

ORA and the Oxford Common Filesystem Layer (OCFL) in Fedora 6

About us

The Bodleian Libraries is [sic] the library service supporting the work of the University of Oxford. Since 1602, it has grown to a collection of more than 13 million printed items.

ORA (Oxford University Research Archive) is the institutional repository for the University of Oxford. We are a permanent and open archive of research materials produced by members of the University of Oxford. This includes articles, conference papers, theses, research data, working papers, posters and more. We hold more than 300,000 works, and over 100,000 of these have public binary files. For many of these works, ORA holds the only digital copy.

Our mission is not just to make these works available without a paywall or subscription today, but as far into the future as we can; to develop ways to make today's digital content as available in 400 years as the books of the 17th century are today.

Towards this end, the Bodleian Libraries is developing an advanced set of Digital Preservation Microservices (DPMS) which aims to monitor and support the preservation of binary content and metadata – e.g., fixity, version checking, virus scanning - as well as establishing Digital Preservation Best practice. This goes beyond fixity, into questions like 'how many PDFs do you hold that are image-only?', or 'what binary files do you hold that do not have supported readers?'. Alongside this, we are building a Digital Preservation Service (DPS) which will preserve a versioned copy of a digital object which will allow the DPMS to monitor, analyse, and support the system. In the first instance, the DPS will support the ORA platform, but we hope to roll the service out across the Bodleian's digital services.

Advantages of OCFL

Using OCFL for the Digital Preservation Service gives us several major benefits:

- It is platform and application agnostic: multiple systems can update and manage the OCFL layer if necessary. We plan to use Fedora6 and OCFL-java to manage the API layer but to use off-the-shelf python and go tools to index and validate those objects.
- In the future, we can use any OCFL-aware application to replace Fedora6 and/or ocf-java.
- The DPS OCFL layer can replace content migrations – instead of migrating content every time we update a management system, we can simply import content into the management system as required.
- It can be backed up and monitored on a system level and optimised using conventional file system management systems (e.g., rsync).
- The system can be rebuilt entirely from a file-system copy.
- The structure of the OCFL layer is clearly documented and human readable.
- All object metadata and binary files are stored in a single parent directory, which can be predictably located on the filesystem from the object's UUID. This means we do not need an index or management application to find or analyse any given object.
- It provides object level versioning: because OCFL allows us to retrieve the state of any object at any point in time, we can use object metadata to designate given timestamps as semantic versions. For example, we can designate 2023-03-14T12:20:12 as 'Version 2' of an object in its metadata, and then retrieve the object from that point in time.

Advantages we are seeing with Fedora6

- A well-documented RESTful API layer
 - o We can manipulate objects and their versions using HTTPS
 - o We can loosely couple the DPS and the services that use it
 - o We can use the API layer for all CRUD (create, read, update, delete) actions, including individual object components, such as binary files
 - o Modification of existing objects and access to point in time versions
 - o It is independent of ORA's business logic or operational system needs. A different service, say our sister project Digital.Bodleian – which has a completely different software stack – can still use the API to manage their content.
- Transaction management: objects being updated are locked against further update, and transactions can be rolled back – especially useful where large files are being ingested
- An authentication and authorisation system.
- We get all of the above 'out of the box'. While optimal configuration has been a challenge, we are able to store a range of 300,000 objects (the largest single object being 500Gb in size) in a reliable, performant way - we have not tested larger - without writing any code.
- Last but not least: we are part of a larger community of users and developers, some of whom provide contracted services for future update and support.

What we're doing next

We have just completed a suite of performance and configuration testing of the Fedora6+OCFL system. Our next step is to export our existing ORA repository into the DPS, and to integrate it with day-to-day operations. Following this, we hope to expand the DPS into other services within the Bodleian Libraries.

Contact us

You can contact me at thomas.wrobel@bodleian.ox.ac.uk, or the team at ora-dev@bodleian.ox.ac.uk