HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

Supplement to:
SOP “Image curation and publication”

This supplemental material can be used to conduct the image curation and publication steps described in the SOP document “Image curation and publication”. Feel free to adjust to your specific needs! All material in this supplement is released under CC-0.

Content:

- A template folder structure to organize image data and metadata
- A task list to document the execution of the “Image curation and publication” SOP
- An image acquisition protocol sheet to fill in prior / during image acquisition
- Data management plan (DMP) sections for image data

References:

- SOP “Image curation and publication”: https://doi.org/10.5281/zenodo.5704846
- SOP “iFDO creation“: https://doi.org/10.5281/zenodo.5681429
- SOP „iFDO creation“ supplement: https://doi.org/10.5281/zenodo.5683081
- FAIR marine images: https://marine-imaging.com/fair

Information about this document:

Title:	SOP „Image curation and publication“ supplement
Authors:	Timm Schoening
Month, Year:	2021/11
Abstract:	This supplemental material can be used to conduct the image curation and publication steps described in the SOP document “Image curation and publication”. That SOP describes how research image data like photos and videos can be published for scientific use.
Note:	This material is kept generic. Adjust it to your needs and include specific information on your infrastructure to adapt it to your institute / workflow.
DOI:	
Keywords:	Image, curation, publication, photo, video, iFDO
License:	CC-0
Related Identifiers:	
Communities:	MareHub of the Helmholtz Association

Revisions:

Version	Date	Author	Comment
v1.0.0	2021/11	Timm Schoening	Initial draft of a public text version of this SOP. Compiled from discussions in the MareHub AG Videos/Images.

Folder structure to organize image data and image metadata

How to structure data on disk should not be enforced by any SOP. However, we recommend the following structure to aid the automation of data curation and publication workflows. It is based on several best-practices guides for handling research data (e.g. https://github.com/drivendata/cookiecutter-data-science). Using this template structure allows you to utilize the full potential of the MarIQT python software for image data curation.

Overall, data is stored in one or many *base-paths*, given in the example below as `/<volume>/<project>`.

```
/<volume>/<project>/                Base-path
├── <event_1>                       One sub-folder per event (sampling station, experiment run, …).
│   │                               Should be a globally unique identifier (no other event is called the same)
│   ├── protocol/                   General information on this event (e.g. ROV deployment plan)
│   ├── <sensor_x>                  One sub-folder per sensor (camera) used during this event
│   │   │                           Should be a globally unique identifier (no other camera is called the same)
│   │   ├── external/               (Optional) External data that affects the creation of raw data
│   │   │                           (e.g. calibration photos)
│   │   ├── raw/                    The raw data as recorded by the sensor (e.g. photos)
│   │   ├── intermediate/           (Optional) Intermediate data that will not be archived.
│   │   │                           Playground or sandbox for working with the raw data
│   │   ├── processed/              (Optional) Processed data that has been QA/QC'd and is ready for publication
│   │   │                           (e.g. color-calibrated photos)
│   │   ├── products/               (Optional) Data products created from the raw or processed data
│   │   │                           for visualization or as combinations of data of several events
│   │   │                           (e.g. 3D reconstructions)
│   │   └── protocol/               Human- and/or machine-readable documentation on how the data was
│   │                               created, curated, processed,  visualized, etc.
│   ├── <sensor_y>                  Same as above for the next sensor deployed during this event
│   │   ├── external/
│   │   ├── raw/
│   │   ├── intermediate/
│   │   ├── processed/
│   │   ├── products/
│   │   └── protocol/
│
└── <event_2>                       The same as above for the next event
    ├── <sensor_x>
    └── <sensor_z>
```

The base-path is likely to be a network drive provided by a research vessel, a research group or technical team. Some disciplines/groups like to split their data by sensor first. This is not recommended but certainly possible. In that case, the paths would look like this:

```
/<volume>/<project_i>/
├── <sensor_x>
│   ├── <event_1>
│   └── <event_2>
└── <sensor_y>
    ├── <event_1>
    └── <event_3>
```

image-project	
image-event	
image-set-name	
image-platform	
image-sensor	
image-creators	

Done	Who	When	What
Preparation phase			
	Scientists	Before project start	Create Data Management Plan (DMP) in cloud
	Scientists	Before image acquisition	Collect project information
Acquisition phase			
	Scientists	Before image acquisition	Begin an acquisition protocol
	Scientists	Image acquisition	Acquire images
	Scientists	Image acquisition	Acquire navigation data
	Scientists	After image acquisition	Complete the acquisition protocol
	Scientists	Before image curation	Structure the data on disk
	Scientists	Before image curation	Split files in size or to subfolders
	Scientists	After image acquisition	Curate image data
	Scientists	After image acquisition	Curate navigation data
	Scientists	After data curation	Create iFDOs for metadata
	Scientists	After iFDO creation	Create Data Processing Report (DPR)
	Scientists	After data processing report	Update Data Management Plan (DMP)
Publication phase			
	Scientists	After iFDO creation	Publish iFDOs
	Scientists	After iFDO publication	Publish navigation data through web interfaces
	Scientists	After workspace creation	Migrate image data to web server
	RDM team	After image migration	Make data available for annotation (optional)
	Scientists	After image migration	Publish data in PANGAEA (optional)
	RDM team	After publication	Update Data Management Plan (DMP)

Note: Lines in green are recommended steps that make your image curation and publication life easier.

image-project	_____	image-event	_____
image-set-name	_____	image-creators	_____
image-platform	_____	image-sensor	_____
Date	_____	Latitude / Longitude	_____

- ◻ Focus is set to: _____
- ◻ Objective lens is cleaned
- ◻ Camera is set to UTC time
- ◻ Cool / dry air is inside pressure housing
- ◻ Power cable is connected
- ◻ Platform origin x-offset: _____ cm (Forward is positive)
- ◻ Platform origin y-offset: _____ cm (Right is positive)
- ◻ Platform origin z-offset: _____ cm (Down is positive)
- ◻ Camera view angle: _____ ° (0° is horizontal, 90° is vertical down, -90° is up)
- ◻ Laser reference is aligned
- ◻ Protective covers are taken off
- ◻ Storage space on recorder: _____ GB
- ◻ Sample images on deck have been taken
- ◻ Switch on platform: _____ date/time
- ◻ Switch on lights
- ◻ Underwater navigation is ready, Beacon ID: ____
- ◻ Start recording images: _____ date/time
- ◻ Annotation protocol is prepared
- ◻ F-Number is set to: _____
- ◻ Dummy plugs are connected
- ◻ Recorder is set to UTC time
- ◻ All seals are greased
- ◻ Data cable is connected
- ◻ Laser reference distance: _____ cm
- ◻ Glass port is cleaned
- ◻ Photos of setup with scale reference taken
- ◻ Calibration images on deck have been taken
- ◻ Switch on camera: _____ date/time
- ◻ Auxiliary sensors (e.g. CTD) are recording
- ◻ Underwater navigation data is recording
- ◻ Start streaming images to network (e.g. VNC)
- ◻ Start annotation

Comments:

image-acquisition: □ photo □ video □ slide
image-quality: □ raw □ processed □ product
image-deployment: □ mapping □ stationary □ survey □ exploration □ experiment □ sampling
image-navigation: □ satellite □ beacon □ transponder □ reconstructed
image-scale-reference: □ 3D camera □ calibrated camera □ laser marker □ optical flow
image-illumination: □ sunlight □ artificial light □ mixed light
image-resolution: □ km □ hm □ dam □ m □ cm □ mm □ μm
image-marine-zone: □ seafloor □ water column □ sea surface □ atmosphere □ laboratory
image-spectral-resolution: □ grayscale □ rgb □ multi-spectral □ hyper-spectral
image-capture-mode: □ timer □ manual □ mixed
image-camera-housing-viewport: □ flatport □ domeport □ other
image-time-synchronisation:
image-item-identification-scheme:
image-target-environment:
image-target-timescale:
image-spatial-constraints:
image-temporal-constraints:
image-objective:

Project information (add table to DMP once)

<Project name / ID>	
Summary of Project	*Description of the project scope and workplan.*
Location	*Geographic location of project (e.g. bounding box latitude / longitude)*
Start / End Date	*Beginning and end of the data creation and curation of the project*
Project Lead	*The one person that coordinates the entire project*
Data Representative	*One person that is responsible for the data during the project (Project Lead or designee).*
Collaborating Institutes	*List all partner institutes with lead representative*
Total data volume	*Estimation of the total data volume of new image sets and derived data products.*

(Project = Campaign = Experiment = Cruise)

New image sets to be acquired (repeat table in DMP for each image set)

Data that does not currently exist and will be collected or generated during the course of the project/campaign for the purpose of creating the final data products and/or project deliverables.

<Image set name>	
Data Representative	*This needs to be someone who has overall responsibility for image data acquisition, processing, quality control, documentation, and preservation. This could be the PI or a designee. (See General Project Information -> Data Representative)*
Description	*Is this a one-time image set, or an ongoing series of measurements? What is the actual or planned temporal coverage of the data? What is the actual or planned geographic coverage of the data?*
Collection Method	*E.g., ROV, AUV, OFOS, Drone, Lander, …*
Data acquisition Standards	*Identify any standard protocols or methodologies that will be used to acquire the data, if available.*
Data Format & Type	*Identify the formats in which the data will be generated, maintained, and made available. E.g., JPEG, MOV, MPEG-4, TIFF, ...*
Licenses / Fees	*Will the project need to license any product to collect or process the data?*
Quality checks	*Identify the procedural steps for ensuring data quality of images*
Data Workflow	*Describe the data curation steps or provide an SOP reference*
Backup & Storage	*Describe the approach for backup and storage of the image data and metadata. How many copies will be maintained and how will they be synchronized? Are the data and backups going to be stored in multiple places and on different media types to protect against a single point of failure? Yes / No? If yes, please explain. Who will routinely check to ensure that backups are being done according to the backup plan you just described? Note: this needs to be someone closely associated with the actual storage and management of your project / campaign data.*
Volume Estimate	*How many MB, GB, TB, or PB*
Data Repository	*Identify repositories where you plan to share your data. Indicate if data will be integrated into an existing image set or offered as a new one. → Elements, PANGAEA*
Metadata (Schema)	*Identify the metadata standard that will be used to describe the data → iFDO*
Metadata URL	*A URL to the metadata schema → https://marine-imaging.com/fair/ifdos/iFDO-overview/*
Embargo	*If a request to limit access for a period of time after project completion is needed, please identify the length of the time and explain the reason for the extension.*
Access / Restrictions	*Identify any limitations on access or reuse (e.g., licenses, sensitive data, restricted data, software with license restrictions, etc.) and provide justification for restriction.*
File Name Convention	*Are file naming conventions going to be used that help describe the data? Yes / No?*
Data Lifespan	*Less than 5 years, 5–10 years, 10–20 years, 20–50 years, 50+ years*
Long-term data format	*Is the data format appropriate for long-term preservation (non- proprietary, machine-readable)? Yes / No? If not, please describe how you intend to convert.*
Additional Documents	*Which of the following will be included with your data? Supporting documents (i.e., fieldbooks), standardized logs, standard operating procedures etc..*
DOI	*To be added afterwards.*

Software (remove table if not needed and repeat table for each Software used)

<Software Name> <Software Version>	
Representative	*The person responsible for this software (project lead or lead developer)*
Description	*Provide a brief description of the software / code and its purpose.*
Source	*If the custom software or code can be accessed via an online repository, provide a link.*
License / Fees	*Identify any limitations on access or reuse.*
Maintenance / Support	*Describe the ongoing support and maintenance of the software tool after the project.*
Languages	*Identify the computing language / framework that was used.*
Environment	*Identify the operating system (e.g. Windows, Linux, MacOS)*

Derived Data Products (Repeat table for each Derived Data Product):

<Data Product Name>	
Data Representative	*The person responsible for collating this Data Product*
Description	*Describe the information that will be produced, including its characteristics, temporal scope and scale, and geographic scope and scale, when available.*
Data Format / Type	*Identify the formats in which the data will be generated, maintained, and made available. E.g., digital numeric, imagery, photographs, video, audio, database, tabular data etc. List format(s): e.g., ArcASCII, AVI, CSV, DOC, DOCX, GeoJSON, JPEG, JSON, MAT, MOV, MPEG-4, netCDF, PDF, SEGY, SHP, TIFF, TXT, XLS, XLSX, XML...*
Quality Checks	*Identify the procedural steps for ensuring data quality.*
Data Workflows	*Describe data processing steps or provide a scientific workflow to create the product.*
Backup & Storage	*Describe the approach for backup and storage of the information associated with the research project during the project. See Image set backup & storage for relevant aspects.*
Volume Estimate	*MB, GB, TB, or PB*
Data Repository	*Identify repositories where you plan to share your data.*
Metadata (Schema)	*Identify the metadata standard that will be used to describe the data.*
Metadata URL	*Please provide a URL to the metadata schema*
Non-existent Metadata	*If metadata are non-existent or non-compliant, please explain:*
Embargo	*If a request to limit access for a period of time after project completion is needed, please identify the length of the time and the reason for the extension.*
Access / Restrictions	*Identify any limitations on access or reuse (e.g., licenses, sensitive data, restricted data, software with license restrictions, etc.) and provide justification for restriction.*
Data Lifespan	*Less than 5 years, 5–10 years, 10–20 years, 20–50 years, 50+ years.*
Long-term data format	*Is the data format appropriate for long-term preservation (non- proprietary, machine-readable)? Yes / No? If not, please describe how you intend to convert.*
Additional documents	*Which of the following will be included? Supporting documents (i.e., fieldbooks), standardized logs, standard operating procedures etc..*
DM Resources	*Describe the proposal resources allocated for data management activities for the new data collected as a level of effort, total Euros allocated, or as a percentage of the total project's cost. Resources could include people's time or prop*
DOI	*To be added afterwards*