stated costs
aws migration	EC2 instance: $1000/month, API calls $10,000, S3 storage: $63,
aws monthly	EC2: $250, API: $85, S3 Storage: $63

WITH THANKS TO

- CottageLabs, providers of many of the components listed here, especially Anusha Ranganathan
- The ORA Development Team
- The Bodleian Digital Systems and Services Infrastructure Team
- The Fedora6 community

Fedora™

MORE INFORMATION, CODE, AND LINKS TO RESOURCES

https://github.com/tomwrobel/OR2024

GET IN TOUCH

thomas.wrobel@bodleian.ox.ac.uk

ORA
OXFORD UNIVERSITY RESEARCH ARCHIVE

UNIVERSITY OF OXFORD