EMBRC reference implementation for provenance of biological material, observations, and data

Version 1.0, 2025-02-14

Authors
Katrina Exter (VLIZ, 0000-0002-5911-1536)
Laurian Van Maldeghem (VLIZ, 0000-0003-0663-5907)

Co-authors
Christina Pavloudi (EMBRC, 0000-0001-5106-6067)
Ioulia Santi (EMBRC, 0000-0002-0202-8256)
Cymon J. Cox (Centro de Ciências do Mar, 0000-0002-4927-979X)
Rudolf Wittner (BBMRI-ERIC, 0000-0002-0003-2024)
Jörg Geiger (University of Würzburg, 0000-0002-7689-531X)
Ibon Cancio (University of the Basue Country, 0000-0003-4841-0079)

Executive summary

Here we present our conceptual model of provenance for *biological material, observations, and their related and derived data*, with a focus on the domain of marine biology. In this model we describe the provenance information necessary to describe the who, what, where, when, and how of collected/created/used biological material, the derived data, and observations. By providing these provenance information, one is providing: trustability in the material, data, and resulting science; transparency and traceability of the processes carried out; and reproducibility and reusability of the material and resulting data.

The model is presented as classes and properties that together describe the range of activities, agents, locations, and entities involved in marine biology. An overview of these classes and properties, and an explanation of how these add to FAIRness is provided here, while the technical details and full descriptions can be found in the technical documentation (see our GitHub repo). Implementations of the model in RDF and in XML (following the EML schema) are also given here, again with the technical details provided in the technical documentation. Finally, four realistic worked examples, which are accompanied by metadata files accessible via GitHub, are provided here; these can be followed as templates by anyone wishing to adopt the model.

While this conceptual model was developed with marine biological material and datasets in mind, it is general enough to be applicable to other marine sciences and other types of biology.
What is not included in the model – but which could be included as extensions – are the properties that are necessary to describe activities and entities that are important to specialised areas of research: Artificial Intelligence, scientific imaging, genomics, etc. For these, it is best that the respective communities propose the information that need to be included as provenance, and these information will most likely be specific to those domains. Finally, while we make suggestions as to which properties should be considered mandatory, this is also a matter that is best tackled by the respective communities.

This work started under the EU Horizon project EOSC-Life and continued under the EU Horizon project MARCO-BOLO. The model was constructed under the auspices of the European Marine Biological Resource Centre (EMBRC) and the Flanders Marine Institute (VLIZ).

Table of Contents

1. Introduction

The work that led to this document was done as part of Work Package 6 “FAIRification and provenance services” of the EU Horizon project EOSC-Life. One output of WP6 was a Common Provenance Model (CPM) for “biological material, data generation, and data processing”, with a particular focus on specimens and their related data relevant to the life sciences, and which takes into account the distributed nature of life-science research. The first part of this has since been published as an ISO standard: “ISO/TS 23494-1:2023 – Biotechnology — Provenance information model for biological material and data — Part 1: Design concepts and general requirements”, and other parts of the standard series are currently under development (i,e, for the underlying provenance model representation and for documenting biological material handling). A description of this can also be found in the deliverable D6.2, and it is explained further in Wittner et al, 2022, Wittner et al, 2024 (see also Wittner et al., 2025, for the latest results).

This CPM describes the classes of provenance information that should be provided and how these information should be structured, how to make the provenance information traceable as specimens and/or data move from place to place (lab to lab, repository to repository), how to make this provenance information verifiable, and how to deal with the provenance of sensitive data. However, the specific details of exactly *what* provenance metadata should be provided within these classes is not part of the CPM, as this depends very much on the scientific (and perhaps commercial and legal) domain in which the specimens and data are gathered, and the purpose for which the provenance is collected. Therefore, the work presented here therefore aims to provide that level of detail for the world of (EMBRC’s) marine biodiversity /science. This is *not* a new model of provenance as a technical concept or construct, rather we have created a set of classes and properties that are necessary to describe the provenance of material and environmental samples and the data derived therefrom, and provide some examples of how to implement the model in different formats and with different existing, and commonly-used, schemas and standards.

This model has been constructed by members of EMBRC as part of EMBRC’s involvement in EOSC-Life (in particular in work packages WP4 *Sensitive Data* and WP6 *Provenance*) and additionally with members of the EMBRC Traceability and E-Infrastructure working groups. It has also drawn heavily on EMBRC’s biodiversity observation network EMO BON project as an example to follow.

1.1. What is provenance?

Provenance is the record of the origin of an object and what has happened to it until the current point in time. For biological material (specimens) and environmental material samples, the provenance will describe where they came from, how and when they were created or retrieved, what material processing steps they had been subjected to, who did which work, and when that was done. For the data that are created from the material – images, measurements, etc – that provenance will include a link to the provenance of the source specimen or sample, and will describe how the data were created, what processing steps those data were subjected to, who did which work, and when that was done.

Why is provenance important? When publishing scientific results based on a study of specimens and the analysis of data, it is critical to explain all the work that was done: this enables quality or fitness-for-purpose assessment of the results, provides reproducibility of the results, leads to trustability in the results, and allows for reuse of the data (and any remaining specimen) for other science. It is common practise to include such explanations in a scientific publication, however this does a disservice to research: it is not only the *science* that is important, but also the *data* that led to that science; in our era of big data and publicly-funded science, it is important that data are published so that they can be found *as data in open/FAIR data repositories*. So that the provenance information can also be found and accessed, it is critical that it is made available *along with* the data. This could be in the form of a link to the publication where the provenance is described, but this provenance is then only consumable by humans: someone has to read the paper (is it open access?) and extract the provenance information; but if that human needs to analyse 100s of datasets, this is an impossible exercise. Therefore, this provenance information should be provided in the form of structured, formatted metadata that both humans and machines can read and understand.

Research usually goes through distinct stages, and the work done in these stages can be distributed among different agents. For example, a sample will be collected in the field and processed, data will be recorded from that processing, then that data may undergo subsequent processing stages and produce new data. Each stage of this workflow must have its own provenance metadata to describe the who, what, where, when, and how of that stage, otherwise the link from the result to the original source is lost. When a dataset is published from stage X, its provenance describes stage X, and when a dataset is then published from stage X+1, its provenance needs only to describe stage X+1 *but* it must also point to the dataset and its provenance from stage X. This allows for a greater efficiency of recording the metadata and allows for more clarity of the provenance of each stage.

The CPM of EOSC-Life recommends that to deal with distributed provenance for a chain of data, it is necessary that the provenance of each dataset includes a link to any relevant source dataset – that on which the current dataset is based – and its provenance metadata. This recommendation requires that the provenance is provided in the form of a digital object that is identified with a persistent identifier (PID), and ideally resolvable online: in other words, as a file that can be accessed, and where the link to the provenance metadata is provided in the datasets's discovery/catalogue metadata record and vice versa.

How to link that provenance chain together is part of the recommendations of the CPM/ISO 23494 series. In this document we focus on the *content* of the provenance metadata for any published dataset from any stage X.

1.2. Scope of this work

In this model, we assert specifically what should be recorded for the provenance of *biological material, observations, and their related and derived data*, and how these recorded elements are related to each other. "Biological material and data" refers to physical and digital artefacts, respectively, arising from the study of specimens taken from or otherwise related to the marine environment. Observations of the environment – measurements of water temperature and salinity for example – are often collected along with biological specimens,

and the model therefore incorporates observations. This document also describes the information that should be provided to describe permits related to the collection of, or working on, material samples taken from the field: this forms part of the regulatory or legal provenance of the material samples and derived data.

We provide four worked examples of the implementation of the provenance model as *metadata,* i.e. as the data that is intended to describe the provenance of some material or scientific data according to the model. Note that the focus is on the metadata related to provenance, and *not* that related to data discovery – which is why fields such as licence, title, publisher, etc are not included.

2. The provenance model

The model was created by first breaking down the activities performed during an EMO BON[1] sampling campaign into a series of steps and then identifying what provenance information was necessary for each step. Next, we considered how to group and link this information into classes. Here we would like to add a proviso: although we have consulted with field-based marine biologists to ensure that the model is inclusive, we are not experts in all of marine biodiversity science. We are more than open to taking corrections and making additions to the model if requested[2].

The classes and their properties that constitute this provenance model are described below, followed by an explanation of how they contribute towards the provenance for biological material, observations, and derived data. A description of two ways to implement the model, in EML 2.2.0[3] and RDF (JSON-LD), follows. Finally, the provenance metadata following the model for five fictive datasets are described and examples in text, JSON-LD (RDF), and EML (XML) for these can be accessed from our GitHub space. The tabulation of technical details of the model and its implementation can be found in the technical documentation (see our GitHub repo).

2.1. The classes of the model

The provenance information is grouped into four areas, being also four abstract classes: information about entities (<Entity>), activities (<Activity>), agents (<Agent>), and locations (<Location>). The classes of these groups are presented here in Figs 1–3, and are described in more technical detail in tables 1–3 of the technical documentation.

The classes are the following:

<Entity> includes the following:

- <Device>, including <SamplingDevice> and <ObservingDevice>, which are used to carry out a physical activity; these can include anything from a bucket to a scanning electron microscope

[1] European Marine Omics Biodiversity Observation Network
[2] Issues can be raised on https://github.com/vliz-be-opsci/embrc-prov-model
[3] Ecological Metadata Language

- <Permit> are for the legal/regulatory permissions to carry out an activity or to use a sample
- <Platform> on which instruments and devices are situated or from which they are launched
- <Protocol> being the procedures followed when performing an activity
- <Sample> being the physical material that is the subject of the activity
- <Software> as used in an activity or a device

<Agent> includes:
- <Person> who carried out the activity
- <Organisation> who is responsible for the person, or for an activity

<Location> consists of information about the location of an activity

<Activity> is split into two further abstract classes, <PhysicalActivity> and <DigitalActivity>, and between them these include:
- <DataAcquiring> is the act of acquiring data from another source (e.g. a person, a data archive)
- <MaterialAcquiring> is the the act of acquiring samples from someone/somewhere else (often this can involve transporting)
- <Transporting> is the act of transporting samples
- <MaterialProcessing> is the act of processing material samples into subsequent material samples or into digital samples (i.e. data)
- <Observing> is the act of making observations and measurements
- <Sampling> is the act of collecting samples in the field
- <DataProcessing> is the use of software, computations, or any manipulations/filtering to produce and process data
- <Storing> and <Biobanking> are the act of storing, for the shorter-term (storing) or more formal biobanking[4].

The subclasses <SamplingDevice> and <ObservingDevice> of <Device> were created to align with SOSA practices, i.e. to fit better with our RDF implementation of the model (described later), but also to accommodate their being conceptually different types of linked activities. Only one of the three needs to be declared for any particular device.

[4] According to ISO 20387, biobanking is: the process of acquisitioning and storing, together with some or all of the activities related to collection, preparation, preservation, testing, analysing and distributing defined biological material as well as related information and data.

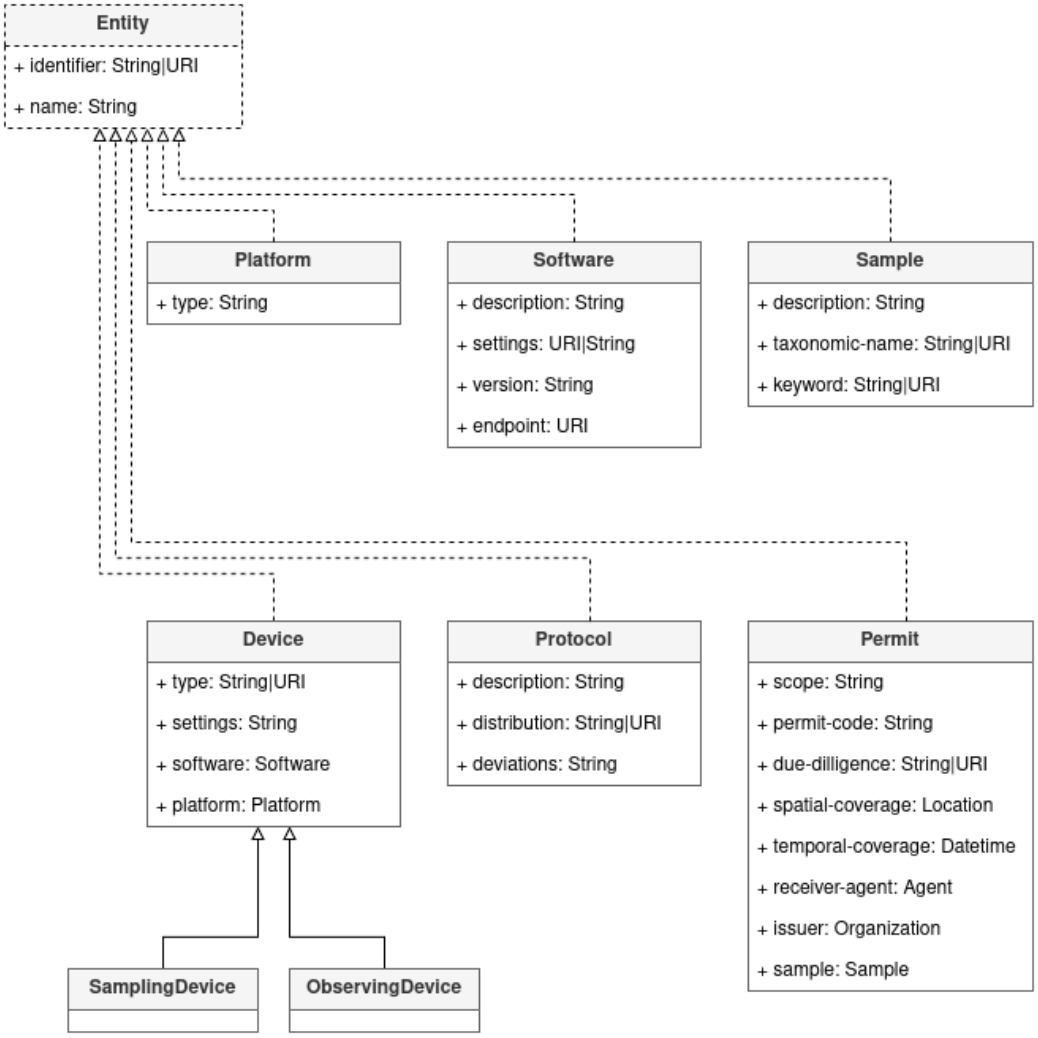

Figure 1 Class diagram of Entity classes.

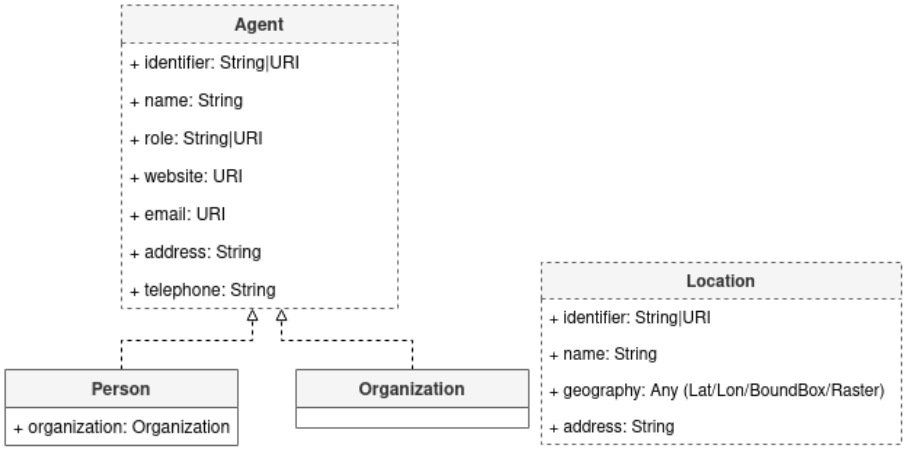

Figure 2 Class diagram of Agent and Location classes.

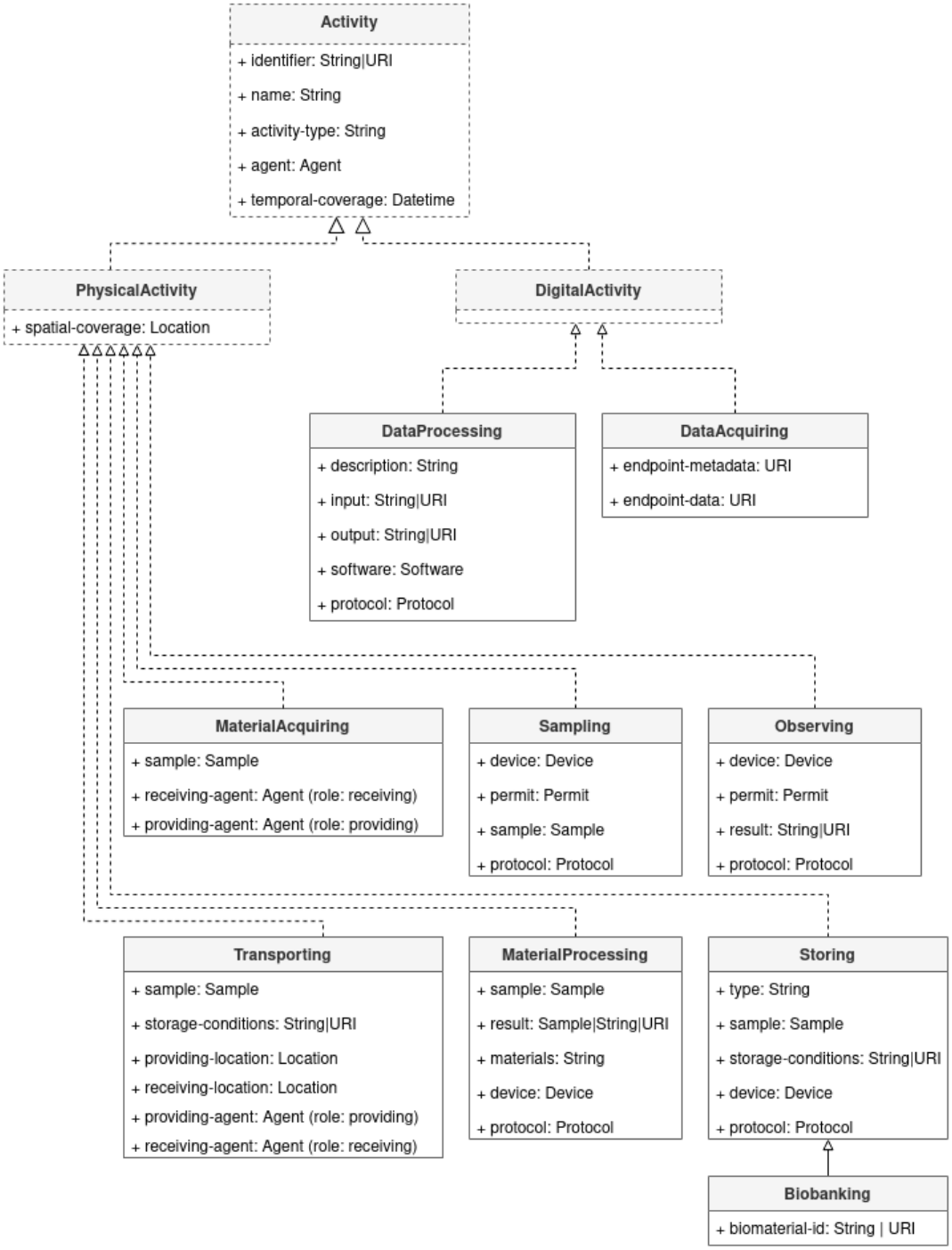

Figure 3 Class diagram of Activity classes.

2.2. The properties of the classes

The properties of each class can be seen in the class diagrams of Figs 1–3, and these properties are elaborated on in more detail in tables 4–6 in the technical documentation (see our GitHub repo): a description and recommendations on the data types to be used for each property. Please see those tables for the lists of the properties for each class. Here, some of the properties that are found over many classes or which need a longer explanation are highlighted.

Comments on properties that are in common to many classes

- The property *identifier* – is found in classes of the model. The role of *identifier* is to allow the Entity or Activity to be identified, within the metadata record itself (i.e. to be able to refer to it from elsewhere) and from external metadata records (which requires that the identifier is globally unique). An *identifier* identifies a thing, but more than that, it should point to only *one* defined thing. This means (1) identifier should be unambiguous, removing any confusion about what that thing is, and (2) an identifier should also be persistent, and (3) ideally also resolvable on the web. For more clarity on identifiers, and examples of identifiers that should be used in *identifier*, see Sec. 2.5. If you do not have an identifier that is permanent and globally unique, it is acceptable to create an identifier that is at least permanent and locally (i.e. within your lab, within the dataset) unique: for example, material sample IDs for samples collected during sampling activities.
- All activities also have the property *name*: this can be used to add some human understanding to complement the *identifier,* and can be used in cases where it is impossible to come up with an identifier. It can also be used to "tag" any Entity or Activity within the metadata record so it can be referred to within that record (see the example datasets described in Sec. 4, and in particular the related examples provided in GitHub). However, a *name* cannot be used to refer to any element in a provenance metadata record from other metadata records since it is unlikely to be globally unique or globally findable (e.g. a name will never be a resolvable web address).
- For some classes (<Permit>, <Protocol>, <Software>) we have both a *name* and a *description (scope* in the case of <Permit>) property: the *description* is to allow the scope of the object to be described, e.g. "The standard operating protocol for collecting and processing EMO BON water samples covering the year 2023", "This permit covers all collected water samples from the Belgian EEZ in the year 2023 by Nice Marine Station". For <Protocol> it could also be used to contain an actual explanation of the protocol, i.e. a listing of the steps, where that is not provided elsewhere.
- The property *sample* and *result* can be found in many classes and is normally described with the properties of the class <Sample>. However, uniquely for <MaterialProcessing>, the *result* can describe a material sample *or* a digital output: the material processed can produce new material or it can produce data such as images, sequences, etc.

Comments on properties for particular classes

Entity

- In <Sample>, the property *taxonomic-name* is to include the most specific taxonomic information about the sample – ideally all of the taxon name, rank, and their identifier in the taxonomic system being used. A single value or list of values are both acceptable. For samples that are (largely) non-biological (e.g. sediment, water) or for which taxonomic information is unknown (e.g. when collecting for environmental DNA), this information can be provided instead as *keyword*, ideally given using a vocabulary such as the Environmental Ontology.
- <Sample> does not have properties to describe where the sample came from or how it was processed: these information are provided via the various Activity and Entity

classes that cover those particular steps. When a sample is processed to create a new sample, then a new instance of <Sample>, with a new *identfier/name* (etc.) should be created.

- For <Platform> and <Device>, we have both *name* and *type*: the *name* is especially useful to use if the object does not have an identifier but does have a name commonly used in the lab; the *type* provides useful information about the type of device, and if possible should be described using a term from a vocabulary (e.g. NVS).
- For <Device> (<ObservingDevice> and <SamplingDevice>) and for <Software>, the property *settings* is to describe the settings/parameters that the device or software was run with (including both default and configured). These settings can be provided in the form of text, a URI (a DOI to a paper, a link to a file on GitHub), or the name of the file in the dataset where those settings can be found (while this latter is less machine-accessible, it is at least uniquely and permanently human-accessible).
- For <Permit>, note that the *sample* property is for the sample that is relevant to the dataset, not all the possible samples that the permit covers.

Agent, Location

- For <Agent> we have a *role*. If one agent is used in different roles, this will mean that the agent will have to be mentioned differently for each instance. Alternatively, all roles taken by the agent in the work can be included in the text of *role*.
- In <Location>, the property *geography* is to provide location information that is as precise as possible (decimal latitude and longitude, a bounding box, or a defined region such as a Marine Region). If the location is an address – for example, a lab, a biobank – then the *address* property should be used instead.

Activity

- For <MaterialAcquiring>, the *identifier* of the *sample* (which is a property of the class <Sample>) is ideally that of the originating organisation: for example, if material is acquired from a biobank, then the biobank identifier should be used. This is to allow for a global traceability: it becomes possible for someone else to investigate the provenance of the material as held within the metadata catalogue of that biobank. If the sample is given a new, local, identifier after being obtained (e.g. to fit in with internal data management processed), then it is recommended that (1) the incoming identifier is used in the *identifier* property of the *sample*, but that it is also recorded that this original identifier is the “same as” the new identifier, which will be used frm then on within the rest of the metadata record. In a basic implementation of the model (i.e. only text, examples of which are given in Sec. 4) you can just write this down, e.g. as comment in the metadata record saying “This original identifier (original-identifier) has been given a new identifier (new-identifier)”. In the RDF implementation a more complex approach is recommended – see Sec. 3 and the technical documentation (see our GitHub repo) for more detail.
- In <DataAcquiring> we have the properties *endpoint-metadata* and *endpoint-data*. The difference is that the former is a metadata record that describes a dataset – a dataset catalogue metadata entry for example – while the latter is directly a dataset. Either or both can be filled.
- In <DataProcessing> we have two properties that can be used to describe the processing activity: *software* (which is of class <Software>) and *protocol* (which is of

class <Protocol>). Where any code is used to perform the processing – a bioinformatics workflow or image recognition software are the two examples given later – then *software* should be used, so the used *settings*, software *version*, and its *endpoint* (URL from where the software is run) can be given. Where the processing activity uses manual methods – for example, filtering out data points that do not meet some threshold using a spreadsheet app – or where the methodology employed is well explained in protocol document, then the *protocol* property can be used and the details entered via the *description, deviation*, and *distribution*. If both software and non-software post-processing is done, then both can be used.

- In <Transporting> the *temporal-coverage*, which is inherited from <Activity>, should be used to indicate the date shipped and received.

Not all of the properties can be filled for all types of datasets, and that is OK: for example, a "Sample" of sediment will not need a "taxonomic-name" but it will need a "keyword". We do not offer advice on how to manage this here, this being a matter of the *implementation* of the model rather than its actuality. Hopefully the examples provided at the end will make it more clear to the user how they could follow the model for their own datasets.

For some of the properties it may be best to link the information to an external file, rather than writing this directly as metadata. For example, where the settings of the software or device are contained in a file, that file can be given as the metadatum: either the URI where the file can be accessed, or the name of that file which is the part of the data being described.

2.3. More detail on Permit and ABSPermit

Permits are a legal or regulatory requirement. Marine biologists who do sampling from the sea will be familiar with the Nagoya Protocol on Access and Benefit Sharing, and the need to obtain other permits to sample in sensitive or commercial areas will also be familiar. How does providing information about the permits obtained to sample or otherwise acquire material, or to carry out a particular experiment, fall under provenance? Essentially, providing information that the necessary permits have been obtained provides assurance that the legal or regulatory requirements have been met, and provides the scope under which the biological material can/cannot be *reused*. This provides transparency and traceability.

For permits that are bilateral and/or otherwise do not make their contents public, it is not necessary to provide that information which is private: in this scenario it will be enough to provide information about the type of permit that was obtained, and to give also the permit's code/number/identifier will allow not just for assurance on re-usability, but also will provide the *proof* that this permit does exist.

For permits related to the Nagoya Protocol, ABS (Access and Benefit Sharing) permits, the minimum property that should be provided is the IRCC number and/or the "proof of due diligence". The IRCC number will exist where an ABS permit was obtained. If the country from where the material was obtained is not party to the Nagoya Protocol, or requires no ABS authorisation for the type of utilisation of the genetic resource that is to be done, there will be no IRCC number; instead, one can provide proof of due diligence. At a practical level,

to provide that you exercised “due diligence”, you need to explain, and show documentation, that proves (1) you know whether the genetic resources you will utilise fall, or do not call, within the scope of the ABS regulation, (2) that you have ascertained whether (or not) the provider country is a party fo the Nagoya Protocol and has legislative and administrative processes in place, and (3) that you did your best to contact the competent authority of that country to obtain an ABS permit, or to be told that this was not necessary (for example a copy of the email, letter, or the link to the web page in which these facts are stated). The other properties that are included – the *spatial* and *temporal* information, the *sample* to indicate which material was the subject of the permit, the agents – that provide fuller assurance of compliance with the permit’s requirements, and information about the scope of the permitted utilisation.

For more information, see the document FAIR Data Management and the Nagoya Protocol that we produced as part of our EMBRC WP6 work in EOSC-Life.

2.4. How this model contributes to provenance

We are well aware that the model is quite demanding compared to current practices, and many will find it difficult to follow the model simply because the information is not routinely collected. However, each of these properties has a good reason to be there: *one has to remember that providing provenance information is necessary to allow the steps taken to be re-traced and repeated, hence to allow the data to be fully trusted and to be re-used for other scientific analyses*.

How do the classes and properties of this model contribute to providing provenance? Provenance information for physical material allows the subsequent user of that physical material (or any replicates thereof) to know how it was collected and processed, from where and when it was collected, how and for how long it was preserved and stored, whom to contact with any questions about that material. The putative user of this material needs to know these facts so that they can determine whether/how they can re-use that material in their project, so that they can reproduce the previously-obtained results, and so that they can trust that the scientific results obtained from the material have a good foundation. By providing provenance information you are making it possible for data to be traced back to the source material from where it came, through all the processes that it underwent, which is particularly useful in the case of *a-posteriori* discovery of mistakes, protocol or device mishaps, etc. Providing information not only allows provenance facts to be established, but the very fact of providing the information allows for a greater degree of trust in the data (“Since you told me who did the work, I believe that the work was actually done”). Provenance information for digital material (data) is the same: in order to be able to trust and reproduce scientific results, or to produce new results, the putative user needs to know what was done to create those data and what processing was done to that data. Providing the information about the provenance provides trustability, traceability and transparency (auditability), and reproducibility, and it enables fitness-for-purpose of the result to be assessed.

The following table outlines the particular role that the model plays in providing provenance.

Table 1 The role of the classes and properties in providing provenance

Class	How this contributes to provenance
<PhysicalActivity>	Providing the details of the *agents* carrying out the activity allows for transparency and traceability (as well as for acknowledgement of work done). We acknowledge that it is unusual to include the property *role* for the agent who carried out the activity, but this provides traceability and to some extent also trustability. The *activity-type* allows context to be given to the activity as described in the metadata file where the provenance is being written – to allow the sequence of events to be understood, especially where multiple activities are recorded. The *temporal-range* of the activity is to provide traceability.
<Sampling>	For Sampling activities, *permit* is a property for reasons described previously: to provide assurance that the legal and regulatory requirements were met. The *sample* property allows the sampling activity to be linked to the actual material sampled, so it is clear to what physical object the metadata applies. While the type of used *device* may be described in the protocol, describing the particular device(s) actually used allows for traceability: e.g. if it later turns out that an instrument has a fault, it will be possible to link this information to material/data that were created with this instrument. For physical activities to be reproduced and trusted, it is necessary to provide the *protocol(s)* used to carry out the physical activities, and to say where and when the activity took place.
<MaterialAcquiring>	This class is for material obtained from somewhere else, i.e. not by you – e.g. ordered from a biobank, taken from the lab's long-term storage, material "created" in the lab. This allows for traceability.
<Storing>, <Biobanking.	When material is stored, it is necessary to provide information about the storage *device*, *times*, and *conditions*, to provide assurance that the material is viable. This is important as general information to provide, but it also allows for traceability, trustability, and reproducibility. (Ideally, the full track of the storage conditions for the timeframe of storage is provided, however we are aware that this rarely collected.) Describing the *sample* being stored allows this storing activity to be linked to the sample (and its properties as described in the *Sample* class). Alternatively, for material obtained from a biobank or long-term storage, its *biobank-id* should allow for that same linking to be made. For the activities to be reproduced and trusted, it is necessary to provide the *protocol(s)* used to carry out the physical activities, and to say where and when the activity took place.
<MaterialProcessing>	Describing the processing done to material is necessary for trustability and reproducibility. The *sample* (or samples) allows one to link the processing activity to the material being processed. While the *device* may be described in the protocol, describing the particular device(s) used allows for traceability (see also *Sampling*). Providing the *result* allows one to link successive steps in a processing chain, and as such provides forward as well as backward provenance for each step. This helps with traceability. If *materials* used in the processing are not included in the protocol (or the protocol is not otherwise open access), then listing them here allows for trustability and reproducibility. For the activities to be reproduced and trusted, it is necessary to provide the *protocol(s)* used to carry out the physical activities, and to say where and when the activity took place.
<Transporting>	When material is transported, it will be subjected to transport *storage-conditions* that should be described to provide traceability, trustability, and reproducibility. The providing and receiving agent and location information provide likewise. The *sample* allows the link to be made between the transporting and the material being transported.
<Observing>	The action of making observations – e.g. environmental measurements – may use *devices* that should be listed to provide trustability and reproducibility, and the *result* will allow a link to be made between the data and what the data were taken from.

	For activities to be reproduced and trusted, it is necessary to provide the *protocol(s)* used to carry out the physical activities, and to say where and when the activity took place.
<DigitalActivity>	Providing the details of the *agents* carrying out the activity allows for transparency and traceability (as well as for acknowledgement of work done). The *activity-type* allows context to be given to the activity as described in the metadata file where the provenance is being written – to allow the sequence of events to be understood, especially where multiple activities are recorded. The *temporal-range* of the activity are relevant in providing traceability.
<DataProcessing>	Any data processing – be that using complex workflows (e.g. bioinformatics), individual pieces of software (e.g.spreadsheets), simple arithmetics (e.g. averaging), or quality control procedures (e.g. data filtering) – needs to be described to allow for trustability and reproducibility. This information is provided via the *software* and *settings* (where code, workflows etc are used) or *protocol* (where more simple or manual procedures using e.g. a spreadsheet app are used, or where the steps are adequately written in a protocol document).
<DataAcquiring>	Data obtained from elsewhere – e.g. taken from a data portal, provided by a colleague – need to have their own provenance description, and the link to that can be provided in *endpoint-metadata* and/or *endpoint-data*. This allows for transparency and reproducibility. It also means that the current user of the data does not need to provide themself the provenance of that acquired data, they can simply refer to the appropriate endpoint.
<Agent>	Agents should be described sufficiently to provide full transparency and trustability. This can include an *identifier, name, website, email, address, telephone*. We acknowledge that it is unusual to include the property *role*: the reason for doing so is to provide traceability and to some extent also trustability. Note that the agent does not have to be a named person: it can be an institute and/or just a role.
<Protocol>	The protocol should be accessible and ideally also open access; giving its *identifier* or *name*, and *distribution* (where it can be found online) will allow it to be found and accessed. The *description* provides context. This provides transparency and reproducibility.
<Device>	While the type of device to be used in an activity may be written into the protocol, the particular device used should be described. So that the device can be identified (traceability), its *identifier* or *name* (as used in the lab) should be given; its (physical) *settings* and any *software* it includes provide reproducibility. If the device is placed on a *platform* this information provides traceability.
<Platform>	As with Device, a platform should be identified with its *identifier,* and the *type* of platform provides context. This provides traceability.
<Sample>	A unique class to describe samples is part of the provenance model for the following reason: several other classes need to refer to the particular material they are relevant to (*Processing, Transporting,* ..), so this material can be described as this *Sample*. If a dataset is based on more than one material sample, then identifying each one allows for traceability and reproducibility. Some of the properties for a sample – *identifier* (e.g. sampling ID, the biobank ID), the *taxonomic-name* – are necessary for traceability and reproducibility, while *keywords* and *name* provide context, with name additionally being useful to provide a link to other classes that are referring to that sample. Naming the *agent* responsible for the sample (one can use *role* in agent to indicate what type of responsibility this is) contributes to transparency.
<Permit>	This is a class for reasons described in a previous section: to provide assurance that the legal requirements were met. The *permit-code* or *name*, and the *temporal-coverage* and *spatial-coverage*. provide more such assurance, as well as transparency.
<Software>	For software that is used to produce or analyse data or to process material (from within a *Device*) the *name, version*, and *endpoint* provide reproducibility.

<Location>	The Location of an activity can be an *address* (e.g. of the marine station) to allow for traceability and transparency; the *geography* of a sampling or measurement event allows for reproducibility.

2.5. A short word on providing interoperable provenance metadata

For human *and* machine interoperability and reusability of provenance metadata, as well as clarity, it is recommended the chosen metadata terms are taken from controlled vocabularies: use the term "https://vocab.lternet.edu/vocab/vocab/xml.php?jsonTema=446" instead of "pressure", or "http://vocab.nerc.ac.uk/collection/P02/current/RFVL/" instead of "water current".

We also recommend that metadata identifying protocols, instruments, platforms, agents, software, locations, and similar are provided as URIs rather than as simple strings (names): that is, as formatted, unique, and permanent identifiers.

What does this mean, and what is a "permanent identifier"? An identifier…identifies a thing – but more than that, it should point to only *one* defined thing, making it unambiguous and so removing confusion about what that thing is. An identifier should also be persistent: it should point to the same thing over time. This expectation leads to the frequent explicit expansion of an identifier to a "PID" (a persistent identifier): a permanent digital reference that uniquely identifies a thing.

Additionally, the format of the identifier should have a recognisable structure that ensures its *global/universal* clarity – meaning nobody else could accidentally use the same string to point to an entirely other thing. Typical examples of such formats are UUIDs, DOIs, URNs, and URIs.

Finally, identifiers are at their best if they are also resolvable on the web: meaning they are themselves URLs, ie web links, or are down to have a direct translation into onw (as doi:X-Y-Z is turned into https://doi/org/x-y-z). This aspect of identifiers makes the consumable by machines.

Some examples of identifiers that we recommended are used in the *identifier* property in this model are:

- An ORCID[5] is an identifier for people and organisations. "Katrina Exter" could be anyone, while https://orcid.org/0000-0002-5911-1536 is a specific instance of Katrina Exter. The ORCID service is providing representations in various formats that are machine-readable, -understandable, and -accessible, and even include links to a wealth of other information about the person.
- IMO[6] numbers uniquely identify ships. IMO: 9622681 is the research vessel called Simon Stevin, while IMO: 9464807 is the pipe-burying vessel also called Simon

[5] Open Researcher and Contributor ID
[6] International Maritime Organization

Stevin. Note: the usage of the “IMO:” prefix to indicate this meaning is known in (a select part) of the marine community, understanding it would thus depend on that context. Unfortunately the IMO numbers do not come as URIs as well, and so they can only be referred to as a unique, but not web-resolvable, ID.

- The thousands of Argo floats each have their own unique identifier, known as the World Meteorological Organization (WMO[7]) float identifiers, which allow the linking of data, location, and time to a particular float. These identifiers are just numbers, but they can be turned into a URI following a pattern (for example, float 15853 can be found on https://www.ncei.noaa.gov/data/oceans/argo/gadr/data/aoml/15853/).
- For devices: many devices have codes assigned by the manufacturer/seller that could be used. However, these codes will not be unique as the same codes are given to every instrument of exactly the same type produced by the manufacturer. Instead, identifiers should be assigned by the lab/marine station: this string should be globally unique, permanent, and managed by the lab. A term from a vocabulary that describes that type of device is not to be used as the *identifier*, as it is not unique to a particular device; use *type* for this instead.

In addition,

- We suggest using DOIs for documentation, ORCIDs for people, MarineRegion geographical IDs for locations, URIs for software (e.g. GitHub or its record in a software repository), etc.
- If a protocol is multi-step it is best to provide the documentation as a PID (e.g. a DOI, URI, a project identifier) as well as providing its name. For simple protocols (“scooped up the water in a bucket”) or where they are not accessible, a text description will suffice. Note that if any instruments or software are mentioned in a protocol, then those should also be described with the *software* and *device* properties.
- For instruments and platforms, if there are no global IDs (e.g. IMO for ships) then those assigned by the marine station/lab should be used. It is to be encouraged that proper ID management is in place: that IDs are also permanent and unique, and linked to online instrument metadata records of some form.
- For permits: if the permit has an identifier assigned by the permit provider, clearly that should be used. If the permit is a digital record held locally and it can be (is allowed to be) shared, a URI pointing to that file should be given.

2.6. Mandatory properties

It is difficult to stamp certain properties as mandatory or optional in a conceptual model, as this depends much on the specifics of the implementation: of the data being described and on the needs of those using the model. Nonetheless, we can outline those properties that ought to be considered mandatory, or at least strongly recommended, in the domain of marine biology, either because it is useful for the internal tracking of the elements of the provenance metadata, or because they are critical for providing the basic level of provenance information necessary for traceability, trustability, or reusability.

[7] https://community.wmo.int/en/buoy-wmo-identification-numbers

Table 2 The properties of classes that should be considered mandatory

Class	Mandatory (internal reasons)	Mandatory (trust, trace, reuse)
Activity	*Identifier* or *name*: so that this "unit of metadata" can be referred to uniquely and unambiguously within the rest of the metadata (or from another metadata file)	*agent* or *role*: to provide trustability and traceability
PhysicalActivity – Sampling	*sample*: to be able to track a sample as it "moves" through the metadata	*protocol* and *device*: to provide trustability and reusability *temporal-coverage* and *spatial-coverage* (with the exception of sensitive data, e.g. endangered species): to provide trustability and traceability *sample*: to provide reusability and trustability, this being the object on which the activity happened
PhysicalActivity – Observing	*result*: to be able to link the result to the activity that produced it	*protocol* and *device*: to provide trustability and reusability *temporal-coverage* and *spatial-coverage*: to provide trustability and traceability *result*: to provide reusability and trustability, this being the object on which the activity happened
PhysicalActivity – MaterialProcessing	*sample* and *result*: to be able to track a sample and result as it "moves" through the metadata	*protocol* and *device*: to provide trustability and reusability *temporal-coverage*: to provide trustability and traceability *sample* and *result*: to provide reusability and trustability, these being the objects on which the activity happened
PhysicalActivity – Storing, Biobanking	*sample*: to be able to track a sample as it "moves" through the metadata	*storage-conditions*: to provide reusability *biomaterial-id:* to provide reusability and traceability *sample*: to provide reusability and trustability, this being the object on which the activity happened
PhysicalActivity – MaterialAcquiring, Transporting	*sample*: to be able to track a sample as it "moves" through the metadata	*providing-agent* and *receiving-agent*: to provide traceability and trustability *sample*: to provide reusability and trustability, this being the object on which the activity happened
DigitalActivity — DataProcessing	*output*: to be able to track an output as it "moves" through the metadata	*software* or *protocol*: to provide reusability *input*: to provide reusability and trustability *output*: to provide traceability
DigitalActivity — DataAcquiring		either *endpoint*: to provide trustability and reusability
Agent	*Identifier* or *name*: so that this "unit of metadata" can be referred to uniquely and unambiguously within the rest of the metadata (or from another metadata file)	*name (and email or identifier)* or *role*: to provide trustability and traceability
Location	*Identifier* or *name*: so that this "unit of metadata" can be referred to uniquely and unambiguously within the rest of the metadata (or from another metadata file)	*geography* or *name/identifier:* to provide trustability and reusability

Entity	*Identifier* or *name*: so that this "unit of metadata" can be referred to uniquely and unambiguously within the rest of the metadata (or from another metadata file)	
Entity – Software		*endpoint* and *settings*: to provide trustability and reusability
Entity – Sample		*taxonomic-name* or *keyword/description*: to provide trustability and reusability
Entity – Protocol		*distribution* or *description*: to provide trustability and reusability
Entity – Permit	*sample*: to be able to track a sample as it "moves" through the metadata	*permit-code* and/or *due-diligence*: to provide trustability and traceability *sample*: to provide trustability, this being the object for which the permit applies

2.7. Limitations of the model

It is important to bear in mind that for certain types of activities and data, specialised types of provenance descriptions may be necessary. For example, for environmental DNA (eDNA) there are many parameters involved in the sequencing and bioinformatics which are important to document to allow the results to be trusted and (re)used. Many projects are looking at recommendations for the proper documentation of eDNA data, i.e. at what is necessary to provide the who, what, where, when, how – the provenance – of the outputs of eDNA activities (see e.g. Klymus et al., 2024; the eDNAquaplan project; Takahashi et al.). Similarly, the specific information that needs to be documented as provenance for scientific imaging activities, or in AI analyses/processes, will also have their own recommendations drawn up by those communities. The model presented here does not include any such specialised information, but can be easily adapted by anyone to accommodate these in the form of extensions.

3. Practical implementations of the model

In this section two implementations of the model are described: in RDF and in XML following the EML (metadata) schema. RDF is a standard model for data exchange on the web. Being a web standard, machine-readable and -understandable, it is an ideal model to allow provenance metadata to be read and understood by a wide web audience. EML is a metadata standard developed for ecology data. It defines a metadata schema to capture and share metadata about ecological datasets, and it is based on the standard XML syntax. IT is used by global biodiversity data publishers such as GBIF[8] and OBIS[9].

3.1. Resource Description Framework (RDF)

For the reference implementation of the provenance model in RDF, we have based ourselves on well-known standards such as the W3C standard PROV-DM, Dublin Core and

[8] Global Biodiversity Information Facility
[9] Ocean Biodiversity Information System

schema.org. Rather than creating new types or a new ontology, we have defined each of our classes and properties as a combination of existing classes in RDF. This approach increases findability and interoperability since third parties are able to identify and query the property paths of the resulting provenance graph.

PROV-DM serves as the backbone for our reference implementation, with the classes selected to describe provenance information being the core structures of PROV-DM: prov:Entity, prov:Activity and prov:Agent (Figure 4).

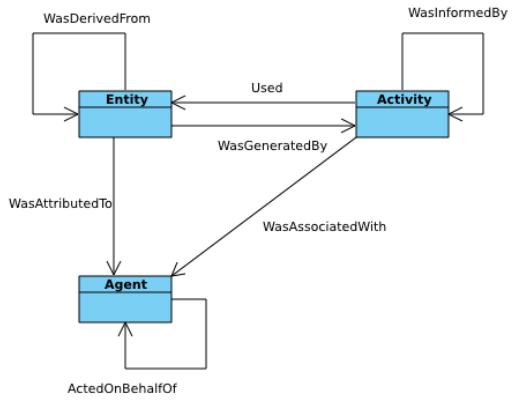

Figure 4 Prov core structures (source: https://www.w3.org/TR/prov-dm/)

Instances of these classes (tables 7–10 of the technical documentation – see our GitHub repo) are described using properties from schema.org and Dublin Core, two well known ontologies that provide general terms for our properties (tables 11–14 of the technical documentation). To describe specific biological aspects of provenance, we used properties of the Semantic Sensor Network.

The properties are our suggestion as to what should be documented in relation to provenance of biological data. The reference implementation has been constructed in such a way that it can be extended with other properties for any use-case not fully covered. We recommend trying to use existing terms as much as possible.

In our GitHub repository we have example files written in JSON-LD, following the example dataset which are described later. These can be consulted as examples of implementing the provenance model in RDF. Note that many of the classes and properties can be instanced in RDF in more than one way: in the example files we chose just one approach.

3.2. Ecological Metadata Language (EML)

This section will be added in the next release of this document.

4. Practical examples of the model

How would this model be applied in practice? Here we give a realistic and detailed demonstration of how the model could be used. We propose a multi-step scenario that produces four different, but related, datasets. For each dataset, we describe the parts of the provenance model that should be used to provide all the necessary provenance metadata for that dataset.

To accompany the four examples here, we provide a plain text, a JSON-LD, and an EML version for each dataset. These can be accessed from our GitHub examples repository. Note that in the JSON-LD files, we have chosen one of the several RDF types that can be used for each class and property (as listed in tables 7–12 in the technical documentation).

The examples written here are quite long: every property of each class is listed even if they are not used; and quite a few of the same properties are used in several places and so the same text is repeated. In the worked examples provided on GitHub, these metadata are *much* more compact.

These are the activities that will create the four datasets:

- Step 1 You go out on a research vessel to collect a sample of water at a depth.
- Step 2 At the same time, you make measurements of temperature and salinity.
- Step 3 You perform a pre-filtering of the water sample (using a large pore-size mesh), split it into two subsamples. One subsample is filtered again using smaller pore-size filter membranes, and resulting filter membrane(s) are placed into tubes. Both subsamples are then put in cold storage on-board.
- Step 4 Arriving at the marine station, the single-filtered subsample is processed through a FlowCam to image the plankton community. The FlowCam produces at set of images.
- Step 5 These images are analysed by the species-identification software within the FlowCam to produce normalised counts of species names.
- Step 6 The twice-filtered subsample is stored in the freezer in the lab, and after some time they are shipped to a genomics facility.
- Step 7 When the genomics facility receives the filters, they are first stored and later processed to extract the DNA. The resulting sequences are stored on their cloud to be shared with others.

Four datasets result from these activities: 1) the log sheet recording data about the sampling activity and the measured values of temperature and salinity; 2) a set of images of the plankton species from the FlowCam, a file with software settings related to the images, and a protocol document that is not provided online; 3) one spreadsheet with the analysis output of the FlowCam, being species names and their normalised abundances, and a file with software settings related to the images; and 4) the sequence on the cloud of the genomics facility. You now want to share these datasets by publishing them for anyone else to find, access, and use. How will these four datasets be described following the provenance model?

In the description that follows, the classes are written as <Class>, and properties are written as *property*.

4.1. Dataset 1: the log sheets

Dataset 1 consists of the log sheets (CSV files) that contain the data collected about and during the sampling campaign, which will include dates, locations, event and sample identifiers, sampling protocol and device names, environmental measurements made at the same time, etc. While we accept that one does not normally publish such log sheets as a dataset, they *are* actually data, and moreover they are the foundation of any subsequent data that is obtained from the physical samples or created with the environmental measurements. Hence the log sheets and their accompanying provenance metadata are the source data for Datasets 2, 3 and 4.

The provenance metadata provided here is related to both the log sheets *and* the resulting material samples. As long as material exists in storage (e.g. water samples or filter membranes) this provenance should be linked to them. Once the material is used and creates digital results (e.g. FlowCam images or DNA sequences), that provenance is transferred to those digital results.

4.1.1. Overview of the steps

Dataset 1 is produced by Steps 1–3. The first step is to collect material, the second to make observations, and the third to process and store the collected material.

Step 1 is a sampling activity, done at a certain place, time, by a person, from a platform and with devices, following protocols, producing some material, and with a permit.

- The class to use for this step is <Sampling> (see table 3 in the technical documentation – see our GitHub repo)
- <Sampling> has properties (see table 6g in the technical documentation), and it also inherits properties from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a)
- Some of these properties themselves need to be described following another class, e.g. the property *agent* points to the class <Agent> which has its own set of properties (table 5b and 5a).

Step 2 is an observing activity by the same person and place as the sampling but with different protocols, producing a set of environmental measurements.

- The class to use for this step is <Observing> (see table 3 in the technical documentation)
- <Observing> has properties (see table 6f in the technical documentation), and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a)
- Here we also have some properties that point to classes, exactly the same as for Step 1.

Step 3 has two material processing activities, followed by two storing activities, done by an agent using devices and protocols.

- The class to use for the processing part is <MaterialProcessing> (see table 3 in the technical documentation)
- <MaterialProcessing> has properties (see table 6e in the technical documentation), and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a).
- The class to use for the storage part is <Storing> (see table 3 in the technical documentation)
- <Storing> has properties (see table 6h in the technical documentation), and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a).
- Here we also have some properties that point to classes, exactly the same as for Step 1.

4.1.2. Writing this down

Here we write out the metadata that describe the provenance of Steps 1–3, that is, the provenance that needs to be provided with Dataset 1. There is a lot of repetition in this worked example because we want each segment to be complete.

Step 1
<Sampling> (inherits from <PhysicalActivity>*)*

- *name*: water sampling
- *activity-type:* sampling
- *spatial-coverage (class* <Location>*)*
    - *identifier*: http://marineregions.org/mrgid/3293
    - *geography:* latitude: 51.249995, longitude: 2.85327
    - not included: *address* (not applicable)
- *temporal-coverage*[10]*:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0001
    - *name*: Jane Smith
    - *email*: jane.smith@somewhere.com
    - *role*: field and laboratory technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *device* (class <SamplingDevice>):
    - *name*: NiceMarineStation rosette#1
    - *type*: Niskin bottles on a Rosette sampler
    - *platform* (class <Platform>)

[10] Can be given just as a start date and end date, or just a single date

        - *identifier*: IMO:1234567
        - *type:* research vessel: https://vocab.nerc.ac.uk/collection/L05/31
        - *name:* RV OurBigShip
    - not included: *settings* and *software* (not applicable), *identifier* (does not have one)
- *protocol* (class <Protocol>)
    - *name*: BigProject_WaterSamples
    - *description*: Collecting and pre-filtering (removal of larger particles) of water samples to be used later for FlowCam and eDNA work
    - *distribution*: https://www.protocols.io/BigProject_WaterSamples.pdf
    - not included: *deviations* (not applicable)
- *permit* (class <Permit>)
    - *name:* ABS permit
    - *permit-code:* ircc2345678
    - *scope*: Samples of sea water from the Belgian EEZ collected over the period 2021-01-01 to 2022-01-01; processed for environmental DNA for taxonomic classification purposes
    - *spatial-coverage* (*class* <Location>)
        - *identifier*: http://marineregions.org/mrgid/3293
        - *geography:* latitude: 51.249995, longitude: 2.85327
        - not included: *address* (not relevant)
    - *temporal-coverage:* 2021-01-01 to 2022-01-01
    - *receiver-agent* (class <Person>)
        - *identifier*: https://orcid.org/0000-0001-0001-0001
        - *name*:Jane Smith
        - *organisation* (class <Organization>)
            - *identifier:* https://edmo.seadatanet.org/report/00000
            - *name:* Nice Marine Station
            - *email:* info@nicemarinestation.eu
            - *website:* https://NiceMarineStation.eu
            - not included: *address* and *telephone* (can be found via the *identifier*)
        - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
    - *issuer* (class <Organization>)
        - *name*: Belgium ABS National Focal Point
        - *website*: https://absch.cbd.int/en/countries/BE
        - not included: *email, address, telephone, identifier* (not relevant/not available)
    - *sample* (class <Sample>):
        - *identifier*: BigProject_belgium_water_10m
        - *description:* water sample to determine plankton community and extract DNA
        - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
        - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)

    - not included: *due-diligence* (not necessary as have provided permit-code), *identifier* (does not have one, *permit-code* used instead)
- *result* (class <Sample>)
    - *identifier*: BigProject_belgium_water_10m[11]
    - *description*: water sample collected from 10m depth
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687[12]
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)

Step 2

<Observing> (inherits from <PhysicalActivity>)

- *name: in-situ* measurement of water properties
- *activity-type:* observing
- *spatial-coverage* (*class* <Location>)
    - *identifier*: http://marineregions.org/mrgid/3293
    - *geography:* latitude: 51.249995, longitude: 2.85327
    - not included: *address* (not applicable)
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0001
    - *name*: Jane Smith
    - *email*: jane.smith@somewhere.com
    - *role*: field and laboratory technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *protocol* (class <Protocol>)
    - *name*: CTD_standard_procedure_2
    - *description*: instructions for using the CTD
    - *distribution*: https://www.mylab.be/CTDinstructions.pdf
    - not included: *deviations* (not applicable)
- *device* (class <ObservingDevice>)
    - *name*: NiceMarineStation CTD#23
    - *type*: CTD: https://vocab.nerc.ac.uk/collection/L05/current/130
    - *platform* (class <Platform>):
        - *identifier* IMO: 1234567
        - *type:* research vessel: https://vocab.nerc.ac.uk/collection/L05/31
        - *name*: RV OurBigShip

[11] This is the material sample ID used by the fictive project collecting our fictive samples - water taken from the coast of Belgium from a depth of 10m

[12] The URL for this is https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1874687, but in this case using the taxon ID recommended by NCBI is acceptable

    - not included: *settings* and *software* (not applicable), *identifier* (does not have one)
- *result:* http://vocab.nerc.ac.uk/collection/P02/current/TEMP, http://vocab.nerc.ac.uk/collection/A05/current/EV_SALIN[13]

Step 3

Four activities are performed in this step: a broad filtering and splitting into two water samples, a further filtering of one subsample, and then storing of each subsample under different conditions.

MaterialProcessing (inherits from <PhysicalActivity>)

- *name:* pre-filtering
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>).
    - *name*: RV OurBigShip on-board lab
    - not included: all other parameters are not applicable
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: field and laboratory technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *protocol* (class <Protocol>)
    - *name*: BigProject_WaterSamples
    - *description*: Collecting and pre-filtering (removal of larger particles) of water samples to be used later for FlowCam and eDNA work
    - *distribution*: https://www.protocols.io/BigProject_WaterSamples.pdf
    - not included: *deviations* (not applicable)
- *sample* (class <Sample>)
    - *identifier*: BigProject_belgium_water_10m
    - *description*: water sample collected from 10m depth
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *result* (class <Sample>)
    - *identifier*: BigProject_belgium_water_10m_1000um_r1
    - *description*: water sample collected from 10m depth and filtered at 1000um

[13] These are the terms for temperature and salinity taken from NVS

    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *result* (class <Sample>)[14]
    - *identifier*: BigProject_belgium_water_10m_1000um_r2
    - *description*: water sample collected from 10m depth and filtered at 1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *device* (class <Device>)
    - *name*: 1000 mesh
    - *type*: nylon mesh of 1000 um pore size
    - not included: *identifier, settings, software, platform* (not relevant)
- not included: *materials* (are included in the protocol)

MaterialProcessing (inherits from <PhysicalActivity>)

- *name:* fine filtering
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>).
    - *name*: RV OurBigShip on-board lab
    - not included: all other parameters are not applicable
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: field and laboratory technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *protocol* (class <Protocol>)
    - *name*: BigProject_eDNAfiltering
    - *description*: instructions for water samples filtering - to capture the biological material on the filter membrane which will be used for DNA extraction
    - *distribution*: https://www.protocols.io/BigProject_eDNAfiltering.pdf
    - not included: *deviations* (not applicable)
- *sample* (class <Sample>)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2

[14] Two replicates are created in this step, hence two resulting samples are described

    - *description*: water sample collected from 10m depth and filtered at 1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *result* (class <Sample>)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2_3um
    - *description*: water sample collected from 10m depth and filtered at 1000um and again 3um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *device* (class <Device>)
    - *name*: membrane filter
    - *type*: polycarbonate membrane filter of 3 um pore size
    - not included: *identifier* (not provided), *settings* (described in protocol), *software, platform* (not relevant)
- *device* (class <Device>)
    - *name*: Peristaltic pump (including pump heads, appropriate silicone tubes, stainless steel filter holders)
    - *type*: Masterflex - EW-07522-20
    - not included: *identifier* (not provided), *settings* (described in protocol), *software, platform* (not relevant)
- not included: *materials* (are included in the protocol)

Storing (inherits from <PhysicalActivity>)

- *name:* cold storage subsample 1
- *activity-type*: storing
- *spatial-coverage* (*class* <Location>)
    - *name*: RV OurBigShip on-board lab
    - not included: all other parameters are not applicable
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: field and laboratory technician
    - *organisation* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *type*: short-term

- *protocol*: not included as there is no protocol, the tubes are just put in cold storage
- *sample* (class *Sample*)
    - *identifier*: BigProject_belgium_water_10m_1000um_r1
    - *description*: water sample collected from 10m depth and filtered at 1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *storage-conditions*: +4 C
- *storage-device* (class <Device>)
    - *name*: Johnson Fridge type XYZ
    - *type*: samples fridge
    - *platform* (class <Platform>):
        - *identifier* IMO: 1234567
        - *type:* research vessel https://vocab.nerc.ac.uk/collection/L05/31
        - *name*: RV OurBigShip
    - not included: *identifier* (not provided), *software* (not relevant)

Storing (inherits from <PhysicalActivity>)

- *name:* cold storage subsample 2
- *activity-type*: storing
- *spatial-coverage* (*class* <Location>)
    - *name*: RV OurBigShip on-board lab
    - not included: all other parameters are not applicable
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: field and laboratory technician
    - *organisation* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *type*: short-term
- *protocol*: not included as there is no protocol, the tubes are just put in cold storage
- *sample* (class *Sample*)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2_3um
    - *description*: water sample collected from 10m depth and filtered at 1000um and again 3um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: t*axonomic-name* (unknown/not applicable), *name* (have identifier instead)

- *storage-conditions*: -20 C
- *storage-device* (class <Device>)
    - *name*: Johnson Freezer type XYZ
    - *type*: samples freezer
    - *platform* (class <Platform>):
        - *identifier* IMO: 1234567
        - *type:* research vessel: https://vocab.nerc.ac.uk/collection/L05/31
        - *name*: RV OurBigShip
    - not included: *identifier* (not provided), *software* (not relevant)

4.2. Dataset 2: the FlowCam images

Datasets 2 (Step 4) and 3 (Step 5) are created by processing one of the sub-samples through a FlowCam. The FlowCam photographs the organisms in the water (passing by as a stream) to produce “raw” images, and the software then analyses these images, comparing them to a library of images with species identifications, to produce a list of species and the normalised abundance. The raw image are Dataset 2 – the images are large and require a specialised data repository and catalogue system to handle them – and the lists of species+abundances are Dataset 3.

Because Dataset 2 is derived from Dataset 1, when Dataset 2 is published it's metadata must contain links to the data and metadata of Dataset 1, as that is the provenance. Bearing in mind that the data and metadata of Datasets 1 and 2 may be published in different online repositories, the links need to be made in the form of URIs or DOIs.

Dataset 2 does not just include the images, but also a protocol file and a settings file. These are documents that in this example could not be provided online (which is likely to be a common scenario) and so they have been included as documents within the dataset. (*How* exactly one would do this is not addressed here: it will depend very much on where these data are published.)

4.2.1. Overview of the steps

Step 4 is the FlowCam work ending with the raw images (i.e. not including the image analysis). This is an activity that starts with material and ends with data. We recommend that in such a case, the class <MaterialProcessing> is used to describe the activity: the *sample* is the sample being processed and the *result* is either a description of the digital results, a URI for that data file/package, or the name of the file(s) that contain the results and which are provided in the data package that you are describing with this provenance metadata. To describe the digital parts of the work, use the property *device*: being of class <Device>, it has the property *software* in which the software and settings can be described.

Step 4 is a material processing done at a certain place, time, by a person, with a device that also has software in it, following protocols and instructions, and producing data as output.

- The first class to use for this step is <MaterialProcessing> (see table 3 in the technical documentation – see our GitHub repo)
- <MaterialProcessing> has properties (see table 6e in the technical documentation), and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a)
- Some of these properties themselves need to be described following another class, e.g. the property *agent* points to the class <Agent> which has its own set of properties (table 5b and 5a).

4.2.2. Writing this down

Here we write out the metadata that describe the provenance of Step 4, that is, the provenance that needs to be provided with Dataset 2. There is a lot of repetition in this worked example because we want each segment to be complete.

Step 4

<MaterialProcessing> (inherits from <PhysicalActivity>)

- *name:* FlowCam imaging
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>)
    - *name*: Nice Marine Station
    - *address*: NiceMarineStation Rd, Ostend, Belgium
    - not included: *geography* (not applicable)
- *temporal-coverage:* date: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: J. Gerschwin
    - *email*: jg@somewhere.com
    - *role*: lab technician
    - *organisation* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *protocol* (class <Protocol>)
    - *name*: FlowCam_1000um
    - *description*: full instructions and settings for using FlowCam with water filtered at 1000um
    - *distribution*: see file FlowCam_1000um.pdf[15]
    - not included: *deviations* (not applicable)
- *sample* (class <Sample>)

[15] Here the protocol is not online, but is provided as a file within the dataset being described by this provenance.

    - *identifier*: BigProject_belgium_water_10m_1000um_r1
    - *description*: water sample collected from 10m depth and filtered at 1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *result*[16]: the first set of images taken by the flowcam are stored in /Users/NMSlab/flowcam/20210101/BigProject/raw
- *device* (class <Device>)
    - *identifier:* NMS_flowcam_2010[17]
    - *name*: FlowCam Micro from CoolLabInstruments Corp., 2010
    - *type*: Flow imaging microscopy: https://vocab.nerc.ac.uk/collection/L05/current/LAB27
    - *settings*: see file flowcam-settings.txt for software and hardware settings
    - *software* (class <Software>)
        - *name*: VisualSpreadsheet FlowCam Micro software
        - *description*: the software that comes with this FlowCam, bought in 2010
        - *settings*: see file flowcam-settings.txt
        - not included: *version, endpoint* (unknown)
    - not included: *platform* (not relevant)
- Not included: *materials* (no extra materials were used)

4.3. Dataset 3: the FlowCam results

Dataset 3 (Step 5) consists of the output of the FlowCam image recognition software: a CSV file with species name and their numbers per unit volume. These data are destined to be published on a biodiversity data portal, and hence constitute a different dataset to Dataset 2. Some of the provenance for the activity of Step 5 is the same as that for the data processing part of Step 4, because the device being used is the same, but here we are describing the parts of the digital work that occur after the initial raw images have been taken. This time we are using <DataProcessing> to describe this work, as at this point in the FlowCam process we are starting with data, not material (water).

Because Dataset 3 is derived from Dataset 2, when Dataset 3 is published it's metadata must contain links to the data and metadata of Dataset 2, as that is its provenance. Bearing in mind that the data and metadata of Datasets 2 and 3 may be published in different online repositories, the links need to be made in the form of URIs or DOIs.

[16] In this example, we follow a common scenario of the raw images produced by the FlowCam being stored only for local access

[17] This is the identifier from a properly-maintained instrument database kept by the marine station. It is not online accessible but this ID will allow other lab technicians to know which instrument was used

4.3.1. Overview of the steps

Step 5 is a digital activity, done at a certain place, time, by a person, with a device and its software, using certain settings.

- The class to use is <DataProcessing> (see table 3 in the technical documentation – see our GitHub repo)
- <DataProcessing> has properties (see table 6k in the technical documentation), and it also inherits some from <DigitalActivity> (table 6a)
- Some of these properties are themselves also classes: the property *agent* points to the class <Agent> which has its own set of properties (table 5b and 5a). This is also the case for the following: *software* (<Software>, table 4g)

4.3.2. Writing this down

Here we write out the metadata that describe the provenance of Step 5, that is, the provenance that needs to be provided with Dataset 3. There is a lot of repetition in this worked example because we want each segment to be complete.

Step 5
<DataProcessing> (inherits from <DigitalActivity>)

- *name*: FlowCam species identification
- *activity-type:* data processing
- *description*: identification of species on images from the FlowCam, using the FlowCam software
- *temporal-coverage*: 2021-01-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: J. Gerschwin
    - *email*: jg@somewhere.com
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *input*: raw images from /Users/NMSlab/flowcam/20210101/BigProject/raw, settings in file flowcam-settings.txt (see dataset), and reference library used was FlowCam_ReferenceLibrary_2024
- *output:* images in zip file Plankton_species_image.zip, results in Plankton_species_numbers.csv
- *software* (class <Software>)
    - *name*: VisualSpreadsheet FlowCam Micro software
    - *description*: the software that comes with this FlowCam, bought in 2010
    - *settings*: file flowcam-settings.txt

    - not included: *version, endpoint* (unknown)
- *protocol*
    - *name*: classification QC
    - *description*[18]: Manual filtering of the flowcam outputs, to remove misidentifications. Method steps: (1) visual inspection of the classified images by a first plankton expert, who suggested reclassifications or removal of the images from the dataset (2) visual inspection and confirmation/rejection of suggestions by a second export (3) reclassification in the image metadata and the spreadsheet for those that both experts agreed on, and agreed rejected results removed from the dataset.
    - not included: *distribution, deviations* (not applicable)

A word about some of the properties of <DataProcessing>

- The property *input* includes the reference image library that was used by the FlowCam software to link images to species names. Creating such a reference library is part of the work one can do with the FlowCam. Ideally, this reference library – which is critical to the provenance of the identification of the species from the images taken from the current water sample (because a different reference library may result in different identifications) – is published in an online resource (e.g. an image repository) and can be referred to via its URI/DOI in that system. Alternatively, the library must have a local identifier ("FlowCam_ReferenceLibrary_2024") so that it can be located and provided upon request.
- For providing the software *settings*: the value can be a list of the settings/parameters that were used when the software was run, or the name of the file (that is part of the dataset) that contains that information, or its URI/DOI. The name of the reference image library that is used to do the image recognition is part of these settings.
- The property *output*: when these images being referred to are eventually published, they will get a DOI(s) or URI(s): at that point you would update this provenance metadata to use that/those links instead of the description given here.
- We have here both a *software* and a *protocol*: the intention here is to show how to include a description of a protocol as a series of method steps, rather than referring to a document. In this example, the analyst first uses the software to get the data and then does a visual quality control on the results. This second step is accommodated in a *protocol*, in which the method steps followed are described in the protocol's *description* property.

4.4. Dataset 4: the DNA dataset

Dataset 4 consists of the DNA (sequences) produced by the genomics facility, and is covered by Steps 6 and 7: first the filters that were created on-board are stored in the lab, then they are shipped to the sequencing facility, which in turn then sequences those samples, and finally produces Dataset 4.

[18] The intention here is not to be have a fully realistic description of a post-processing procedure for FlowCam work, rather to have an acceptable description to demonstrate how to do this within the model.

In the penultimate stage of the sequencing, we start with material and end with data. We recommend that in such a case, the class <MaterialProcessing> is used to describe the activity: the *sample* is the sample being processed and the *result* is either a description of the digital results, a URI for that data file/package, or the name of the file(s) that contain the results and which are provided in the data package that you are describing with this provenance metadata. To describe the digital parts of the work, use the property *device*: being of class <Device>, it has the property *software* in which the software and settings can be described.

As Dataset 4 is based on Dataset 1, when Dataset 4 is published it must include in its metadata record the endpoint of Dataset 1 and of the provenance metadata of Dataset 1 (which may be part of the metadata description of Dataset 1 or may be included as part of Dataset 1 itself).

4.4.1. Overview of the steps

Step 6 is a storing activity and then a shipping activity.

- The first class to use is <Storing> (see table 3 in the technical documentation – see our GitHub repo)
- <Storing> has properties (see table 6h in the technical documentation), and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from Activity (table 6a).
- The next class to use is <Transporting> (see table 3 in the technical documentation)
- <Transporting> has properties (see table 6i in the technical documentation, and it also inherits some from <PhysicalActivity> (table 6b), which in its turn inherits from <Activity> (table 6a).
- Here we also have some properties that point to classes, as for Step 1, e.g. *agent* (<Person>, table 5b and 5c).

Step 7 There are many stages in DNA processing, and there will be differences depending on the type of sample, the sequencing facility doing the work, and the type of DNA extracted. In order to not make this document longer than it already is, we will compact Step 7 into the following:

- Samples are received by the sequencing facility and stored until used: this is <Transporting>, <MaterialAcquiring>, and <Storing> (see table 3 in the technical documentation). All inherit from properties from <PhysicalActivity> (table 6b) in addition to their own (tables 6d, 6c and 6h). As before, some of the properties point to classes.
- The samples are then subjected to multiple stages of <MaterialProcessing> (see table 3 in the technical documentation), which for compactness we have condensed into three: cleaning and lysing, quantification, and library preparation. <MaterialProcessing> inherits the properties from <PhysicalActivity> (table 6b) in addition to its own (table 6e).
- The final sample is then put into a sequencer and data in the form of sequences are produced. Following the example of the FlowCam (Step 4), we will document this as <MaterialProcessing>.

- The sequences are then subjected to one stage of quality control. This is a <DataProcessing> step (see table 3 in the technical documentation), which inherits properties from <DigitalActivity> (table 6a) as well as its own (table 6k)

4.4.2. Writing this down

Here we write out the metadata that describe the provenance of Steps 6 and 7, that is, the provenance that needs to be provided with Dataset 4. There is a lot of repetition in this worked example because we want each segment to be complete.

Step 6
Three lab activities are performed in this step: storing the filters created on-board until they can be shipped to the sequencing facility, and then shipping those filters.

<Storing> (inherits from <PhysicalActivity>)

- *name:* cold storage at Nice Marine Station
- activity-type: storing
- *spatial-coverage* (*class* <Location>)
    - *name*: Nice Marine Station
    - *address*: NiceMarineStation Rd, Ostend, Belgium
    - not included: *geography* (not applicable)
- *temporal-coverage:* start: 2021-01-01, end: 2021-06-01
- *agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *type*: medium-term
- *protocol* (class <Protocol>)
    - *name*: BigProject_eDNAstorage
    - *description*: filters to be stored in tubes at -80
    - not included: *deviations* (not applicable), *distribution* (not available/applicable)
- *sample* (class *Sample*)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2_3um
    - *description*: water sample collected from 10m depth and filtered between 3-1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687

    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *storage-conditions*: -80 C
- *storage-device* (class <Device>)
    - *name*: Johnson Freezer type XYZ
    - *type*: freezer
    - not included: *identifier* (not provided), *platform*, *software* (not relevant)

<Transporting> (inherits from <PhysicalActivity>)

- *name*: transporting DNA filters
- *activity-type*: transporting
- *temporal-coverage:* date: 2021-06-01
- *providing-agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *receiving-agent* (*class* <Person>)
    - *name*: SequencingIsUs HQ
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - *address*: Rue DNA 3322, Paris, France
        - not included: *geography* (not applicable)
    - not included: *identifier, email, address, telephone* and *website* (information not provided by the organisation)
- *providing-location* (*class* <Location>)
    - *name*: Nice Marine Station
    - *address*: NiceMarineStation Rd, Ostend, Belgium
    - not included: *geography* (not applicable)
- *receiving-location* (*class* <Location>)
    - *name*: SequencingIsUs
    - *address*: Rue DNA 3322, Paris, France
    - not included: *geography* (not applicable)
- *storage-conditions*: dry ice, -80 C
- *sample* (class *Sample*)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2_3um
    - *description*: water sample collected from 10m depth and filtered between 3-1000um

    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)

Step 7

In this final step we have many stages: the acquisition of material (SequencingIsUs gets the material that was shipped by Nice Marine Station), storing material, several stages of material processing, and finally one stage of data processing.

<MaterialAcquiring> (inherits from <PhysicalActivity>)

- *name:* transferring ownership of DNA filter
- *activity-type*: material acquiring
- *temporal-coverage:* date: 2021-06-01
- *sample* (class *Sample*)
    - *identifier*: BigProject_belgium_water_10m_1000um_r2_3um
    - *description*: water sample collected from 10m depth and filtered between 3-1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have identifier instead)
- *providing-agent* (*class* <Person>)
    - *identifier*: https://orcid.org/0000-0001-0001-0002
    - *name*: F. Lee
    - *email*: f.lee@somewhere.com
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *identifier:* https://edmo.seadatanet.org/report/00000
        - *name:* Nice Marine Station
        - *email:* info@nicemarinestation.eu
        - *website:* https://NiceMarineStation.eu
        - not included: *address* and *telephone* (can be found via the *identifier*)
    - not included: *website* (is same as *identifier*), *address, telephone* (not necessary)
- *receiving-agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, email, address, telephone* and *website* (information not provided by the organisation)

Note: the sample with identifier BigProject_belgium_water_10m_1000um_r2_3um was given a new, internal identifier SiU_BP_0001[19] and that will be used henceforth.

<Storing> (inherits from <PhysicalActivity>)

- *name:* cold storage at SequencingIsUs
- *activity-type:* storing
- *spatial-coverage* (*class* <Location>)
    - *name*: SequencingIsUs
    - *address*: Paris, France
    - not included: *geography* (not applicable)
- *temporal-coverage:* start: 2021-06-01, end: 2021-10-08
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, email, address, telephone* and *website* (information not provided by the organisation)
- *type*: medium-term
- *protocol* (class <Protocol>)
    - *description*: storing filter membranes in dry freezer at -80Cuntil further processing
    - not included: *name, distribution* (not provided), *deviations* (not applicable)
- *sample* (class *Sample*)
    - *identifier*: SiU_BP_0001
    - *description*: water sample collected from 10m depth and filtered between 3-1000um. Note: is same as BigProject_belgium_water_10m_1000um_r2_3um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have *identifier* instead)
- *storage-conditions*: -80 C
- *storage-device* (class <Device>)
    - *type*: Johnson Freezer type XYZ
    - not included: *name, identifier* (not provided), *platform*, *software* (not relevant)

<MaterialProcessing> (inherits from <PhysicalActivity>)
We have four activities here: DNA extraction, quantification, library preparation, and then sequencing.

First

[19] Once the material arrives as the sequencing centre, it is given a new, local, ID. Both are reported here but only the new one will continue to be used in the subsequent steps.

- *name:* DNA extraction
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>)
    - *name*: SequencingIsUs
    - *address*: Paris, France
    - not included: *geography* (not applicable)
- *temporal-coverage:* 2021-10-08
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, organization, email, address, telephone* and *website* (information not provided by the organisation)
- *protocol* (class <Protocol>)
    - *description*[20]: Extraction of DNA from polycarbonate filters using the Alberti et al. (2017) protocol, followed by DNA purification using NucleoSpin RNA kits combined with the NucleoSpin RNA/DNA buffer set (Macherey-Nagel, Düren, Germany).
    - not included: *name, deviations* (not applicable), *distribution* (not provided)
- *sample* (class Sample)
    - *identifier*: SiU_BP_0001
    - *description*: water sample collected from 10m depth and filtered between 3-1000um
    - *keyword*: coastal sea water: http://purl.obolibrary.org/obo/ENVO_00002150, marine plankton metagenome: NCBI:txid1874687
    - not included: *taxonomic-name* (unknown/not applicable), *name* (have *identifier* instead)
- *result* (class <Sample>)
    - *identifier*: SiU_BP_0001_a
    - description: DNA extracted from filter
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *device* (class <Device>):
    - *name*: centrifuge
    - *type*: centrifuge: https://vocab.nerc.ac.uk/collection/L05/current/83/
    - not included: *type, identifier, settings* (not provided), *software, platform* (not relevant)
- *device* (class <Device>):
    - *name*: cryogenic grinder
    - not included: *type, identifier, settings* (not provided), *software, platform* (not relevant)
- not included: *materials* (not provided)

[20] This is an example of a protocol written directly into the provenance metadata, rather than referring to a document

To keep this example realistic, we have listed the main *devices* used in the work but do not give any details (most sequencing companies would not provide you with this information; in reality it is unlikely even to get any information about devices). We have also not included a *protocol* document but rather summarised the steps in the description. This same applies to the following steps.

Second

- *name:* DNA quantification
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>)
    - *name*: SequencingIsUs
    - *address*: Paris, France
    - not included: *geography* (not applicable)
- *temporal-coverage:* 2021-10-08
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, organization, email, address, telephone* and *website* (information not provided by the organisation)
- *protocol* (class <Protocol>)
    - *description*: Quantification of the abundance of DNA using DNA-Binding Fluorescent Dyes: minimum threshold 4 ng (according to the Qubit dsDNA BR (Broad range) Assay kit (ThermoFisher Scientific, Waltham, MA))
    - not included: *name, deviations* (not applicable), *distribution* (not provided)
- *sample* (class <Sample>)
    - *identifier*: SiU_BP_0001_a
    - description: DNA extracted from water filter
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *result* (class <Sample>)
    - *identifier*: SiU_BP_0001_b
    - description: DNA extracted from filter after quantification
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *device* (class <Device>)
    - *name*: Qubit 4 Fluorometer
    - *type*: Fluorometer: https://vocab.nerc.ac.uk/collection/L05/current/113/
    - not included: *settings, identifier* (not provided), *software, platform* (not relevant)
- not included: *materials* (not provided)

Third

- *name:* Library preparation (18S rRNA)
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>)

    - *name*: SequencingIsUs
    - *address*: Paris, France
    - not included: *geography* (not applicable)
- *temporal-coverage:* 2021-10-08
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, organization, email, address, telephone* and *website* (information not provided by the organisation)
- *protocol* (class <Protocol>)
    - *name*: amplicon library preparation
    - *description*: BID strategy, as described in Belser et al. (2023)
    - not included: *deviations, distribution* (not applicable)
- *sample* (class <Sample>)
    - *identifier*: SiU_BP_001_b
    - description: DNA extracted from filter after quantification (18S)
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *result* (class <Sample>)
    - *identifier*: SiU_BP_001_c
    - description: DNA prepared for sequencing platform
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *device* (class <Device>)
    - *name*: T100 Thermal Cycler (BIORAD)
    - *type*: Thermal Cycler: https://vocab.nerc.ac.uk/collection/L05/current/LAB50/
    - not included: *settings, identifier* (not provided), *software, platform* (not relevant)
- not included: *materials* (not provided)

Fourth
<MaterialProcessing> (inherits from <PhysicalActivity>)

- *name:* sequencing
- *activity-type*: material processing
- *spatial-coverage* (*class* <Location>)
    - *name*: SequencingIsUs
    - *address*: Paris, France
    - not included: *geography* (not applicable)
- *temporal-coverage:* 2021-10-08
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs

        - not included: the rest (not provided)
    - not included: *identifier, organization, email, address, telephone* and *website* (information not provided by the organisation)
- *protocol* (class <Protocol>)
    - *name*: NovaSeq 6000 Reagent Kit
    - *description*: SP Flowcell (Illumina, San Diego, CA, USA)
    - *distribution*: https://science-docs.illumina.com/documents/Instruments/novaseq-6000-spec-sheet-html-770-2016-025/Content/Source/Instruments/NovaSeq/novaseq-6000-spec-sheet-770-2016-025/novaseq-system-spec-sheet-html-770-2016-025.html
    - not included: *deviations* (not applicable)
- *sample* (class <Sample>)
    - *identifier*: SiU_BP_001_c
    - description: DNA prepared for sequencing platform
    - not included: *taxonomic-name* (unknown/not applicable), *keyword* (not relevant at this stage), *name* (have *identifier* instead)
- *device* (class <Device>)
    - *name*: JARVIS
    - *type*: Illumina NovaSeq 6000
    - *settings*: Number of cycles = 150
    - *software* (class <Software>)
        - *name*: Real-Time Analysis (RTA) software
        - *description*: Built-in software operating during cycles of sequencing chemistry and imaging, providing base calls and associated quality scores representing the primary structure of DNA or RNA strands, and performing primary data analysis on Illumina sequencing systems automatically. See https://emea.illumina.com/informatics/sequencing-data-analysis.html for more information on the software from this device. Note that in the outputs the Optional sequencing summary metrics have been set to include the “percentage of PhiX”, the “percentage of aligned reads” - the percent of reads that are aligned to PhiX should be close to the percent of PhiX spiked in, the “percentage of clusters passing the filter” and the “cluster density”.
        - not included: *version, settings* (not provided), *endpoint* (specific endpoint to the actual software used not provided)
    - not included: *platform* (not relevant), *identifier* (not provided)
- *result*: SiU_BP_0001_c_data. Digital sequences, ready for QC. Data available upon request to BigProject (in folder BigProject_2021/fromSequencer).
- Not included: *materials* (no extra materials were used)

The result of this final <MaterialProcessing> step is data, being the (raw) sequences from the one sample. In our scenario, these data are kept on the SequencingIsUs cloud, to be made available upon request (they should later be published in an online repository such as the European Nucleotide Archive, but the example is already long enough so we skipped that step!). The description given in *result* allows this to be known. This information is certainly not machine-readable; it is not always possible to achieve that.

<DataProcessing> (inherits from <DigitalActivity>)

- *name*: Illumina filtering
- *activity-type*: material processing
- *description*: QC: filtering of raw data to remove unwanted clusters and produce the final 18s sequences
- *spatial-coverage*: 2021-10-12
- *agent* (*class* <Person>)
    - *name*: DNA technician
    - *role*: lab technician
    - *organization* (class <Organization>)
        - *name*: SequencingIsUs
        - not included: the rest (not provided)
    - not included: *identifier, organization, email, address, telephone* and *website* (information not provided by the organisation)
- *input*: Raw digital sequences SiU_BP_0001_c_data. Data available upon request to BigProject (in folder BigProject_2021/fromSequencer)
- *output:* Filtered digital sequences produced by the filtering step. Data available upon request to BigProject (in folder BigProject_2021/QC/fromIllumina)
- *protocol* (class <Protocol>)
    - *name*: Illumina filtering with proprietary software
    - *description*: Filtering of raw data to remove clusters that have "too much" intensity corresponding to bases other than the call-base. By default, the purity of the signal from each cluster is examined over the first 25 cycles and calculated as Chastity = Highest_Intensity / (Highest_Intensity + Next_Highest_Intensity) for each cycle. The default filtering implemented at the base calling stage allows at most one cycle that is less than the Chastity threshold (0.6).
    - not included: *settings, version, endpoint* (not provided)
- not included: *software* (used protocol instead)

Why did we use *protocol* in the final step rather than *software*, despite the fact that software was used? Mainly this is to give an example of its use in <DataProcessing>; given the fact that our example proposes that all information about the software (*version, endpoint, settings*) are not provided by the SequencingIsUs company, it is acceptable to use *protocol* to describe the series of steps carried out in this processing activity, rather than *software*.

Acknowledgements and references

EOSC-Life has received funding from the European Union's Horizon 2020 programme under grant agreement number 824087.
MARCO-BOLO is funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101082021 (MARCO-BOLO).
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Katrina Exter and Laurian Van Maldeghem would like to extend a huge Thank You to the co-authors and the members of the Open Science Team at the Data Centre of the Flander's Marine Institute (VLIZ) for their patience and assistance.

Klymus, K.E., Baker, J.D., Abbott, C.L., *et al*. The MIEM guidelines: Minimum information for reporting of environmental metabarcoding data. Metabarcoding and Metagenomics 8: e128689. https://doi.org/10.3897/mbmg.8.128689 (2024)

Takahashi, M., Frøslev, T.G., Paupério, J., *et al*. A metadata checklist and data formatting guidelines to make eDNA FAIR (Findable, Accessible, Interoperable and Reusable). In review.

Wittner, R., Holub, P., Mascia, C. *et al*. Toward a common standard for data and specimen provenance in life sciences. Learn Health Sys., 8:e10365 (2024)

Wittner, R., Matej, G., Frexia, F. *et al*. Common Provenance Framework for Multi-organizational Environments, preprint under review, https://zenodo.org/records/14526108 (2025)

Wittner, R., Mascia, C., Gallo, M. *et al.* Lightweight Distributed Provenance Model for Complex Real–world Environments. *Sci Data* 9, 503 (2022). https://doi.org/10.1038/s41597-022-01537-6

Exter et al., FAIR Data Management and the Nagoya Protocol to be found on https://github.com/vliz-be-opsci/embrc-prov-model/blob/main/docs/FAIR_DataManagement_Provenance_and_the_Nagoya_Protocol.pdf, 2023