IV Lifecycle in Detail

This section expands the **IV Lifecycle Overview** of the alignment between the information viewpoint and the data lifecycle. The descriptions uses the **IV Information Objects** and **IV Information Action Types** to a greater extent providing a deeper insight into the processing of information objects by the RI.

The notation for the diagrams in this section is as follows. The rounded rectangles represent IV actions on data and the straight rectangles represent instances of information objects at different stages. The arrow lines link actions and objects as follows: arrows leaving an action connect to IV objects created by the action while arrows entering an action connect IV objects to actions using them.

Data Acquisition

The data acquisition phase encompasses the actions defined for the observation/experimentation, storage, identification and backup of measurements/observations (raw data).

The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data acquisition.

> **Note**
>
> This example is provided for illustrative purposes. The example shows one of many alternatives for performing data acquisition. Other IV actions and IV objects can be introduced at this stage. Additional actions and objects not described in the IV of the ENVRI RM can also be incorporated.

Specify investigation design: Before a measurement or observation can be started the design (or setup) must be defined, including the working hypothesis and scientific question, method of the selection of sites (stratified / random), necessary precision of the observation or measurement, boundary conditions, etc. For correctly using the resulting data, details about their processing, and the parameters defined have to be available (e.g. if a stratified selection of sites according to parameter A is done, the resulting value of parameter A can not be evaluated in the same way as other results).

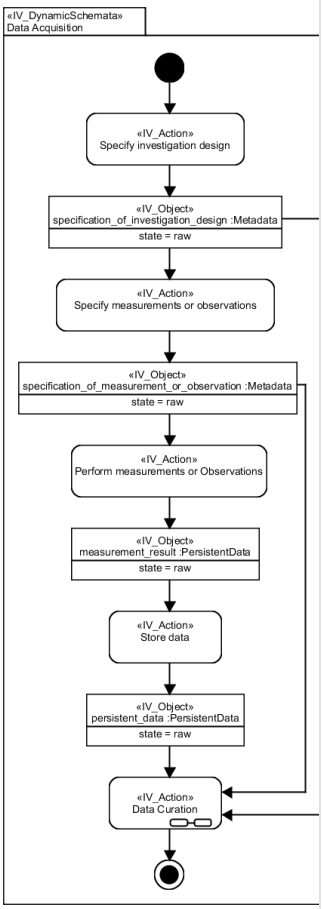
Specify measurement or observation: After defining the overall design of measurements or observations, the measurement method, complying with the design, including devices which should be used, standards / protocols which should be followed, and other details have to be specified. The details of the process and the parameters used have to preserved to guarantee correct interpretation of the resulting data (e.g. when modelling a dependency of parameter B of a parallel measured wind velocity, the limit of detection of the used anemometer influences the range of values of possible assertions).

Perform measurement or observation: After the measurement or observation method is defined, the experiment can be performed, producing measurement result(s) which is a form of persistent data in a raw state.

Store data: The measurement result data is stored.This action can be very simple when using a measurement device, which periodically sends the data to the data management system, but this can also be a sophisticated harvesting process or e.g. in case of biodiversity observations a process done by humans. The storage process is the first step in the lifecycle of data that makes data accessible in digital form.

Data curation: Once data is stored, the next phase of the data lifecycle is data curation.

Data Acquisition



Notation

Data Curation

The data curation phase encompasses the actions that support the the long term preservation and use of research data. The main product of this set of actions is persistent data in a stable state (annotated data). The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

> **Note**
>
> This example is provided for illustrative purposes. The example shows one of many alternatives for performing data curation. Other IV actions and IV objects can be introduced at this stage, for instance: Check quality, Register metadata, or Publish metadata. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

Data Acquisition: The first action is Data Acquisition, the phase of the data lifecycle that precedes data curation. This action produces three IV Objects: PersistentData, SpecificationOfMeasurementsOrObservations and SpecificationOfInvestigationDesign.

Carry out backup: As soon as data are available to the RI a backup can be made, independently of the state of the persisted data. This can be done locally or remotely, by the data owners or by dedicated data archiving centres.

Assign Unique Identifier: Data needs to be uniquely identified for correct retrieval and processing, the unique identifier can be local to the RI or global, to be used from outside the RI. As such it can be a simple numerical value assigned by the RI DBMS or a specific PID assigned following the standards of an external PID provider.

Add metadata: This action uses the specifications of investigation and measurements to facilitate the understanding of the associated persistent data object. In addition to this data the RI can add timestamps, and other identification data as metadata. Once the data is correctly stored and identified, and the corresponding metadata has been also created, persistent data can be linked to metadata.

Annotate data: Data is further enriched with additional metadata which can correspond to a specific ontology for the research field.

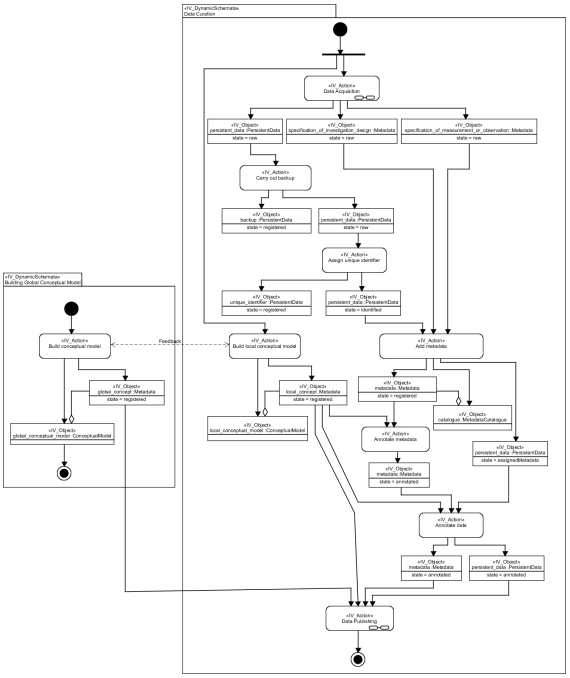
Annotate metadata: Metadata can also be further enriched with additional metadata which can correspond to a specific ontology for the research field.

Build conceptual model: The building of a **local conceptual model** mirrors the wider research community efforts to build a global conceptual model. In this set of activities concept are added to the local conceptual model of the RI. The conceptual model is made of the composition of concepts, which are used to help people know, und erstand, or simulate a subject the model represents. The pairing of data and metadata using semantic annotations creates a local concept (a new metadata object) and changes the state of the persistent data object to annotated.

Global conceptual models are ontologies, thesauri, dictionaries, or hierarchies built by a larger communities than a single RI, such as GEMET, DOLCE, SWEET. This action normally happens outside of the RI's main activities. Trough feedback mechanisms RIs participate in the creation of global conceptual models while developing their own models..

Data Publishing: Once data have been curated, the next phase of the data lifecycle is data publishing.

Data Curation



Notation

Data Publishing

The data publishing phase encompasses the actions that make the data available for entities (people and systems) outside the RI. The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

> **Note**
>
> This example is provided for illustrative purposes. The example shows one of many alternatives for performing data curation. Other IV actions and IV objects can be introduced at this stage, for instance: QualityAssurance. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

Data Curation: The first action is Data Curation, the phase of the data lifecycle that precedes data Publishing. This action produces four IV Objects: PersistentData, LocalConceptualModel, LocalConcept, and Metadata.

Finally Review Data: Persistent data that is in the process of publishing needs to be reviewed before proceeding to publising. It is important to clearly specify what the "finallyReviewed" state means. In some RIs it can mean, that those data will never change again, the optimum for the outside user. For other RIs it might also mean, that only under certain circumstances those data will be changed. In this case it is important to know what "certain circumstances" mean.

Build Global Conceptual Model: The construction of a global conceptual model makes sure that there is an appropriate fit between the persistent data to be published and their metadata (including the local conceptual model) with other models existing outside the RI. The GlobalConceptualModel is the representation of how that outside world looks to the RI.

Semantic Harmonisation: unifies data (and knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability. This complex activity is performed in two stages: setup mapping rules and perform mapping, defined as follows.

Setup Mapping Rule: The Global model is used to generate a set of mapping rules to enable linking the RI data and metadata to global semantics. This may include simple conversions, such as conversions of units, but may also imply more sophisticated transformations like transformations of code lists, descriptions, measurement descriptions, and data provenance.

Perform mapping: This action carries out the linking of data and metadata to one or more global models.

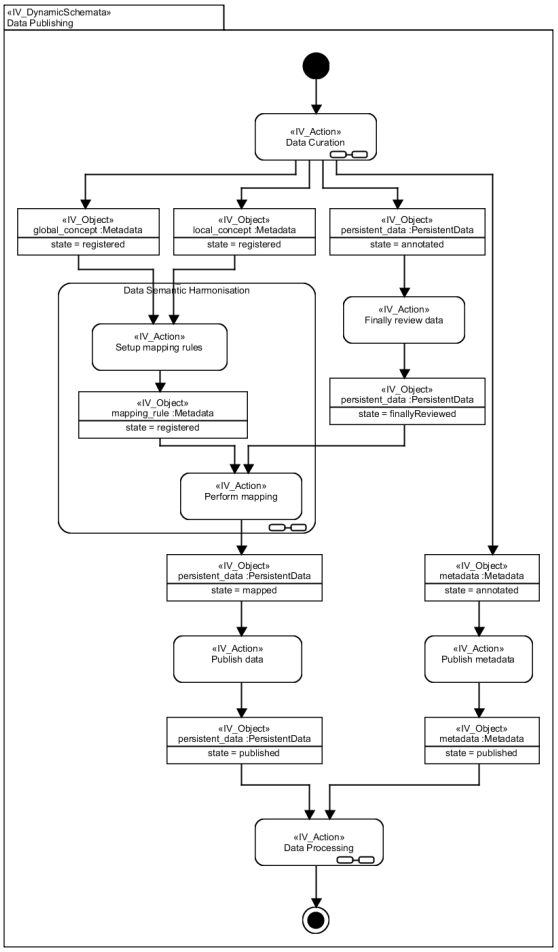
Publish data: Mapped data is made available to the outside world. The PID is the main identifier of the data but the data can also be located by querying metadata.

Publish metadata: Metadata is also mapped and published to enable more sophisticated data querying.

Data Processing: Once data have been published, the next phase of the data lifecycle is data processing.

Data can be made directly accessible or indirectly. Direct access means, that a data request to a data server (query data) gets the data or an error message as answer. Indirect access means, initially accessing metadata (query metadata), searching for a fitting data set and then querying on the resulting data set. Those two steps can be extended further, when intermediate steps are involved. The multi-step approach is often used for data, which are not open, making metadata open but not data itself. For queries touching several data sets and/or filtering the data (like e.g. give me all NOx air measurement where O3 exceeds a level of Y ppb) the multy-step approach can be seen blocker.

Data Publishing



Notation

Data Processing

The data processing phase encompasses the actions that support making use of the RI published data. The following paragraphs explain the detailed diagram of how the IV actions can be combined to support data curation.

> **Note**
>
> This example is provided for illustrative purposes. The example shows one of many alternatives for performing data processing. Other IV actions and IV objects can be introduced at this stage. Actions and objects not described in the IV of the ENVRI RM can also be incorporated.

Data Publishing: The first action is Data publishing, the phase of the data lifecycle that precedes Data Processing. This action produces three IV Objects: PersistentData, GlobalConceptualModel, and Metadata which are used in the data processing phase to access and process data.

Provenance Tracking: Is the action that keeps a log about the the actions and the data state changes as data evolves through the RI systems. The resulting provenance data is a form of metadata which may be of interest for referencing and citing the use of data within and outside the RI.

Query data: This action requests specific persisted data from the RI.

Do data mining: This action implies the execution of a sequence of metadata/data request/interpret/result/request which automatically produce or find patterns in the data being analysed. Usually this sequence helps to deepen the knowledge about the data.

Resolve annotation: This action implies finding a specific data set from a set of semantic annotation and constrains on those annotations. If the annotation is resolved the result should be a link or a set of links to specific data sets, if not the result is an empty set.

Query metadata: This action requests specific persisted data from the RI using metadata as additional parameters for narrowing down the search.

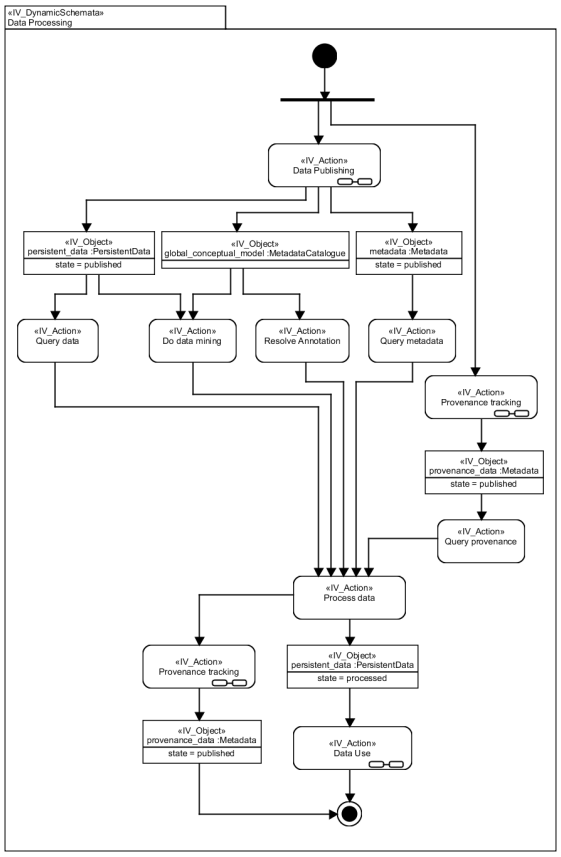
Query provenance: This action requests specific persisted data about the provenance of some data or metadata. This is usually done to determine the origin and validity of data but can also be helpful for citation and referencing.

Process data: The performance of any of the five actions listed before, is automatically detected as a form of data processing by the RI system. This should result in changing the state of the data to "processed". The processed state can mean several things such as: the data has been consulted, the data has been referenced, the data has been downloaded, the data has been used as input for an external process, etc.

Data use: Once data have been processed, the next phase of the data lifecycle is data use, which to some extent overlaps with processing.

Provenance Tracking: As described in the overview, the provenance tracking action tracks all changes in the states of persistent data. This is an important action which has wide use inside and outside the RI.

Data Processing



Notation

Data Use

The data use phase is a bridge phase which sits between processing and acquisition. In this phase, the data is used and may produce new data (raw data) which can in turn be persisted by an RI. The actions that act on data at this point can be provided by same RI exposing the data or by external entities (RIs or other).

In the use phase, the RI system is open to the outside world. Users (persons or external systems) can use the services provided to produce new data products.

> **Note**
>
> This example is provided for illustrative purposes. The example shows one of many alternatives for performing data processing. Other IV actions and IV objects can be introduced at this stage. Actions and objects not described in the IV of the ENVRI RM can also be incorporated. In the diagram and associated descriptions below, cite data, convert data, produce model, and visualise data are some examples of these types of actions.

Data Publishing: The first action is Data publishing, the phase of the data lifecycle that precedes data Processing. This action produces three IV Objects: PersistentData, GlobalConceptualModel, and Metadata which are used in the Data Use phase to access and process data.

Provenance Tracking: Provenance tracking keeps a log about the the actions and the data state changes as data evolves through the RI systems. The resulting provenance data is a form of metadata which may be of interest for referencing and citing the use of data within and outside the RI.

Data Processing: Data Processing produces Persistent Data IV Objects.

Query data: This action requests specific persisted data from the RI.

Do data mining: This action implies the execution of a sequence of metadata/data request/interpret/result/request which automatically produce or find patterns in the data being analysed. Usually this sequence helps to deepen the knowledge about the data.

Resolve annotation: This action implies finding a specific data set from a set of semantic annotation and constrains on those annotations. If the annotation is resolved the result should be a link or a set of links to specific data sets, if not the result is an empty set.

Query metadata: This action requests specific persisted data from the RI using metadata as additional parameters for narrowing down the search.

Query provenance: This action requests specific persisted data about the provenance of some data or metadata. This is usually done to determine the origin and validity of data but can also be helpful for citation and referencing.

Cite data: Produce a reference to persistent data or metadata.

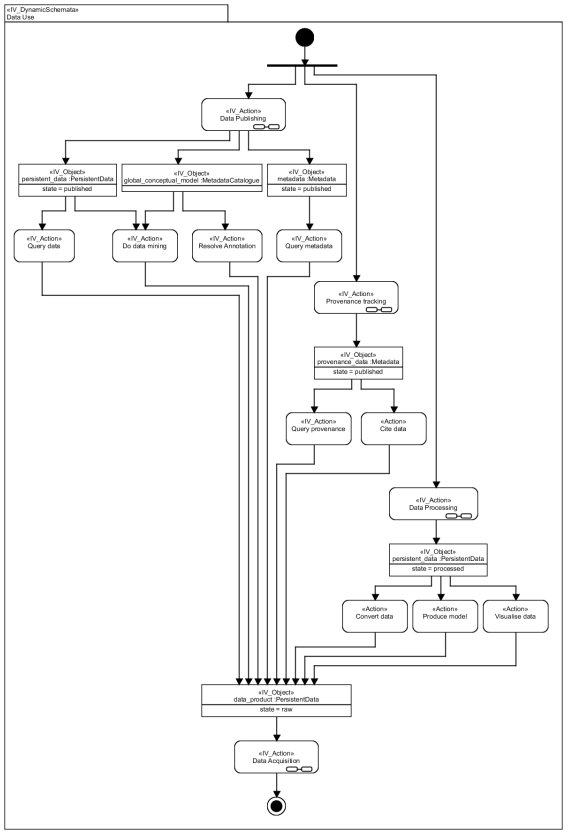
Convert data: converting and generating data products, for instance translating to a different format.

Produce model: creation of statistical models, simulation models or summaries with the data provided.

Visualise data: creating visual models which display data alpha-numerically, graphically, or geographically.

Data Acquisition: Use of data has the potential for creating data products which may need to be persisted, re-initiating the data lifecycle. For this reason, the the last action after Data Use actions is Data Acquisition.

Data Use



Notation