The gocfl Workflow for Research Data

Reporting and publication of clustered data packages

The Oxford Common File Layout (OCFL[1]) is an application-independent specification for the structured, transparent storage of digital information. Its internal structure is geared towards the requirements of long-term storage and makes it easier to check the completeness and integrity of OCFL-data (-packages). The main goal is to keep knowledge about data available, even after active use, e.g. after completion of a research project.

The open-source software GOCFL[2] is programmed in GO and fulfills the OCFL standard. It provides a routine for the automated and sustainable packaging of data. The tool enables resource-saving reporting. In addition, people who were not actively involved in the editing processes can also keep an eye on the data situation (content, usage and other property rights) with the help of the web- and pdf-based human-readable reports. Optionally, both the OCFL packages and the reports (web format + PDF) can be published FAIRly.

This document outlines the implementation of the GOCFL workflows at the Academy of Art and Design Basel FHNW. It explains the core-functionality of this specific open-source software. Graphically, the GOCFL workflow can be anchored in the research data cycle as follows:

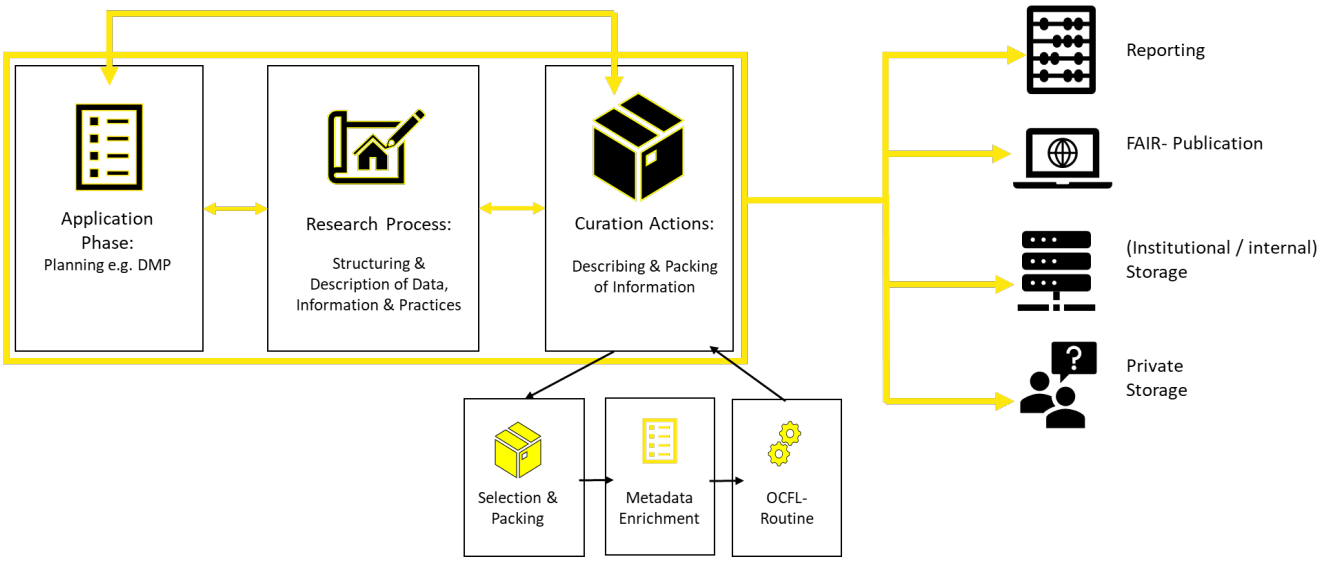


Fig. 1: GOCFL workflow as implemented within the research data cycles at HGK Basel FHNW

[1] https://ocfl.io/1.1/spec/

[2] https://github.com/je4/gocfl

Mediathek
HGK FHNW Basel

Freilager-Platz 1
4002 Basel

T +41 61 228 43 65

tabea.lurk@fhnw.ch
www.fhnw.ch/hgk

Implementing of GOCFL at HGK Basel FHWN

The open-source software GOCFL was integrated within the research data cycle in 2023 at HGK Basel FHNW. The main aim was to improve both the quality of data management regarding data overview and the publication options for closed data. Since then, GOCFL has been one of the institutionally funded services used in the final research phase when a project is completed (see Fig. 1).

A classic application for the use of GOCFL is research data that should not or cannot be freely published after completion of a project. The reports comment on data protection, copyright, and personal rights. They offer researchers an efficient way of representing the diversity of material identified in the data management plans, while they enable research managers to obtain an overview of the status quo of the data available at their institute.

Functionalities of GOCFL

As a software library, GOCFL extends the OCFL standard according to specified rules with software routines that contribute to sustainability, reporting and thus improve FAIR-ness. GOCFL zips folders of structured data and preserves the files and their structure. It extracts technical metadata, builds checksums and documents the applied routines. GOCFL allows later updates and validation through versioning. Of special interest are the reports GOCFL creates: They provide clear and human-readable information about the available data packages as browser-based (full length) and PDF-compatible (short/abbreviated) versions.

Implemented as a modular software with indexing cascades, further functionalities can be added or removed to meet specific requirements. To give an example, if required, GOCFL might encrypt filenames.

In general, all OCFL objects contain a so-called object root, in which data and an inventory are contained.

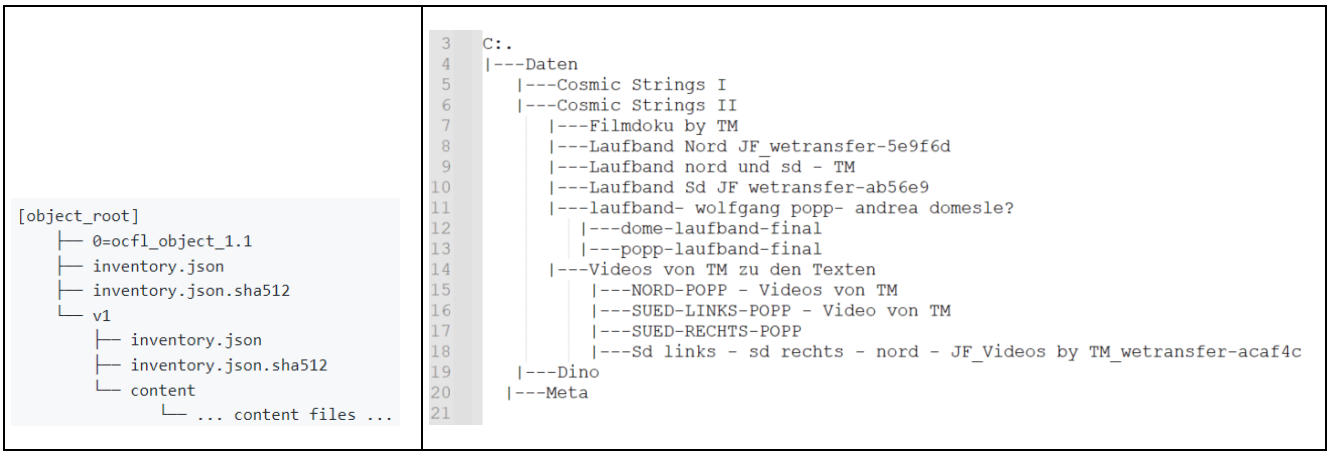


Fig. 2: OCFL object structure: Left side according to the standard[3], right as sample of HGK Basel

Since filenames and filing structures often contain semantic information regarding content and context such as the research process (dates, persons or places involved), filenames and -structures are preserved. Corresponding data is extracted and can be displayed after the routine.

[3] https://ocfl.io/1.1/spec/

Since special characters in file names can jeopardize sustainability, one of the GOCFL-routines renames them by converting unsupported characters to utf8. All changes are documented.

Due to its technical structure as cascade of single software libraries, hich can be added or removed as wished, GOCFL can be extend for carrying out additional requirements.

Currently GOCFL extracts the following metadata during the technical analysis:

- File size
- Creation and modification date
- Filename and position in the directory structure
- Characterization of the data type and the underlying operating system
- Determination of the file format using PRONOM Persistent Unique Identifier (PUID)
- Creation of checksums to identify the files
- Depending on the configuration, further metadata
- Removal of special characters from file names to increase sustainability
- Creation of a ZIP file[4].

To create an easy-to-understand report, GOCFL also offers the following functionalities:

- Statistical evaluation of the data according to data type and format
- Creation of thumbnails
- Creation of a web-based overall report
- Creation of a short version of the report that is laid out for common PDF format
- Filename and position in the directory structure
- Characterization of the data type and the underlying operating system

[4] The OCFL packages are available as a zip file, as archive memories are generally not very good at handling many small files.

Web based report

The index and technical metadata of the OCFL object can be viewed using the browser:

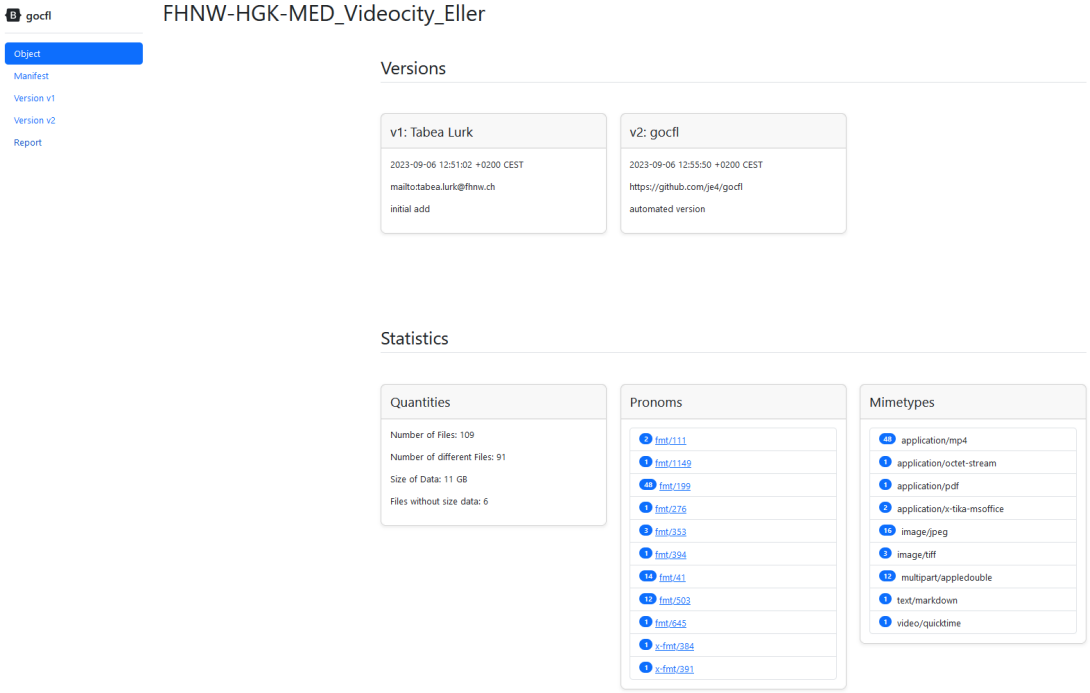


Fig. 3: Web-based report of the packed OCFL object

Fig. 4: Manifest of the package with list of files, linkage to the full index information, file format and size information

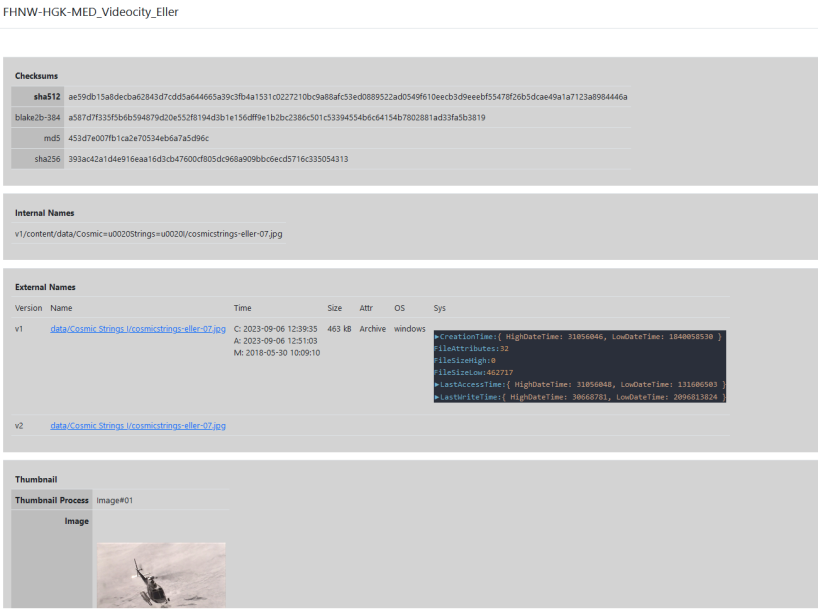
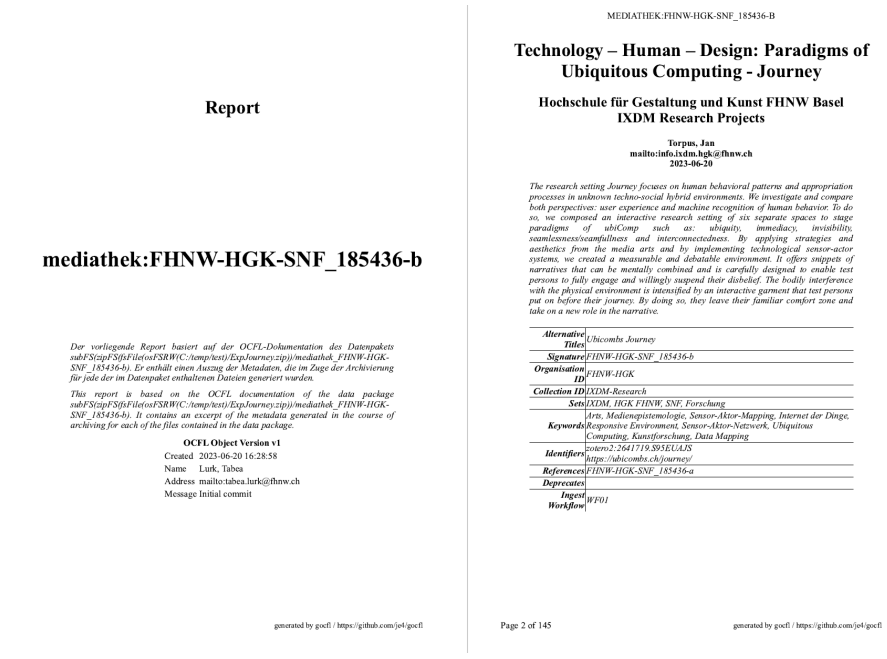


Fig. 5: Excerpt of the detail view of the analysis, including a thumbnail of the stored images

PDF-based report

The PDF report contains a laid-out extract of the technically extracted metadata. The cover page with the info.json (see below) is followed by a statistical list of the file formats contained in the package, the directory tree and finally a short version of the extracted metadata.



Report

mediathek:FHNW-HGK-SNF_185436-b

Der vorliegende Report basiert auf der OCFL-Dokumentation des Datenpakets subFS(zipFS(fsFile(osFSRW(C:/temp/test)/ExpJourney.zip))/mediathek_FHNW-HGK-SNF_185436-b). Er enthält einen Auszug der Metadaten, die im Zuge der Archivierung für jede der im Datenpaket enthaltenen Dateien generiert wurden.

This report is based on the OCFL documentation of the data package subFS(zipFS(fsFile(osFSRW(C:/temp/test)/ExpJourney.zip))/mediathek_FHNW-HGK-SNF_185436-b). It contains an excerpt of the metadata generated in the course of archiving for each of the files contained in the data package.

OCFL Object Version v1

Created 2023-06-20 16:28:58
Name Lurk, Tabea
Address mailto:tabea.lurk@fhnw.ch
Message Initial commit

generated by gocfl / https://github.com/je4/gocfl

MEDIATHEK:FHNW-HGK-SNF_185436-B

Technology – Human – Design: Paradigms of Ubiquitous Computing - Journey

Hochschule für Gestaltung und Kunst FHNW Basel
IXDM Research Projects

Torpus, Jan
mailto:info.ixdm.hgk@fhnw.ch
2023-06-20

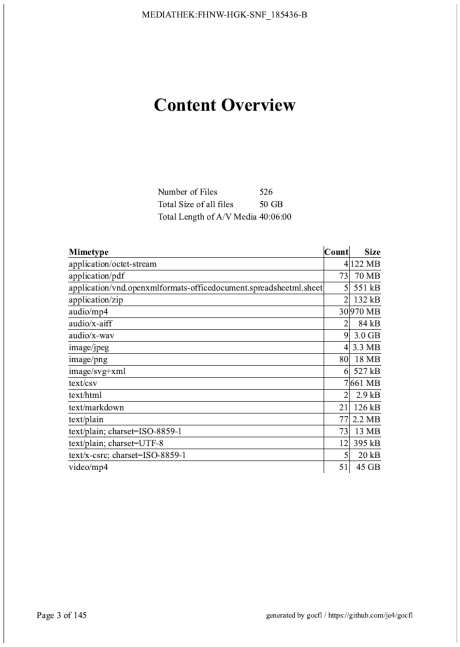
The research setting Journey focuses on human behavioral patterns and appropriation processes in unknown techno-social hybrid environments. We investigate and compare both perspectives: user experience and machine recognition of human behavior. To do so, we composed an interactive research setting of six separate spaces to stage paradigms of ubiComp such as: ubiquity, immediacy, invisibility, seamlessness/seamfullness and interconnectedness. By applying strategies and aesthetics from the media arts and by implementing technological sensor-actor systems, we created a measurable and debatable environment. It offers snippets of narratives that can be mentally combined and is carefully designed to enable test persons to fully engage and willingly suspend their disbelief. The bodily interference with the physical environment is intensified by an interactive garment that test persons put on before their journey. By doing so, they leave their familiar comfort zone and take on a new role in the narrative.

Alternative Titles	*Ubicombs Journey*
Signature	*FHNW-HGK-SNF_185436-b*
Organisation ID	*FHNW-HGK*
Collection ID	*IXDM-Research*
Sets	*IXDM, HGK FHNW, SNF, Forschung*
Keywords	*Arts, Medienepistemologie, Sensor-Aktor-Mapping, Internet der Dinge, Responsive Environment, Sensor-Aktor-Netzwerk, Ubiquitous Computing, Kunstforschung, Data Mapping*
Identifiers	*zotero2:2641719.S95EUAJS https://ubicombs.ch/journey/*
References	*FHNW-HGK-SNF_185436-a*
Deprecates	
Ingest Workflow	*WF01*

Page 2 of 145

generated by gocfl / https://github.com/je4/gocfl

Fig. 6: Intro sequence of the PDF-based report (page 1) with information about the project and persons in charge (page 2)



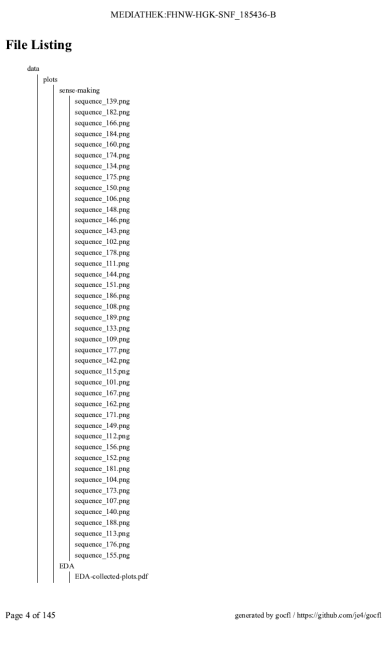
MEDIATHEK:FHNW-HGK-SNF_185436-B

Content Overview

Number of Files 526
Total Size of all files 50 GB
Total Length of A/V Media 40:06:00

Mimetype	Count	Size
application/octet-stream	4	122 MB
application/pdf	73	70 MB
application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	5	551 kB
application/zip	2	132 kB
audio/mp4	30	970 MB
audio/x-aiff	2	84 kB
audio/x-wav	9	3.0 GB
image/jpeg	4	3.3 MB
image/png	80	18 MB
image/svg+xml	6	527 kB
text/csv	7	661 MB
text/html	2	2.9 kB
text/markdown	21	126 kB
text/plain	77	2.2 MB
text/plain; charset=ISO-8859-1	73	13 MB
text/plain; charset=UTF-8	12	395 kB
text/x-csrc; charset=ISO-8859-1	5	20 kB
video/mp4	51	45 GB

Page 3 of 145 generated by gocfl / https://github.com/je4/gocfl



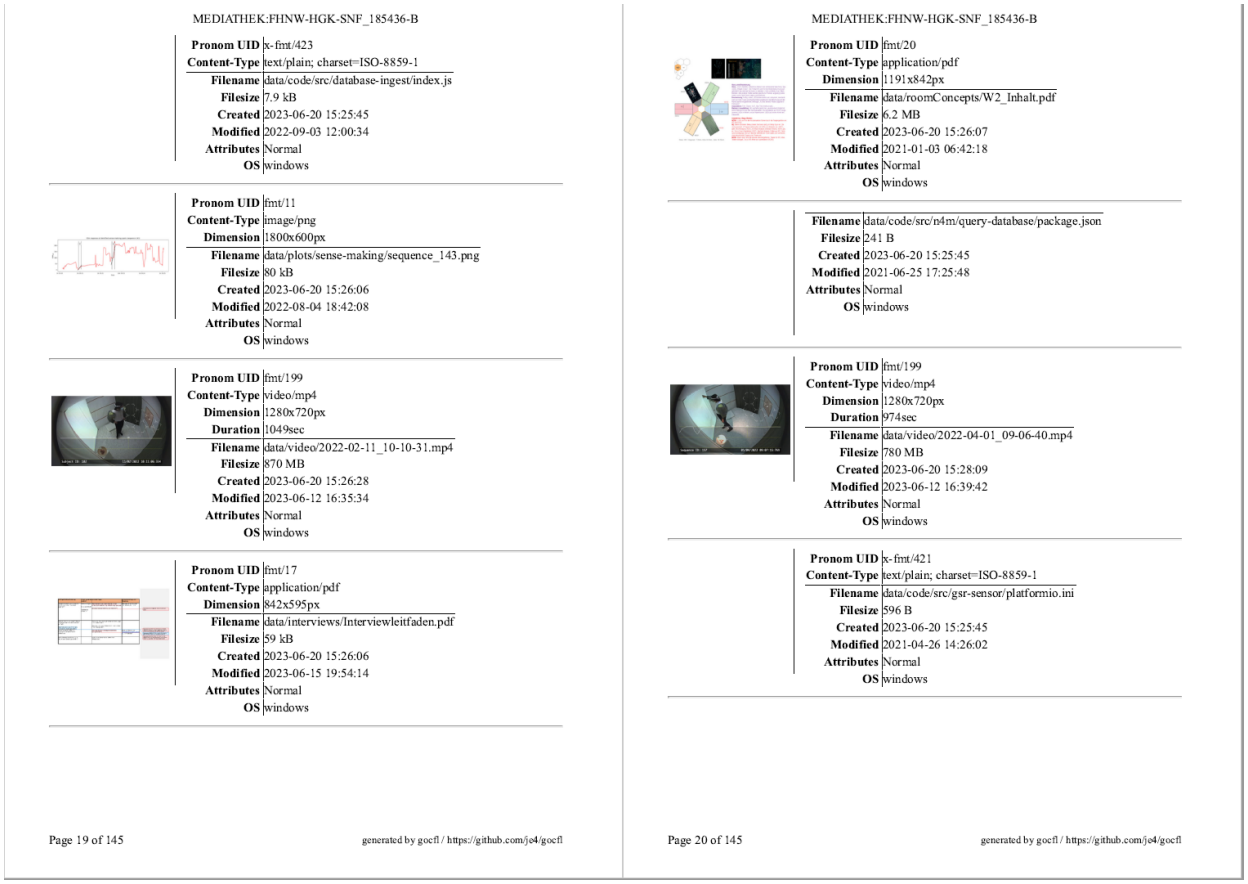
MEDIATHEK:FHNW-HGK-SNF_185436-B

File Listing

data
- plots
  - sense-making
    - sequence_139.png
    - sequence_182.png
    - sequence_166.png
    - sequence_184.png
    - sequence_160.png
    - sequence_174.png
    - sequence_134.png
    - sequence_175.png
    - sequence_150.png
    - sequence_106.png
    - sequence_148.png
    - sequence_146.png
    - sequence_143.png
    - sequence_102.png
    - sequence_178.png
    - sequence_111.png
    - sequence_144.png
    - sequence_151.png
    - sequence_186.png
    - sequence_108.png
    - sequence_189.png
    - sequence_133.png
    - sequence_109.png
    - sequence_177.png
    - sequence_142.png
    - sequence_115.png
    - sequence_101.png
    - sequence_167.png
    - sequence_162.png
    - sequence_171.png
    - sequence_149.png
    - sequence_112.png
    - sequence_156.png
    - sequence_152.png
    - sequence_181.png
    - sequence_104.png
    - sequence_173.png
    - sequence_107.png
    - sequence_140.png
    - sequence_188.png
    - sequence_113.png
    - sequence_176.png
    - sequence_155.png
  - EDA
    - EDA-collected-plots.pdf

Page 4 of 145 generated by gocfl / https://github.com/je4/gocfl

Fig. 7: Listing of file formats according to frequency and file tree



MEDIATHEK:FHNW-HGK-SNF_185436-B

Pronom UID	x-fmt/423
Content-Type	text/plain; charset=ISO-8859-1
Filename	data/code/src/database-ingest/index.js
Filesize	7.9 kB
Created	2023-06-20 15:25:45
Modified	2022-09-03 12:00:34
Attributes	Normal
OS	windows

Pronom UID	fmt/11
Content-Type	image/png
Dimension	1800x600px
Filename	data/plots/sense-making/sequence_143.png
Filesize	80 kB
Created	2023-06-20 15:26:06
Modified	2022-08-04 18:42:08
Attributes	Normal
OS	windows

Pronom UID	fmt/199
Content-Type	video/mp4
Dimension	1280x720px
Duration	1049sec
Filename	data/video/2022-02-11_10-10-31.mp4
Filesize	870 MB
Created	2023-06-20 15:26:28
Modified	2023-06-12 16:35:34
Attributes	Normal
OS	windows

Pronom UID	fmt/17
Content-Type	application/pdf
Dimension	842x595px
Filename	data/interviews/Interviewleitfaden.pdf
Filesize	59 kB
Created	2023-06-20 15:26:06
Modified	2023-06-15 19:54:14
Attributes	Normal
OS	windows

Page 19 of 145 generated by gocfl / https://github.com/je4/gocfl

MEDIATHEK:FHNW-HGK-SNF_185436-B

Pronom UID	fmt/20
Content-Type	application/pdf
Dimension	1191x842px
Filename	data/roomConcepts/W2_Inhalt.pdf
Filesize	6.2 MB
Created	2023-06-20 15:26:07
Modified	2021-01-03 06:42:18
Attributes	Normal
OS	windows

Filename	data/code/src/n4m/query-database/package.json
Filesize	241 B
Created	2023-06-20 15:25:45
Modified	2021-06-25 17:25:48
Attributes	Normal
OS	windows

Pronom UID	fmt/199
Content-Type	video/mp4
Dimension	1280x720px
Duration	974sec
Filename	data/video/2022-04-01_09-06-40.mp4
Filesize	780 MB
Created	2023-06-20 15:28:09
Modified	2023-06-12 16:39:42
Attributes	Normal
OS	windows

Pronom UID	x-fmt/421
Content-Type	text/plain; charset=ISO-8859-1
Filename	data/code/src/gsr-sensor/platformio.ini
Filesize	596 B
Created	2023-06-20 15:25:45
Modified	2021-04-26 14:26:02
Attributes	Normal
OS	windows

Page 20 of 145 generated by gocfl / https://github.com/je4/gocfl

Fig. 8: Sequence with thumbnails and metadata information about the file such as PUID

Workflow

The GOCFL workflow requires the following steps:

- Curation of data
  - Store files in one or more folders
  - Describe/characterize data clusters with a readme file (see below)
  - Add basic record information in an info.json file (see below)
- Configure GOCFL (once per executing instance [computer/server])
- Start help tools (Tika, Powershell)
- Execute the GOCFL script
- (Re-)View report and archive package and forward to a publication instance if necessary.

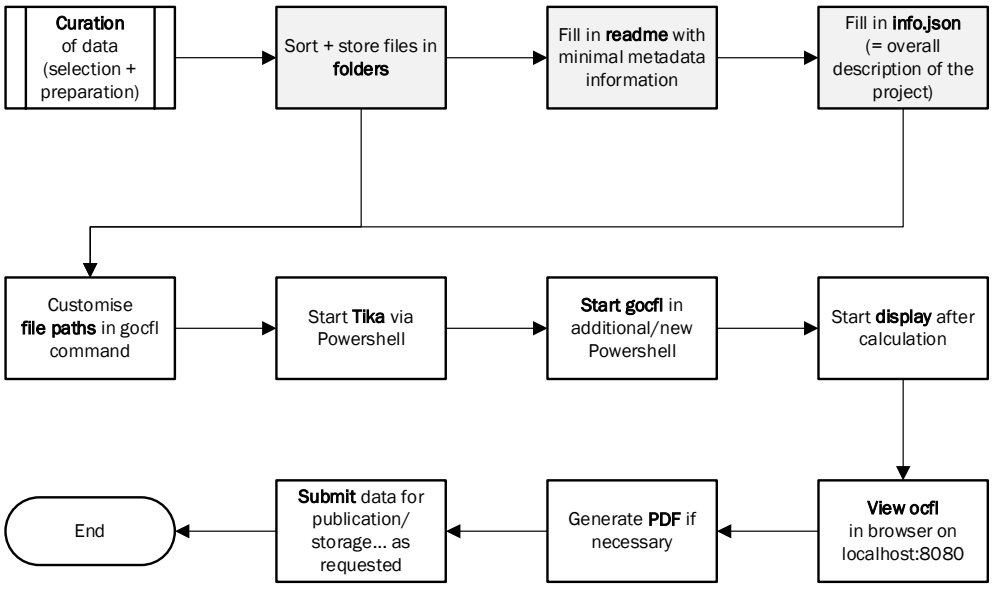


Fig. 9: OCFL-workflow

Description

Researchers are responsible for determining the depth, quality and format of description. The following explanation may thus be regarded as recommendations for a minimum amount of information, still increasing the FAIRness of data.

Nevertheless, we recommend creating at least one readme file per superordinate folder/file path that provides information about the contents of the directory. If different legal areas are affected, these should be mentioned. In principle, two levels of description should be used:

- Superordinate information about the project, the persons responsible and the creation --> info.json
- Content characterization of the stored files --> readme.

info.json

The requirements for info.json (Fig. 10) are stricter, as they must be generally meaningful. They are prominently placed on the cover page of the report and should contain formal information such as the title of the project, responsibilities and legal issues beyond the active research period.

Readme

For the readme (Fig. 10), we recommend structuring the data in terms of content according to Dublin Core. Formatting conventions vary by nature. A suitable storage format might be raw text, containing e. g. a Marcdown structure, or JSON.

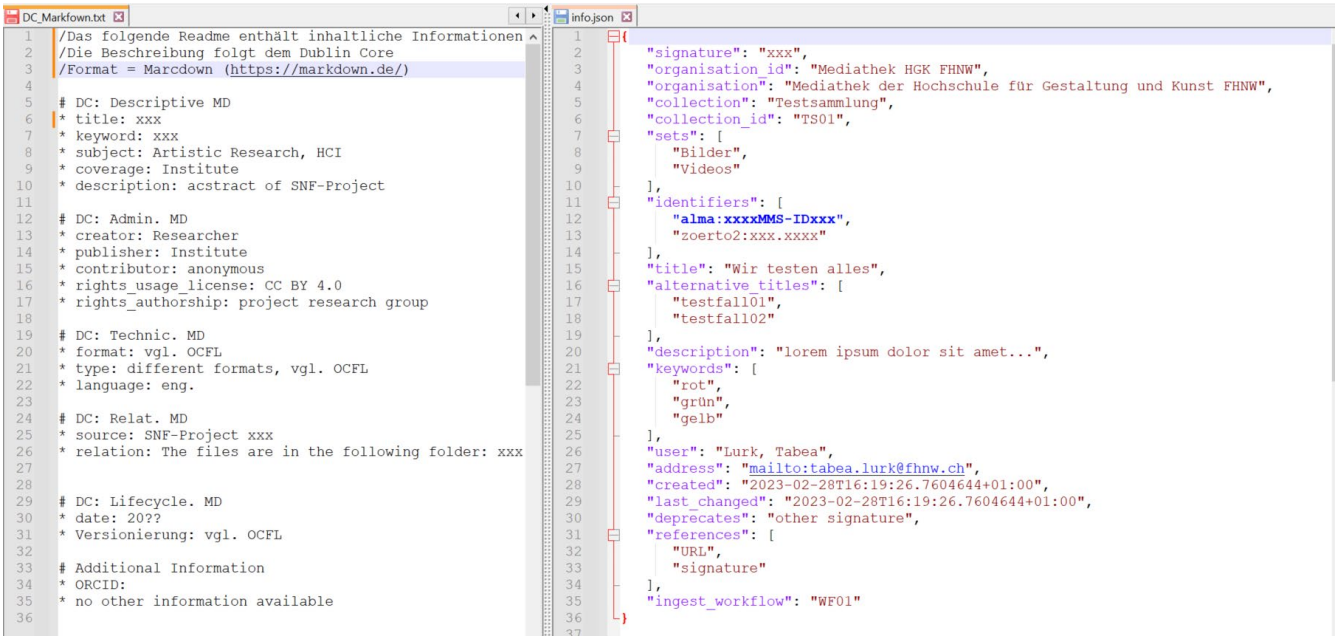


Fig. 10: Suggestions: left - content description of data, right - formal information of context

At the media library of the HGK Basel FHNW, GOCFL is mainly used in the areas of research support and for clarifying copyrights in the course of the data transfer for special collections.

License

GOCFL was developed by Dipl. Inform. Jürgen Enge[5] and is released with Apache-2.0 License.

[5] https://ub.unibas.ch/de/kontakt/.