PITHIA-NRF
Plasmasphere Ionosphere Thermosphere Integrated Research
Environment and Access services:
a Network of Research Facilities

# Data Resource Registration

at PITHIA e-Science Centre

*User Guide*

Version 0.5
December 23, 2022

The PITHIA-NRF project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101007599

# 1. Terminology and Abbreviations

Acquisition

[*standard ISO vocabulary*]: Interaction of the *Instrument* with the *Feature of Interest* to obtain its *Observed Properties.* (Step #7 of the registration)

Catalogue

[*standard PITHIA vocabulary*]: A listing of events or investigations assembled to aid users in locating data of interest. Each Entry in a *Catalogue* has distinct *begin* and *end* times and a list of registered Data Subsets with optional DOIs to their persistent storage.

Computation

[*standard ISO vocabulary*]: Numerical calculations without interacting with the *Feature of Interest*; characterised by numerical input and output. (Step #10 of the registration).

Data Collection

[*standard PITHIA vocabulary*]: top-level metadata document for registration of provided measurements and model computations (final Step #12 of the registration)

Data Level

Level of information processing ranging from Level 0 (unprocessed) to Level 4 (derived by secondary analysis of lower-level data or by modelling).

Data Resource

Single data service item and its associated metadata, available through the PITHIA-NRF system.

Dataset

Pre-computed or pre-processed data resource available for download.

DQ

Data Quality [flag].

DQF

Data Quality Flag.

Data Subset

A portion of a *Data Collection* for registration in a *Catalogue* of particular events or targeted investigations

Feature of Interest

[*standard ISO vocabulary*]: Real-world object that carries the property which is observed or modelled to produce a *Data Collection*

GUID

Global Unique Identifier, generated on demand using an algorithm that does not have to consult with a centralised authority to issue the identifier.

ISO

International Standards Organisation

Metadata Model

[*science-neutral*]: Specification of different documents and their contents that are required for registration of data resources

Ontology

[*science-specific*]: A set of standard vocabularies for the selected domain of science

Observed Property

[*standard ISO vocabulary*]: description of a physical *Phenomenon*. *Observed Property* is obtained by means of observation or modelling that generates an estimate of the *Phenomenon's Measurand* value. Technical details of generating *Observed Property* values are described by *Process*.

Phenomenon

[*standard ISO vocabulary*]: A physical observable (a.k.a. “Mother Nature”). Not to confuse with *events*; phenomena are not defined in time or space. The top-level phenomenon categories are Field, Particle, and Wave.

Process

[*standard ISO vocabulary*]: A designated procedure used by the action of observation in order to assign a number, term, or other symbols to a *Phenomenon* generating the observation result. (Step #11 of the registration).

Registration

A 2-phase operation of adding science metadata to PITHIA e-Science Centre data search engine. Phase 1: building XML files describing the data collection (“12 steps”). Phase 2: ingesting the XML files in the e-Science Centre system using its online web submission page.

XML

eXtensible Markup Language: plain-text data format for long-term storing and exchanging of information. Developed by IBM for the task of multi-decade preservation of arbitrary structured data and came into common use for the exchange of data over the Internet.

XSD

XML Schema Definition: a reference document that defines the standard content rules for XML documents beyond their syntax correctness. Schema definitions are open to public access over the Internet to allow testing XML documents for compliance with the XSD rules.

## 2. Background

- All PITHIA-NRF data resources are registered with the e-Science Centre using the International Standards Organisation guidelines for Observations and Measurements (ISO 19156:2011).
- PITHIA-NRF leverages metadata designs for space physics data registration developed by the ESPAS consortium [Belehaki et al., 2016] in 2012-2015.
- At the time of the PITHIA-NRF project performance, the governing ISO standards prescribe using XML as the physical format. Therefore, all PITHIA data registrations are done using XML as the metadata format.
- PITHIA's Metadata Model and Space Physics Ontology are building blocks for the data resource registrations.
- Data providers build XML files using
  - this guide,
  - example XML files as the templates, and
  - learning tools available at PITHIA e-Science Center
    - PITHIA ontology browser
    - PITHIA resource browser
    - PITHIA search engine
- Data providers upload the XML files to the resource registration page of the e-Science Centre where the submission is instantly validated for
  - Good XML syntax of each document
  - Compliance to PITHIA metadata model and ISO/W3 standards
    - [accomplished using XSD schemas]
  - Validity of links to the PITHIA ontology vocabulary terms
    - [accomplished by automatically testing the response of the ontology server to the queries for stated URLs]
  - Integrity of cross-links between provided XML records
    - [accomplished by ensuring the dependencies are all available and valid per the metadata model diagram in Figure 1.]

## 3. Twelve Steps of Data Resource Registration

- Each registration involves preparing up to 12 types of XML documents (Figure 1).
- It is recommended to edit provided document templates for each step. Editing replaces the example content with resource-specific information.
  - Each template is pre-structured according to the PITHIA metadata model.

To edit provided templates, replace the example text with the relevant text:

1. Use free text for descriptions, titles, names, postal addresses, etc.,
2. Use an online ontology browser to find standard URLs to the ontology terms,
3. Link documents to other XML documents in the data collection document set, and
4. Build document identifiers

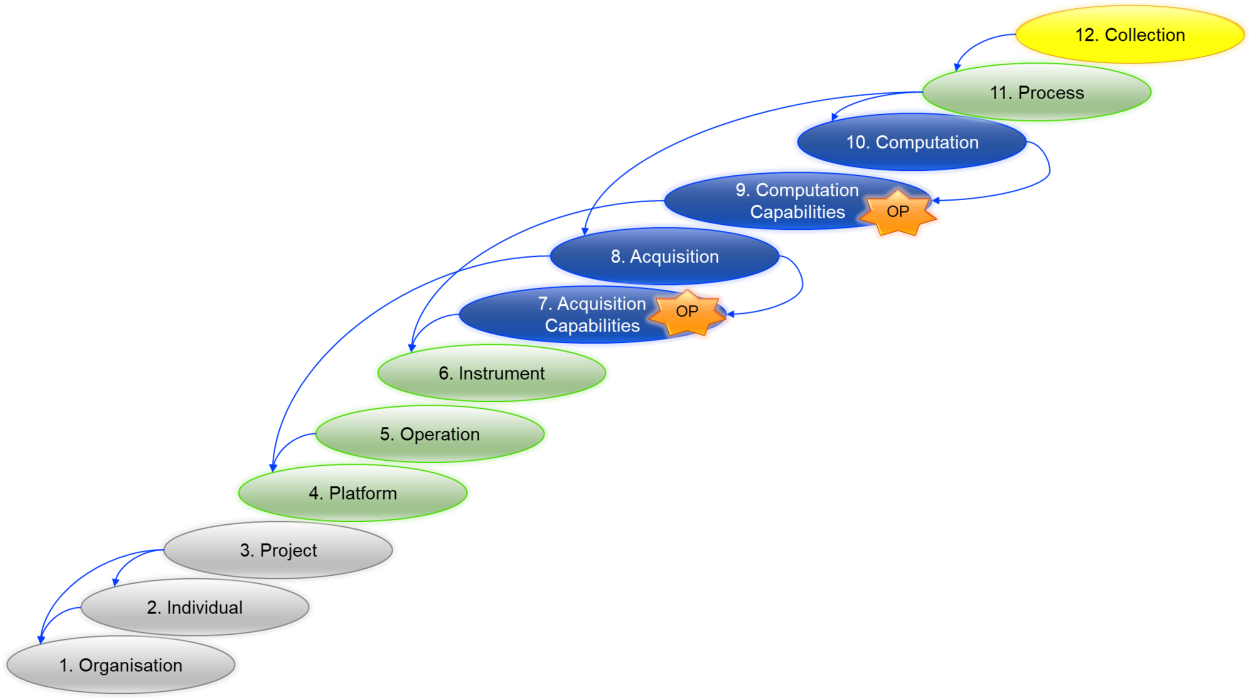


**Figure 1.** Twelve steps of PITHIA data resource registration. Different colours are used to indicate the complexity of the definitions; blue ovals are harder to define. OP is *Observed Property* information commonly used for content-aware searches.

Certain steps of resource registration will require more effort than others; the colour scheme in Figure 1 suggests green background for the easier steps.

**It is important to follow the steps in the correct order.** The PITHIA-NRF validation algorithm will not permit adding a document if it is linked to a missing document.

# 4. Providing standard Identifiers to registration documents

## 4.1. Local PITHIA identifier

Each registration document has a local PITHIA identifier element. For example:

```
<PITHIA_Identifier>
   <localID>Platform_Athens_Greece</localID>
   <namespace>noa</namespace>
   <version>1</version>
   <creationDate>2022-03-14T15:00:00Z</creationDate>
   <lastModificationDate>2022-03-14T15:00:00Z</lastModificationDate>
</PITHIA_Identifier>
```

All items of the standard PITHIA_Identifier are the responsibility of data providers.

### 4.1.1 Local ID

*Local IDs* are selected using the following recommendations:

- Local ID must be unique
- Local ID does not have to be *global unique ID (GUID)*
- Local ID must start with the short name of the resource:

| Step # | Resource | Local ID Prefix |
|---|---|---|
| 1 | Organisation | Organisation_ |
| 2 | Individual | Individual_ |
| 3 | Project | Project_ |
| 4 | Platform | Platform_ |
| 5 | Operation | Operation_ |
| 6 | Instrument | Instrument_ |
| 7 | AcquisitionCapabilities | AcquisitionCapabilities_ |
| 8 | Acquisition | Acquisition_ |
| 9 | ComputationCapabilities | ComputationCapabilities_ |
| 10 | Computation | Computation_ |
| 11 | CompositeProcess | CompositeProcess_ |
| 12 | Data Collection | DataCollection_ |

- The second part of the *Local ID* shall describe the specific resource by its short name.
    - For example, *Organisation_NOA*
- The optional third, fourth, etc. parts of the Local ID can be added to provide further detail of the resource registration under sub-categories.
- Add an underscore separator between subsequent parts of the Local ID.
    - For example, *Computation_Raytracing_Analytical_Midpoint-CQP*
- If any part of the Local ID includes multiple words:
    - Capitalise the first letter of every word in the part
        - e.g., *Computation_IonogramScaling_ConfidenceEvaluation*
    - Consider using a hyphen symbol to separate abbreviations, for example, “DataCollection_RayTRIX-CQP” (rather than DataCollection_RayTRIXCQP”.
- Special case: **Platform** (an observatory, satellite, balloon, ship, aircraft.,.)
    - Do not include standard identifier(s) in the Local ID if the Platform is registered with a particular authority. Such authority-specific IDs (for example, the 5-symbol URSI Code of the observatory) are defined inside the document.
    - Include country for the ground observatories

        - For example, *Platform_Athens_Greece*
    - Include mission name for the satellite payloads, and add “Constellation” for the multi-spacecraft missions
        - For example, *Platform_IMAGE, Platform_ClusterConstellation*

### 4.1.2 Choice of namespace

*Namespace* is used to sort registration documents into organised sets by attributing them to a specific organisation. In most cases, the namespace is defined as a short name of your organisation in the lower case, e.g., *noa* for National Observatory of Athens, with the following exceptions when the documents must be attributed to the top-level *pithia* namespace:

- **Data Collection** documents – use *pithia* to simplify online browsing of all collections
- **Organisation** documents - use *pithia* to avoid “one document in the folder” arrangement
- **General-purpose resources** that can <u>also be found outside your organization</u>, for example,
    - Universally applicable *ComputationCapabilities* such as “basic ionogram autoscaling” should go in *pithia* namespace
        - For example, several autoscaling algorithms extract a basic set of URSI characteristics from ionogram images. References to that basic autoscaling process should be placed in the *pithia* namespace to be used by different ComputationCapabilities descriptions.
        - However, use your custom namespace for *ComputationCapabilities* that are unique to the instrument

### 4.1.3 Naming XML files

XML file name shall be *<Local ID>.xml*

- For example, “ComputationCapabilities_IonogramScaling_Basic.xml”

## 4.2. Persistent, globally unique metadata and data identifiers

### 4.2.1 GUID versus DOI

Globally unique IDs are 128-bit digital labels that are, for practical purposes, unique without depending on a central ID authority. GUID can be generated automatically by e-Science Centre..

Digital Object Identifier (DOI) is a centrally registered unique ID assigned by a specific authority (publisher, repository host, etc.). Issue of a new DOI requires submission of a request form to the relevant authority. There are costs associated with issuing a DOI.

### 4.2.2 PITHIA metadata: GUID but no DOI

Each metadata document registered with PITHIA e-Science Centre will be provided by GUID in order to satisfy FAIRness requirements. The GUID issue is the responsibility of the e-Science Centre. PITHIA metadata documents are not registered with DOI authorities..

### 4.2.3 PITHIA data subsets: DOI

A provision is made to attach previously issued DOIs to specific subsets of data collections listed in Catalogues. A strong consideration is given to registering EGI as a DOI authority to help data providers with the process of registering DOI to important subsets of their data.

# 5. Individual registration steps

## STEP ONE: ORGANISATION

Use the provided template to edit the name, description, address, and phone number.

## STEP TWO: INDIVIDUAL

Use the provided template to edit the person's name and contact information. Ensure that the organisation xlink points to a valid Organisation document.

## STEP THREE: PROJECT

*Project* is a research program funded and hosted by *Organisation*. By design, Data Collections are linked to projects (related to *scientific investigations*), not organisations (related to *administrative operations*).

Use the provided template to edit the project name, abstract, related parties (personnel), status, and keywords. Provide references to relevant academic publications on the research topic of the project (using the <Citation> element). Ensure that Related Party elements point to the valid Organisation and Individual documents already registered with PITHIA-NRF. The Status element has to use a valid term in the PITHIA ontology dictionary; please refer to the online ontology browser to look up available terms.

## STEP 4-8: REGISTRATION OF MEASUREMENT DATA

For those *Data Collections* that hold measurement data (as opposed to numerical modelling data), Steps 4-8 describe the process of acquiring the measurements. If a *Data Collection* refers to numerical modelling data that do not involve measurements, steps 4-8 are bypassed.

### BACKGROUND: PLATFORM + INSTRUMENT

Definitions

- *Instrument* is a human-made device that interacts with the *Feature of Interest* (Mother Nature) in order to estimate its *Observed Property* values.
- *Platform* is an identifiable object that brings the instruments to the appropriate environment (e.g., aircraft, ground station, satellite).

In most cases, the *Platform* simply refers to the observatory that hosts the instrument whose measurements are registered. In case the Platform is a moving object such as a satellite or a ship, the *Operation* document (Step 5) describes how the platform location can be determined as a function of time.

### One Platform, Many Instruments

A single Platform can accommodate multiple sensor instruments of different types. Only one Platform document requires registration. In the <description> element of the Platform, feel free to describe instruments that operate on the platform (all of them or only the key ones). However, the Platform itself (and its Local ID in particular) should not be specific to any specific Instrument or its attributes (for example, a standard identifier issued to one of the instruments). Use <standardIdentifier> element to list all specific IDs that the Platform was assigned by different authorities. Multiple standard IDs issued by the same authority are allowed. For example,

```
<standardIdentifier authority="URSI">RO041</standardIdentifier>
<standardIdentifier authority="URSI">RM041</standardIdentifier>
```

## Acquisition versus Computation

### Definition

*Process* is an ISO term for specifying the technical details of observing and modelling the *Feature of Interest*. In the PITHIA metadata model, we use *CompositeProcess* documents to describe the process in terms of its *Acquisition* and *Computation* components:

- **Acquisition (Step 8)** involves an Instrument: a device that interacts with the *Feature of Interest* to obtain estimates of the observed properties;
- **Computation (Step 10)** does not interact with the *Feature of Interest* but instead involves only numerical computation with particular input and output.

## Linking Instruments and Platforms: use Acquisition

To define which instruments operate on which platforms, use the *Acquisition* document (Step 8). *Acquisition* provides <capabilityLinks> element with a list of Platform-Instrument pairs. Each <capabilityLink> inside <capabilityLinks> defines one pair of *Platform* and *Instrument*. The <capabilityLink> may also include <timeSpan> to define when the *Instrument* was installed on the *Platform* and whether it is still in operation.

### Sensor Networks and Spacecraft Constellations

Special arrangements are made to support platform networks or constellations that manage multiple coordinated platforms with centralised data management.

When a *Data Collection* registers data from such networks/constellations:

- Each participating Platform is registered individually
- The *Acquisition* document is issued for the **network**
    - The network *Acquisition* XML includes multiple <capabilityLink> items inside the <capabilityLinks> element for each participating Platform and matching Instrument.
    - Use <timeSpan> to capture history of the instrument operation, with upgrades and replacements

Both *Acquisition* and *Computation* generate *Observed Properties* (ISO term) – physical characteristics of the Phenomenon whose values are estimated and presented in Data Collections.

## STEP 4: PLATFORM

Individual *Platform* registration: edit name, description, type, documentation Citations, and related parties. Use your organisation namespace.

Registration of composite *Platforms* (sensor networks or spacecraft constellations): ensure all <childPlatform> links point to the right documents in their correct namespaces. Use *pithia* namespace for the Composite Platforms. Because composite *Platform* definitions apply to coordinated measurements by similar instrumentation, their Local ID should include a short descriptor of the instrument, for example, Platform_Ionosondes_GIRO.

## STEP FIVE: OPERATION

Missions of a specific period of performance or with a varying platform location may be provided with this *Operation* document to specify begin and end times of the mission or various Citations to point to the orbital/location data. Please ensure that *Operation* is linked to a valid *Platform* document.

## STEP SIX: INSTRUMENT

### Instrument Types

Specification of the *Instrument:* ensure that the category of the instrument is available in the "InstrumentType" vocabulary of PITHIA ontology specifications. If a new category is needed, please contact Ivan.Galkin@borealis-designs.org.

### Instrument Operational Modes

Multi-modal instrumentation becomes more common; if differences in the operational modes are significant enough to result in different *Data Collections*, please define all individual modes separately using the <operationalMode> element in *Instrument*. For example, separate definitions can be provided for active vs passive, wide-angle vs beam-forming, steering vs static, high vs low resolution, high vs low frequency band, fixed-frequency vs stepping frequency, etc., etc.  This registration will help identify how the measurement data are generated.

If the Instrument always operates in one standard mode, **define that one mode** in <operationalMode>. This definition is now required in order to link Platform and Instrument: the <capabilityLink> in the *Acquisition* document uses <instrumentModePair>. So, *Platform* and *Instrument* are linked using the appropriate operational mode of the *Instrument*.

## DEFINING PROCESS CAPABILITIES

Multiple observed properties can be sampled/generated during the acquisition/computation process. The number of observed properties may vary from measurement to measurement or from model run to model run. Here it is necessary to compile a list of each possible property, regardless of whether it is observed or modeled every time.

Use the <processCapability> element that contains

- <observedProperty> (a link to the PITHIA ontology vocabulary)
- <dimensionalityInstance> and <dimensionalityTimeline> (2 links to the ontology)
- <cadence> to define time resolution of sampling
- <vectorRepresentation> and <crs> (2 links to the ontology) – for the vector properties
- <units> (a link to the ontology)

### Ontology of Observed Properties

Please browse the online space physics ontology system for definitions of various Observed Properties. Should your instrument or numerical model generate an observed property that is not listed in the ontology, please reach out to the ontology definitions team via the PITHIA-NRF email distribution list.

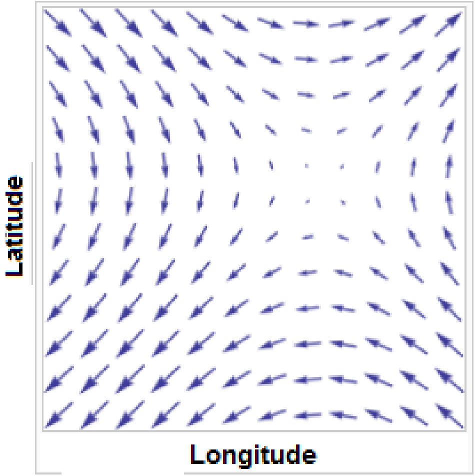
### Ontology of Dimensionalities

Dimensionality describes the *data domain* of the observed property that is spanned by its independent variables. Two dimensionalities are provided for the data domain (1) at any instance of time and (2) when viewed in time progression.

For example, the data domain of an outdoor thermometer has *0D.Point* instance dimensionality and *1D.TimeSeries.Single* timeline dimensionality. For the ionogram-derived Electron Density Profiles (EDP), the instance dimensionality is *1D.Profile.Altitude.Vertical* and the timeline dimensionality is *2D.Image.Profilogram*.

### Ontology of Vector Representations

Vector Representation defines whether a vector Observed Property is represented partially or completely by the observation process. It commonly includes the system of coordinates selected for the resulting data resource. For example, the 3D vector wind velocity may be represented as a 2D vector field at a particular altitude (see Figure 2).



Observed Property: **Neutral Wind Velocity**
Phenomenon: **Particle.Neutral.Air**
Measurand: **Velocity.Flow**
Feature of Interest: **Earth.Atmosphere.Troposphere**

Parameter: **Position.Altitude**

DimensionalityInstance: **2D.Map**
DimensionalityTimeline: **2D.Animation**
Vector Representation: **Projection.Horizontal**
crs = **GEOSpherical**

**Figure 2.** Example of a partial representation of a vector *Observed Property* “neutral wind velocity”. The sensor is capable of recording the horizontal projection of the true vector.

### Data Levels

A vocabulary of standard Data Level definitions is provided in PITHIA ontology to characterise the output observed properties of each *Acquisition* and *Computation*:

| Data Level | Description |
|---|---|
| L0 | Raw samples acquired by sensor instrumentation in their entirety and full resolution, presented in the instrument-specific units. |
| L1 | Instrument-specific observed properties derived from Level 0 data by a sequence of computational processes to calibrate, clean, reduce volume, enhance quality, or convert the presentation to physical units.<br>Typical example processes for transitioning to Level 1 are: enhancement of signal-to-noise ratio (pulse compression, excision of interference and outliers, look integration, synthetic aperture beamforming), cross-channel equalization (for the multi-channel or multi-element observations), compensation of distortions such as aberration or defocusing, derivation of secondary physical quantities of the probing signal (e.g., angle of arrival from multi-channel reception), spectral analysis of the observed signal, data reduction to retain only important quantities (e.g., resolution reduction or data compression, thresholding to detect principal signal components). |
| L2 | Geophysical properties of the primary Feature of Interest derived from Level 0/1 instrument data by a sequence of computation processes. The spatial extent and time resolution of Level 2 data are usually constrained to those of the source instrument data. Transition to Level 2 commonly includes geolocation of the Feature of Interest and may rely on underlying model assumptions about the feature (such as the slant-to-vertical total electron density transformation or tomography computations). Most commonly, it is the Level 2 resource that the instrument teams release to the science community at large for reasoning about the Feature of Interest. |
| L3 | Same as Level 2, but additionally processed to map observed properties to a uniform spatial and temporal grid. Transition to Level 3 is commonly based on underlying model assumptions about Feature of Interest's spatial/temporal extent over areas/times outside the sensor coverage. Typical Level 3 processes are spatial interpolation and extrapolation onto a 2D surface grid and compensation of the short-term measurement latency in real-time weather applications. |
| L4 | Model computations generated with or without analysis of the lower-level data. May include additional resources external to the data collection (such as characterizations of the geospace activity or additional observations in the multi-instrument experiments). |

### Data Quality Flag

PITHIA-NRF facility nodes are responsible for the quality of the digital research data before publishing them to the e-Science Centre.

The quality of digital research data is determined by:

- Their intrinsic scientific quality;
- The quality of metadata that describe the research data;
- The quality of data resources.

Data providers are responsible for the description of the intrinsic scientific quality of their collections by reporting the Data Quality Flag (DQF).

**Data quality flag (DQF)** describes measures taken to *clean* and *validate* the data, as well as characterise the residual data noise. It is distinct from another data qualifier used in the specification of Acquisition and Computation called *Data Level* (see above) characterises the amount of data processing applied to the measurements to obtain higher level data products (in terms of the observed properties). Commonly, *Data Level* 1 refers to observed properties of the instrument probing signal while *Data Level* 2 corresponds to the derived geophysical properties of the Feature of Interest, etc.

Data Quality Flag accepts 5 different values:

| **DQF** | **Name** | **Description** |
|---|---|---|
| 0 | RAW | Raw output of Acquisition or model Calculation with no regard to its quality |
| 1 | CLEAN | Automatic data conditioning is applied |
| 2 | EVALUATED | Provided with automatically computed confidence and uncertainty metrics |
| 3 | VERIFIED-CLEAN | Post-processed manually to ensure removal of data noise |
| 4 | VALIDATED | Validated against independent measurements or models |

DQF gradations are not mutually exclusive; Data Collection, Acquisition, Computation, and Catalogue Subset documents may assign multiple data quality flags for the same data product. For example,

- Volcano Eruption Catalogue: a Subset of the ionogram-derived measurements is registered in the Catalogue that was VERIFIED-CLEAN by manual editing of ionograms and VALIDATED against other measurements of the same study.
- Space Weather Monitoring: assimilative data models that assess the quality of their input data may require data products that are CLEAN and EVALUATED in order to use them in a Kalman-filter assimilation procedure.

- Manually Edited Ionograms: although manual editing ensures that ionogram scaling is correct (VERIFIED-CLEAN), the profile inversion Computation may use an ensemble of software algorithms to evaluate the uncertainty of true height values (EVALUATED).

**Data Quality Flag 0 (RAW)**

When no consideration is made to the evaluation of the data product quality, DQF is zero.

**Data Quality Flag 1 (CLEAN)**

The CLEAN flag is assigned to report the *data conditioning* capability of the Computation that applies automatic measures to exclude data noise.

Example data conditioning algorithms are

- Detection/removal of data outliers,
- Filtering to exclude data jitter,
- Content sanity checks against physical criteria (e.g., exclusion of negative density or altitude values, or other comparisons against established threshold values).

Ensuring that data collection is CLEAN is important in space weather monitoring scenarios where data noise may disrupt operations of an assimilative forecasting model.

**Data Quality Flag 2 (EVALUATED)**

The EVALUATED flag is assigned to those Data Collections and their Computation steps that provide confidence and uncertainty metrics evaluated automatically. The metrics may be related to

- observed data precision (as expressed by the standard deviation of repeated measurements),
- previous statistical error analysis,
- inter-comparisons of ensemble computations running in parallel.

[Note: use of the "error bar" language is discouraged for the automatically calculated metrics because the "error" can be stated only when the true value is known. "Uncertainty bounds" language is recommended.]

EVALUATED data are required in assimilative models based on the Kalman filter.

**Data Quality Flag 3 (VERIFIED-CLEAN)**

The VERIFIED-CLEAN flag is given to data collections and their computation steps that involve human experts to ensure the removal of data noise. A typical example would be manual scaling of ionograms to remove the artefacts of autoscaling.

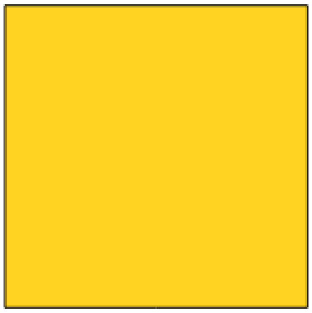
**Data Quality Flag 4 (VALIDATED)**

Scientific data of the best quality are additionally validated by comparisons against independent measurements or models. Such Data Collections and Catalogue Subsets are part of the consortium of models and measurements used collectively for analysis of the Feature of Interest and confirmed to agree in their descriptions. Typical examples arise in event studies involving multiple instruments and models, or specific “CalVal” campaigns to validate novel instrumentation. For example, joint analysis of the peak density height measurements in the ionosphere as observed by ionosonde, incoherent radar, and radio occultation network can result in a Catalogue Entry Subset with the assigned VALIDATED flag.

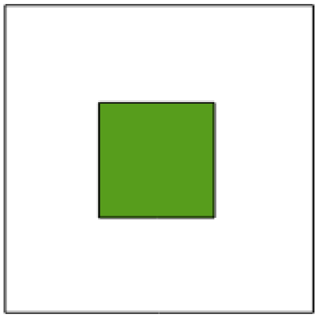
### Describing Data Collections of Mixed Quality

It is common for data collections to hold scientific information of varying quality, ranging from DQF-0 to DQF-4 in the flag classification. The current recommendation to PITHIA-NRF data owners is to prepare multiple registrations of the same data resource in 2+ DataCollection documents, sorted by their data quality.

Suppose we are registering a *Data Collection* that starts with automatic *Acquisition* and *Computation* processes. The DQF of such a collection (shown in yellow) could be RAW or CLEAN/EVALUATED:



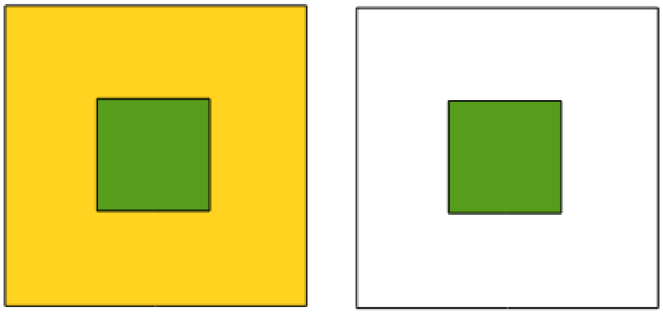
Additional processing was applied to a subset of this collection; so the smaller subset of the original data attains a higher level, VERIFIED or VALIDATED (shown in green):



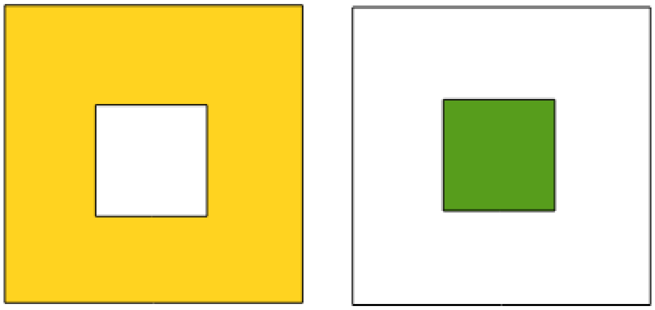
Note that the values of the data might be the same, but e.g. in the case of ionograms the value of a scaled characteristic according to a human scaler might be different than the original raw value. Thus, yellow and green values at the same place in the square might be the same or different.

So, two data collections are required. There are, however, several ways these two could be conceived:

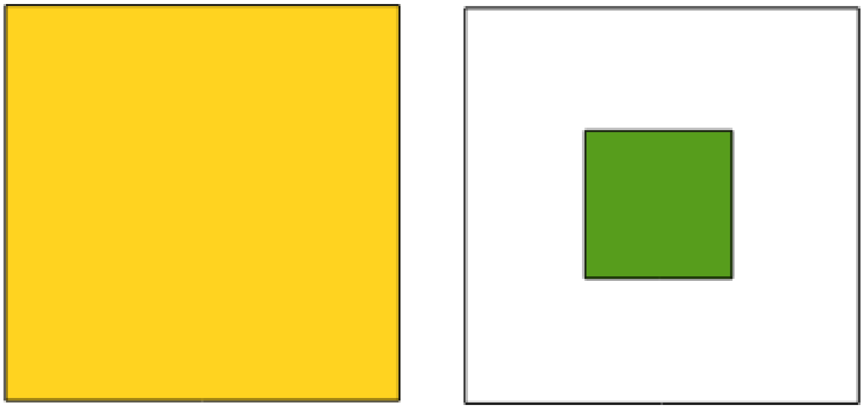
**Option 1**: Allow non-uniform quality flags, i.e., best available with a lower bound (each collection gets labelled with the lowest flag in it, even if higher quality data is also part of it):



**Option 2**: Each collection only has data of uniform quality, users can combine different collections themselves to get the entire time series:



**Option 3**: Different levels of quality are available side by side for the same parameter, same time, same instrument. Users can implement the “best available” collection themselves by merging, but also can conduct statistical studies of the automatic data quality in comparison to the verified/validated reference:



The current recommendation is to use **Option 3**.

### Metadata Quality

Data Collections can also be assigned multiple metadata quality flags:

| MQF | Name | Description |
|---|---|---|
| 1 | USAGE | MQ1D = Descriptive (i.e., date); MQ1S = Structural (10 steps); MQ1A = Administrative (identifiers) |
| 2 | SCOPE | Agreeable content and vocabularies |
| 3 | PROVENANCE | Provided with a track of acquisitions/computations that resulted in the Data Collection (repeatable) |
| 4 | PERSISTENCE | Provided with unique global identifiers and pointing to data that are also provided with such identifiers |
| 5 | AGGREGATIONS | Capable of grouping different resources into sets by event or investigation |
| 6 | STANDARTISATION | Provided with the data model and compliant with its schema |
| 7 | INTEROPERABILITY | Open for harvesting in interdisciplinary applications |
| 8 | QUALITY | High quality |
| 9 | EARLINESS | Built automatically as soon as data are available |

## STEP SEVEN: ACQUISITION CAPABILITIES

The primary content of *AcquisitionCapabilities* is the list of Process Capabilities related to the Instrument operation.

### Instrument-related Computations

Many instruments engage a simple acquisition procedure followed by a series of intricate computations that then comprise the Process of generating Data Collection. In particular, optical instruments may acquire an image in a straightforward manner but then employ a series of secondary analyses to extract and interpret visible signatures in the image that are related to the feature of interest.

To simplify the registration process and keep the *Process* document (Step 11) simple, we attribute all computational steps that deal with the instrument itself to the *AcquisitionCapabilities*. In this case, such instrument-bound process steps do not have to be registered as individual Computation documents. The only requirement is that these steps do not generate the properties of the feature of interest but rather manipulate sampled data to improve or characterise the sensor instrument's capability.

For example, computation and application of the cross-channel phase differences, elimination of the measurement biases, computation of the probing signal spectra, derivation of the angle of signal arrival from multi-channel reception data, or protection measures against interference – are all considered parts of the *Acquisition*: they remain within the sensor instrument domain, even though certain interpretations of thus acquired sensor values may be involved.

For example, a sounder that detects an echo of its probing signal and computes its time of flight does not require formal definitions of the involved computations; its Acquisition document simply states “signal time of flight” as the observed property. However, an ionosonde that converts its transmission frequency to the ionospheric plasma density at the reflection point must define the Computation procedure because it results in a property of the *Feature of Interest*.

### Data Level

The Acquisition documents are expected to report Data Level 1. Level-0 data (raw samples) are usually considered to be a prerogative of the instrument team with sufficient engineering expertise; their registration is not necessary.

### Data Level

### Use of Namespace in AcquisitionCapabilities

Use your organisation’s namespace for unique instrumentation developed by your research and engineering team. More commonly used and generally accepted *AcquisitionCapabilities* may be placed into *pithia* namespace so that other *Data Collections* can reuse them.

## STEP EIGHT: ACQUISITION

The primary objective of the *Acquisition* document is to **link Platform(s) and Instrument(s)** with their acquisition capabilities. It holds one <capabilityLinks> element with 1 or more <capabilityLink> sub-elements inside.

### One Platform, One Instrument, Single Installation

In case of one unique *Instrument* continuously operating on one *Platform*, just include one <capabilityLink>, for example:

```
<!-- connect platform and instrument acquisition capabilities -->
<capabilityLinks>
  <capabilityLink>
    <platform xlink:href="https://metadata.pithia.eu/resources/2.2/platform/eiscat/Platform_EISCAT_Svalbard"/>
    <acquisitionCapabilities xlink:href="https://metadata.pithia.eu/resources/2.2/acquisitionCapabilities/eiscat/AcquisitionCapabilities_EISCAT_ESR"/>
    <timeSpan>
      <gml:beginPosition>1996-08-22</gml:beginPosition>
      <gml:endPosition indeterminatePosition="after"/>
    </timeSpan>
  </capabilityLink>
</capabilityLinks>
```

### One Platform, One Instrument, History of upgrades

In case of one *Instrument* continuously operating on one *Platform*, but with history of science and engineering developments that resulted in significant hardware upgrades, use several <capabilityLink> to capture the history, for example:

```
<!-- connect platform and instrument acquisition capabilities -->
<capabilityLinks>
  <capabilityLink>
    <platform xlink:href="https://metadata.pithia.eu/resources/2.2/platform/uit/Platform_EISCAT_Tromso"/>
    <acquisitionCapabilities xlink:href="https://metadata.pithia.eu/resources/2.2/acquisitionCapabilities/eiscat/AcquisitionCapabilities_EISCAT_UHF_Old"/>
    <timeSpan>
      <gml:beginPosition>1981-01-01</gml:beginPosition>
      <gml:endPosition>2000-01-01</gml:endPosition>
    </timeSpan>
  </capabilityLink>
  <capabilityLink>
    <platform xlink:href="https://metadata.pithia.eu/resources/2.2/platform/uit/Platform_EISCAT_Tromso"/>
    <acquisitionCapabilities xlink:href="https://metadata.pithia.eu/resources/2.2/acquisitionCapabilities/eiscat/AcquisitionCapabilities_EISCAT_UHF"/>
    <timeSpan>
      <gml:beginPosition>2000-01-01</gml:beginPosition>
      <gml:endPosition indeterminatePosition="after"/>
    </timeSpan>
  </capabilityLink>
</capabilityLinks>
```

[Side comment]: similarly, if a *Computation* component of the *CompositeProcess* was updated, history of modifications is captured inside the *Computation* document.

In those cases when the upgrades were so significant that two different *DataCollections* will have to be registered with PITHIA, make two different Acquisition documents.

### Many Platforms, One Instrument

If your instruments are all of the same model and *AcquisitionCapabilities* and it is only that they are placed on multiple platforms (e.g., Cluster 1,2,3,4), the easiest way would be to build one *Instrument*, one *AcquisitionCapabilities* for the instrument, and many *Platform* descriptions first. Then, write one *Acquisition* document whose <capabilityLinks> element has entries for each platform.

### Many Platforms, Several Instrument models, History of upgrades

When your *DataCollection* is made using a network of instruments on different platforms, use <capabilityLinks> to link platforms and instruments accordingly.

NOTE: <capabilityLink> allows optional <standardIdentifier> sub-element for those cases when a *Platform* was assigned different identifiers for different instrument models. The *Platform* itself does not carry the <timeSpan> descriptor; it is the *Acquisition* (and also *Computation*) that allows specification which *Platform* identifier to use inside the <capabilityLink>.

## STEP NINE: COMPUTATION CAPABILITIES

It is critical for *DataCollection* registrations to define every "observed property" item that will be searchable *by content* using the online PITHIA user interface. For the measurement-based data collections, searching for the instrument "Level 1" observed properties is less important than the derived ones that are attributed to the feature of interest. The derived observed properties are computed without additional interaction of the *Instrument* with the *Feature of Interest*.

All derived *ObservedProperty* items are enlisted in *ComputationCapabilities* documents using <processCapability> elements.

### Use of Namespace in ComputationCapabilities documents

Use your organisation's namespace for those *ComputationCapabilities* definitions that are unique to your *DataCollection*. More commonly used and generally accepted computations shall be placed into *pithia* namespace so that other *DataCollections* can refer to them.

## STEP TEN: COMPUTATION

*Computation* document is very similar to *Acquisition* in that it uses the same <capabilityLinks> element to associate *Platform* and *ComputationCapabilities*.

For the numerical modeling *DataCollections*, the Platform is not included in the <capabilityLink>.

## STEP ELEVEN: PROCESS

All *Acquisition* and *Computation* documents prepared in Steps 8 and 10 of the registration are listed in the Process document.

## STEP TWELVE: DATA COLLECTION

*Data Collection* documents include four important components:

1. **URLs to data**
   a. Use <pithia:CollectionResults> to list all URLs to data services
   b. Each URL is placed inside one <source> element as a single <OnlineResource> element
2. **Feature of Interest**
   a. Use PITHIA ontology to enlist each item as a <namedRegion>
   b. This description will be instrumental for those *search-by-content* queries that refer to a specific region of space
3. **Data access licence**
   a. The licence definition must be added to the PITHIA ontology vocabulary of licences (this is not a free-text item!)
   b. Consider using standard licence definitions, e.g., Creative Commons
4. **Input Parameters**

  a. *Parameter* (*ISO vocabulary*) is an external attribute of a *Data Collection* used to control its *Process*, either manually by project investigators or automatically by involving independent data resources. Most commonly, activity indices (e.g., sunspot number, $F_{10.7}$, $K$p, $D$st, etc.) are input parameters impacting the *Computation* component of the *Process*. Sometimes, *Data Collection* may include an instrument setting or a software configuration as its input parameter.
  b. *Parameter* (input) shall not be confused with *Observed Property* (output).
  c. Define input parameters if their management is required for the *Process* to run correctly. For example, executable models that run automatically at the PITHIA e-Science Centre may require up-to-date knowledge of the activity indices or other computations.

Additionally, make sure to include proper links to

- Process (Step 11) using <procedure> element
- Research Project at your organization (Step 3) using <project> element, and
- Organisation and Individuals (Steps 1 and 2) responsible for the collection using <relatedParty> elements

## 6. Catalogues

Catalogues are listings of events or investigations assembled to aid users in locating data of interest. Each Catalogue entry has distinct begin and end times and an optional DOI to the Data Subset in the permanent storage.

*Catalogue* is not part of the standard *Data Collection* registration. The catalogues are managed separately, based on the Data Collections in PITHIA eScience Centre. There are three types of components of each *Catalogue:*

1. Top-level Catalogue Category (e.g., Catalog_VolcanoEruption)
    a. Catalogue category is ontology-controlled
2. Individual entries of the Catalogue describing each specific event or investigation
    a. Each entry has a description and PhenomenonTime
    b. Each entry is linked to the Catalogue Category document
3. *Data Subset* items
    a. Each *Data Subset* refers to a Data Collection
    b. Each *Data Subset* includes <resultTime> to define intervals of time that the subset spans
    c. Optionally, the data provider may specify a DOI for the persistent storage of the *Data Subset*

## References

Belehaki, A., James, S., Hapgood, M., Ventouras, S., Galkin, I., Lembesis, A., Tsagouri, I., Charisi, A., Spogli, L., Berdermann, J. and Häggström, I., (2016). The ESPAS e-infrastructure: access to data from near-Earth space. *Advances in Space Research*, *58*(7), pp.1177-1200. https://doi.org/10.1016/j.asr.2016.06.014