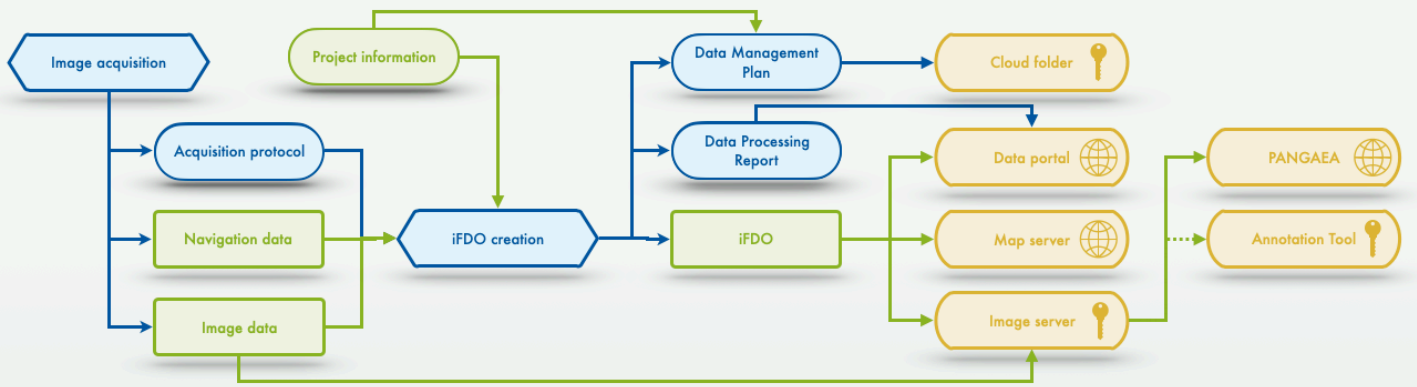
SOP: Image curation and publication

The **purpose** of this Standard Operating Procedure (SOP) is to generally describe how to **publish marine research image data** such as photos and videos for scientific use.

The **goal** of this document is to **enable all scientists** to provide **FAIR and open** image data using the infrastructure available to them.

The **scope** of this SOP includes the steps necessary to **make existing images** (photos and videos) **available** to open scientific use **after acquisition**. This includes a) providing a core set of image *metadata*, b) saving files and metadata in a *central storage* location with backup and long-term archival, c) *disclosure* of image existence through public databases by publishing the image metadata, d) *enabling access* to image data for authorized stakeholders, e) enabling *scientific interpretation* of image, and f) FAIR and open *publication* of the image data itself.

Completing the SOP will **result** in **well-curated image** data that adheres to **open standards**.



Overview of a generic image curation and publication procedure. In **blue: processes** and process documentation, in **green: data** entities, in **yellow: infrastructure** to manage image data. The infrastructures can be access-controlled (key symbol) or public. The image acquisition is the sole responsibility of the scientists. Ideally, an RDM team provides support for executing the workflow steps. The supplement material of this SOP provides templates for acquisition protocols and data management plans. Each image acquisition should trigger an independent execution of the SOP workflow. That means that for each acquisition – i.e. image data set – also an acquisition protocol, an iFDO, and data processing report are created. The iFDO metadata format and file facilitate FAIRness of the image data. Creating a valid, quality-controlled iFDO is the most important part of the workflow.

Preparation phase

1. Scientists create a data management plan as a living document. → FAIR

   Result: *A **docx file in a cloud folder** that all project partners can access.*

   Info: Almost all third-party funding nowadays requires fundees to maintain a data management plan (DMP). This is a living document that ideally starts its life cycle before handing in a research proposal and is continuously updated until beyond the project lifetime to document the data lifecycle. Image data that is being acquired needs to be documented in the DMP. The supplement of this SOP contains an example DMP for image data.

2. Scientists collect basic project information in machine-readable format. → FAIR

   Result: *A **json/yaml file in a Git repository** that all data stakeholders can access.*

   Info: To create an iFDO, information about the context of the image acquisition, i.e. the project information, is required once. This information can be copied e.g. from the proposal or a cruise's web information. Required are the following fields `image-project` (Number / name of the project or experiment within which these images were created. Could be a cruise number, e.g. "SO268"); `image-context` (Wider context of the project or experiment or cruise, e.g. "Mining Impact 2"); and `image- pi` (Full name, email and ORCID of principal investigator (project lead), not necessarily the data creators.).

Acquisition phase

3. Scientists document each image acquisition by an acquisition protocol. → FAIR

   Result: *A **PDF** alongside the images **on an external disk**.*

   Info: Detailed digital documentation of each image acquisition is crucial. This includes core information on the gear used as well as information on the intent to acquire the images. The acquisition protocol should be made available publicly in digital format (e.g. as a scan or photo of the manually populated sheet).

   Tip: See supplement to this document for a template.

4. Scientists acquire images, navigation data and metadata.

   Result: ***Terabytes of images on an external disk.***

   Info: Acquiring image data is the sole responsibility of the scientists. You know how to best acquire data for your research purposes. You know the procedures, the best-practices, the optimal settings. The RDM team will not interfere with your proven scientific demands.

   Tip: If you assume there is potential for improvement of your image acquisition procedures, you can ask the MareHub working group on Videos/Images or the international Marine Imaging Community for input.

5. Scientists structure the data. → FAIR

   Result: *Data is stored in a **consistent file and folder structure on disk.***

   Info: Keeping track of data in the future is supported by a consistent file naming scheme and folder structure. You can create your own, but then stick to it for all your data.

   Tip: See supplement to this document for a recommended folder structure template.

6. Scientists curate the image data. → FAIR

Result: *Terabytes of **properly named images,** containing a UUID metadata value, **stored on a network drive.***

Info: Image data is huge and unstructured. Name image files properly, ideally already during acquisition. A valid image name includes at least an identifier of the acquisition event (usually including the project number), an identifier of the sensor used for acquisition and the UTC date and time of acquisition.
An example could be: `SO268-1_21-1_OFOS_SO_CAM-1_20190304_083834.JPG` where the red part is the event including the project, the blue part is the sensor identifier and the grey part is the date and time information.
To make images FAIR, it is also essential, that each image file contains one specific value in its metadata header: a random UUID which uniquely identifies the specific file. You can create this UUID yourself.
You should also consider removing images from your data set that were taken during pre-acquisition checks or with erroneous settings that led to corrupt data. No need to publish those but don't be too strict, other users might still be interested in the data.

Tip: To make working with big image data more manageable, split huge video files to chunks (e.g. of 15 minutes length) and split huge image folders to subfolders (e.g. thousand files per subfolder). Your operating system and analysis software speeds will increase.

7. Scientists curate the navigation data. → FAIR

Result: *A **machine-readable file** (yaml, json, xml, csv, …) containing curated **4D navigation** data for each photo or second of a video, stored alongside the curated image data.*

Info: All FAIR image data needs 4D acquisition coordinates (3D spatial, 1D temporal) and an estimate of the spatial coordinate uncertainty. Times must always be given in UTC. Images of experiments in the lab are assigned the location of the lab. Images of samples are assigned the sampling location. Images of fixed observatories are assigned a fixed coordinate. Images of moving platforms are assigned a varying coordinate. Should underwater navigation not be available use the next-best estimate (e.g. ship coordinate). Quantification of navigation data uncertainty is required per image. As raw navigation data is error-prone, a data curation step is essential.

8. Scientists create an iFDO file for the image data set. → FAIR

Result: *A **machine-readable file** (**`event_sensor_date_time`**_iFDO.yaml) containing all metadata for the image set, stored alongside the curated image data.*

Info: This is the most important step to make your images FAIR. See https://marine-imaging.com/fair for details on iFDOs. And see https://doi.org/10.5281/zenodo.5681429 for a detailed SOP on this specific step.

An iFDO file has to contain two sections: the **`image-set-header`** and the **`image-set-items`** section. All metadata that attains the same fixed value for all the images in the data set can be added once to the **`image-set-header`** section. All metadata that attains varying values for the individual images has to be added to the **`image-set-items`** section. The field names in both sections have to follow the standardized iFDO nomenclature. You can add more fields in case there is none that suits your needs for your specific domain metadata. The field names in the iFDO are split into three groups: iFDO core, iFDO capture and iFDO content. The iFDO core fields are required to create a valid iFDO. You have to provide all the metadata values for all iFDO core fields. The iFDO capture and content groups contain optional, yet recommended metadata fields. You will only gain visibility and credit for your image data with the recommended capture fields populated. And you will only have awesome image data in case you also populate the content fields. The capture fields contain information on how and why the images were created. The content features provide information on the pixel of semantic content of the images.

Tip: The MarIQT python package and jupyter notebooks provide a lot of functionality to help you create iFDOs for your data and check their validity. Of course, any other way to create the iFDOs is fine as well.

9. Scientists create a data processing report. → FAIR

Result: *A **human-readable file** (pdf, yaml, html, …) containing all information needed to recreate the curated image and metadata as well as iFDO, stored alongside the curated image data.*

Info: The data processing report collects all information that is needed to recreate the curated image data and curated metadata from the used raw data. In an ideal world this includes the virtual environment used to run algorithms and the exact versions of this algorithms. In a realistic setting this is a description of the tools and parameters used of manual interference, of the applied workflow of processing steps and locations of data.

The data processing report needs to be available to data users, so it is essential to make it publicly available. You have to add it to the dataset upon publication in PANGAEA.

Tip: Base the processing report on this SOP and follow its section structure to prevent re-writing info that is already here and just link to the DOI of this SOP. That way you can keep your data processing report short.

10. Scientists update the data management plan. → FAIR

Result: *__Additional sections in the DMP__ which is stored as a living document in a cloud.*

Info: Update the DMP with all relevant information from the acquisition process and its documentation.

Publication phase

11. Scientists publish the iFDO. → FAIR

Result: *The iFDO file is publicly available for __download__.*

Info: To make your image data FAIR and public it is essential (and required) to provide access to the metadata early on. Use a database like OSIS to share your data early on with your project partners.

12. Scientists publish the navigation data with a web map interface. → FAIR

Result: *Image __coordinates__ are __available__ to GIS applications through __web interfaces__.*

Info: GIS applications like QGIS or ArcGIS can access position data through web interfaces like WFS, WMS. Using such an infrastructure makes it easy to use basic information about image acquisition.

13. Scientists migrate the image data to a public web server. → FAIR

Result: *Curated data is stored with __backup__ in a server infrastructure __and accessible__ through network folders & URLs.*

Info: Copy the image data and metadata from your network drive (or external disk) to a publicly accessible server. Create separate folders for each event / station / deployment / dive / experiment / acquisition. Name these folders properly to include the project / cruise identifier, and the event identifier.

14. Scientists make the image data available for analysis (optional). → FAIR

Result: *__Image__ data is accessible __in an annotation__ tool for semantic annotation.*

Info: Web-based image annotation software, such as BIIGLE, can be used to share image data with others, to mark objects and regions of interest and to conduct analyses.

15. Scientists publish image data in PANGAEA (optional). → FAIR

Result: *Image data and metadata are __migrated to PANGAEA__ and can be referenced by a __DOI__.*

Info: PANGAEA must physically host the image data to provide a DOI for it. This step is optional and only required in case you need a DOI for your image data.

16. Scientists update the data management plan. → FAIR

Result: *__Additional information in the DMP__ which is stored as a living document in a cloud.*

Info: Provide the remaining information, e.g. DOIs, handles to the data management plan.

Tip: Store a static copy (e.g. a PDF) of the final DMP as a reference for future projects and collaborations. Some say, this should be published as well. Consider uploading it to Zenodo to make it citeable by DOI.

Appendix:

References:

- SOP “Image curation and publication” supplement: https://doi.org/10.5281/zenodo.5704844
- 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

Glossary:

DMP: Acronym for Data Management Plan – a living document that collects information on planned and acquired datasets as well as models, software and more

FAIR: Acronym for Findable, Accessible, Interoperable, Reusable – describing the leading principle in data management to increase the value of data

iFDO: Image FAIR Digital Object – a standardized format for the description of image metadata

Image: Photo (still image) or video (moving image)

RDM: Acronym for Research Data Management – referring to the process and also the team of highly-trained people

SOP: Acronym for Standard Operation Procedure – a static or dynamic document describing a sequence of tasks acting on data entities to reach a defined goal

Information about this document:

Title: Image curation and publication

Authors: Timm Schoening

Month, Year: 2021/11

Abstract: This Standard Operating Procedure document describes how research image data like photos and videos can be published for scientific use. The goal of this document is to enable scientists to provide FAIR and open image data.
Its scope includes making images (photos and videos) available to scientific use. This includes a) providing a core set of image metadata, b) saving files and metadata in a central storage location accessible by all authorized stakeholders, c) disclosing image existence through public databases by publishing the image metadata, d) publishing the image data itself, e) making image data available for scientific interpretation through web interfaces.

Note: This SOP provides a general overview of the topic. 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.