Preservation Tools Techniques and Policies

Preserving Scientific Annotation WG

CC BY-SA 4.0

The Oxford Common File Layout

Neil Jefferies, Bodleian Libraries, University of Oxford

CC BY-SA 4.0

What is it?

The Oxford Common File Layout (OCFL) specification describes an **application-independent** approach to the storage of digital information in a structured, transparent, and predictable manner. It is designed to promote long-term object management best practices within digital repositories.

Observations

- Archived objects change relatively slowly than archival software.
- Filesystems (and in particular POSIX filesystems) have been the most consistently implemented and widely tested API's for accessing storage in any form.
- Migration by export/ingest is slow and risky
- MOAB, BagIT, RDF DataBank as antecedents…

CC BY-SA 4.0

Key attributes

- **Complete.** All the information about a digital object in an OCFL compliant repository should be serialised in the OCFL.
- **Application Independent.** Consequently, a repository should be rebuildable from just the data in an OCFL. Even if it is not the source repository.
- **Human parsable.** An OCFL should be understandable to a person (with a little effort). With basic filesystem tools they should be able to identify digital objects and their versions and contents.
- **Portable.** OCFL requires a minimal set of filesystem capabilities so it can be implemented on most filesystems, and be portable between them.
- **Provenance and Versioning**. OCFL allows the capture of a version history for objects and provide for the implementation of an audit trail.
- **Deduplication.** OCFL allows for deduplication of content between object versions so that unchanged parts
- **Fixity.** OCFL provides fixity as a by-product of its use of content-based addressing.
- **Burn-Line.** OCFL permits the loss of all systems except for a basic file-system storage node and still provide full recoverability.

CC BY-SA 4.0

Who?

- Andrew Hankinson (Bodleian Libraries, University of Oxford)
- Neil Jefferies (Bodleian Libraries, University of Oxford)
- Rosalyn Metz (Emory University)
- Julian Morley (Stanford University)
- Simeon Warner (Cornell University)
- Andrew Woods (DuraSpace)

- https://groups.google.com/forum/#!forum/ocfl-community

- “Oxford” because of a Samvera meeting at Oxford…

CC BY-SA 4.0

OCFL Features

- Focus is a digital object composed of multiple files
  - File format agnostic, data and metadata are treated equally
- Objects are represented by a filesystem directory which contains:
  - Sub-directories that correspond to each version of the object
    - OCFL supports delta versioning so each version only contains changes from the previous one
  - A JSON inventory file that indicates how each version is derived from previous versions
    - And some basic preservation information such as fixity
- All object directories are contained in a top level folder called the OCFL Root
  - The path to an individual object directory is derived from an object PID (e.g. Pair Tree)
  - The root contains OCFL version information and documentation to allow recovery

CC BY-SA 4.0

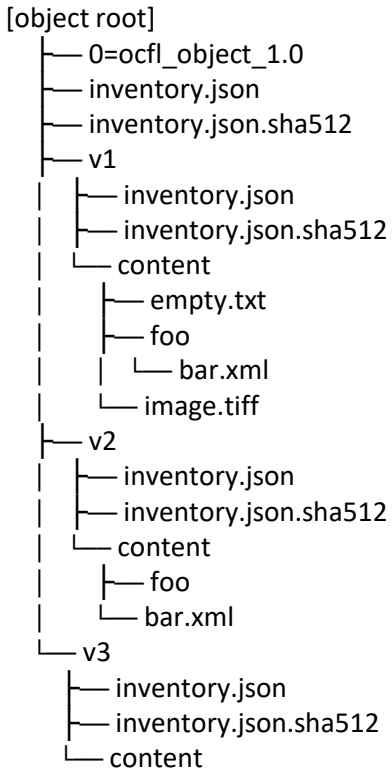
What does it look like?

```
[storage_root]
├── 0=ocfl_1.0
├── ocfl_1.0.txt (optional OCFL spec)
├── d45
│   └── be6
│       └── 26e
│           ├──d45be626e024
│           │   ├── 0=ocfl_object_1.0
│           │   └── ...
│           └──d45be626e036
│               ├── 0=ocfl_object_1.0
│               └── ...
├── 310
│   └── 4ed
│       └── f03
│           └── 3104edf0363a
│               ├── 0=ocfl_object_1.0
│               └── ...
└── ...
```

```
[object root]
├── 0=ocfl_object_1.0
├── inventory.json
├── inventory.json.sha512
├── v1
│   ├── inventory.json
│   ├── inventory.json.sha512
│   └── content
│       └── myfirstbag
│           ├── bagit.txt
│           ├── data
│           │   └── 27613-h
│           │       └── images
│           │           ├── q172.png
│           │           └── q172.txt
│           └── manifest-md5.txt
└── v2
    ├── inventory.json
    ├── inventory.json.sha512
    └── content
        └── myfirstbag
            ├── data
            │   └── 27614-h
            │       └── images
            │           ├── q173.png
            │           └── q173.txt
            └── manifest-md5.txt
```

CC BY-SA 4.0

Inventory




```
{
 "digestAlgorithm": "sha512",
 "head": "v3",
 "id": "ark:/12345/bcd987",
 "manifest": {
  "4d27c8...b53": [ "v2/content/foo/bar.xml" ],
  "7dcc35...c31": [ "v1/content/foo/bar.xml" ],
  "cf83e1...a3e": [ "v1/content/empty.txt" ],
  "ffccf6...62e": [ "v1/content/image.tiff" ]
 },
 "type": "Object",
 "versions": {
  "v1": {
   "created": "2018-01-01T01:01:01Z",
   "message": "Initial import",
   "state": {
    "7dcc35...c31": [ "foo/bar.xml" ],
    "cf83e1...a3e": [ "empty.txt" ],
    "ffccf6...62e": [ "image.tiff" ]
   },
   "type": "Version",
   "user": {
    "address": "alice@example.com",
    "name": "Alice"
   }
  },
  "v2": {
   "created": "2018-02-02T02:02:02Z",
   "message": "Fix bar.xml, remove image.tiff, add empty2.txt",
   "state": {
    "4d27c8...b53": [ "foo/bar.xml" ],
    "cf83e1...a3e": [ "empty.txt", "empty2.txt" ]
   },
   "type": "Version",
   "user": {
    "address": "bob@example.com",
    "name": "Bob"
   }
  },
  "v3": {
   "created": "2018-03-03T03:03:03Z",
   "message": "Reinstate image.tiff, delete empty.txt",
   "state": {
    "4d27c8...b53": [ "foo/bar.xml" ],
    "cf83e1...a3e": [ "empty2.txt" ],
    "ffccf6...62e": [ "image.tiff" ]
   },
   "type": "Version",
   "user": {
    "address": "cecilia@example.com",
    "name": "Cecilia“ }}}}
```


CC BY-SA 4.0

Current Status

- V0.1 Alpha specification https://ocfl.io/
  - Feedback/issues to https://github.com/OCFL/spec/issues
- Mailing Lists
  - pasig-discuss@mail.asis.org Broader communications, announcements, meeting notifications
  - ocfl-community@googlegroups.com OCFL community discussions, announcements, support
  - ocfl-tech@googlegroups.com Software implementer discussions, spec interpretation, technical discussions
- Implementation work at Oxford, Cornell, Data Futures...

CC BY-SA 4.0