

## Deliverable D5.4: Report on data interoperability

**Markus Fiebig, Ewan O'Connor, Simo Tukiainen, Giuseppe D'Amico, Richard Olav Rud,  
Sabine Philippin, Cathrine Lund Myhre**

<b>Work package no</b>	<b>WP5</b>
<b>Deliverable no.</b>	<b>D5.4</b>
<b>Lead beneficiary</b>	<b>NILU</b>
<b>Deliverable type</b>	<input checked="" type="checkbox"/> R (Document, report) <input type="checkbox"/> DEC (Websites, patent fillings, videos, etc.) <input type="checkbox"/> OTHER: please specify .....
<b>Dissemination level</b>	<input checked="" type="checkbox"/> PU (public) <input type="checkbox"/> CO (confidential, only for members of the Consortium, incl Commission)
<b>Estimated delivery date</b>	<b>Month 36</b>
<b>Actual delivery date</b>	<b>30/05/2018</b>
<b>Version</b>	
<b>Comments</b>	

## 1. Scope

The numerous measurement methodologies applied within ACTRIS result in a considerable diversity of the data collected. In accordance with these requirements, the ACTRIS Data Centre consists of three topical data repositories archiving the measurement data, which are all linked through the ACTRIS data portal to provide a single access point to all data. This deliverable describes the efforts towards data interoperability between ACTRIS and the European environmental Research Infrastructures (RIs) in the atmospheric domain. These RIs, the *In-service Aircraft for a Global Observing System* (IAGOS) and the *Integrated Carbon Observation System* (ICOS), the *European Incoherent Scatter Scientific Association* (EISCAT-3D), and the *Svalbard Integrated Observing System* (SIOS) besides ACTRIS, collaborate on data interoperability within the framework of the ENVRIplus project, an umbrella project covering all European environmental research infrastructures and focussing on common aspects of management between RIs. The deliverable covers the topic within the atmospheric domain.

Data interoperability covers at least the following aspects:

- **machine-to-machine interface for metadata**, allowing remote access to metadata, e.g. by portals or virtual research environments (VREs) for harvesting and searching the data of an archive.
- **machine-to-machine interface for data**, allowing to retrieve datasets from the archive found by searching the metadata.
- **standardisation of metadata format and specifications of format use (profile)**, needed to allow machine controlled interpretation of metadata. The metadata standard may go beyond basic discovery metadata to include use metadata, e.g. uncertainties and quality assurance (QA) information.
- **standardised format and format profile for data** in the machine-to-machine interface, needed to interpret accessed data automatically.
- **standardised vocabulary used in metadata**, another requirement for automatic interpretation, alternatively mapping between vocabularies
- **standardised ontology where possible**, alternatively mapping between ontologies, where an ontology is the hierarchy of terms describing data and infrastructure
- **common or compatible access conditions to data**, i.e. compatible use of data policies and data licenses.
- **common options for user authentication** to facilitate single-sign-on to access restricted data within and across infrastructures.
- **data citation service**, i.e. the option of quoting data by a unique and globally resolvable identifier, thereby giving credit to entities involved in the data production chain.
- **facilitate data discovery** by registration in relevant discovery portals, e.g. WMO Information System (WIS), WMO International Global Observing System (WIGOS), Global Earth Observation System of Systems (GEOSS).

This ACTRIS-2 deliverable documents work of the ACTRIS-2 and ENVRIplus projects when the collaboration of the ACTRIS, IAGOS, and ICOS RIs was established. This work focussed on essential aspects of those mentioned above needed to obtain first valuable results in inter-RI data interoperability, which are:

- metadata machine-to-machine interfaces
- data machine-to-machine interfaces
- metadata standard and profile
- vocabulary
- access conditions
- authentication
- registration in portals.

Since then, requirements for inter-operability have been formalised, e.g. the FAIR criteria of Findability, Accessibility, Interoperability, and Re-usability, advocated by the FORCE 11 group and adopted by the European Commission, e.g. in the declaration of the European Open Science Cloud (EOSC). These will be the guiding principles of the continuation and extension of the work presented here.

## 2. Steps Towards Interoperability Between Atmospheric RIs

### 2.1. Machine-to-machine interfaces for metadata

All of the RIs involved here work, fully or partially, in the atmospheric domain. The atmospheric RIs decided to have the common goal of implementing metadata servers conforming to a common standard. Concerning interfaces and standards for metadata exchange, the standards set by the World Meteorological Organisation are therefore relevant. For metadata exchange, the WMO Information System (WIS) uses the OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) standard as transport layer. The metadata themselves are encoded as XML files following the ISO 19115 standard and the WMO profile for that standard.

Summary of the current implementation status:

- ACTRIS:
  - surface in situ: standard conforming interface implemented:
    - <http://ebasoi.nilu.no/oai>, e.g.
    - <http://ebasoi.nilu.no/oai?verb=ListSets>
  - aerosol remote sensing:
    - THREDDS catalogue service available at:
      - <https://login.earlinet.org:8443/thredds/catalog.html>
      - ISO 19115 protocol:
        - [https://login.earlinet.org:8443/thredds/iso/\\*](https://login.earlinet.org:8443/thredds/iso/*), e.g.
        - <https://login.earlinet.org:8443/thredds/iso/earlinedbscan/ab/f2002/ab0201041108.b355.html>
        - NetCDF Markup Language (NcML):
          - [https://login.earlinet.org:8443/thredds/ncml/\\*](https://login.earlinet.org:8443/thredds/ncml/*), e.g.
          - <https://login.earlinet.org:8443/thredds/ncml/earlinedbscan/ab/f2002/ab0201041108.b355.html>
      - OAI-PMH service currently being implemented.
    - cloud remote sensing: standard conforming interface to be implemented, customised interface based on RESTful API available
  - IAGOS: interface serving metadata conforming to ISO 19115 to be implemented, in addition to existing OGC-CSW (Open Geospatial Consortium - Catalog Service for the Web) interface, e.g.
    - <http://sedoo.sedoo.fr/geosource/srv/fre/csw-aeris-envriplus?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities>
  - ICOS: THREDDS catalogue service available at.
    - <http://thredds.icos-cp.eu/thredds/catalog.html>

ACTRIS currently uses both, standard conforming and customised, auto-updated text file based metadata interfaces. The ACTRIS Data Centre (DC) roadmap plans a transition to metadata interfaces complying with WIS or WIGOS standards for both, internal and external metadata exchange.

### 2.2. Machine-to-machine interfaces for data

In the context of ENVRIplus, ACTRIS-2 seeks to agree with the atmospheric RIs to serve their data by standard conforming interfaces, but leaving the type of the interface open. Data describing properties or constitution of the atmosphere can have very different morphology. Data from Earth surface observatories are usually time series, observations by column remote sensing are 2D time series, model output as 3D time series, and aircraft-based observations as time series along trajectories. Consequently,

all these data types may correspond to different machine interfaces that best meet their specific requirements. Options include a number of commonly used data machine-to-machine servers, e.g. OPeNDAP (Open-source Project for a Network Data Access Protocol) and OGC-WCS (Open Geospatial Consortium – Web Coverage Service). The OPeNDAP interface is recommended since it allows streaming of data, which improves performance of web-based interoperability applications. A commonly used server software package that offers these services, and more, based on a file archive in NetCDF (Network Common Data Form) format, is the THREDDS (Thematic Real-time Environmental Distributed Data Services) server package.

The implementation status is summarised as follows:

- ACTRIS:
  - surface in situ (OPeNDAP service):  
<http://dev-ebas-pydap.nilu.no/>
  - aerosol remote sensing (OPeNDAP service):  
<https://login.earlinet.org:8443/thredds/dodsC/>\*, e.g.  
<https://login.earlinet.org:8443/thredds/dodsC/earlinedbscan/ab/f2002/ab0201041108.b355.html>
  - cloud remote sensing: OPeNDAP service to be implemented, currently using custom RESTful API
- IAGOS: OPeNDAP, OGC-WCS, and OGC-WMS services to be implemented via THREDDS data server.
- ICOS: data machine-to-machine interfaces implemented by THREDDS server, e.g.  
<http://thredds.icos-cp.eu/thredds/dodsC/>

For the exchange of data between the three ACTRIS DC topical data repositories and the ACTRIS data portal, customised non-standard interfaces are currently in use for machine-to-machine data exchange. However, all ACTRIS DC topical data repositories are in the process of setting up interfaces based on the THREDDS data server package, or have already done so. The corresponding data services will be used to serve ACTRIS data both inside and outside the RI.

### 2.3. Standard and profile for metadata

For data in the atmospheric domain, WMO has set 2 standards:

1. **WMO Information System (WIS) standard:** The WIS standard is meant for data discovery, and therefore contains only the corresponding metadata items. It uses the ISO19115 standard with pertaining XML implementation of ISO19139 and the WMO implementation profile. The metadata are served via an OAI-PMH service. In the WMO profile, variable names defined in the Climate and Forecast (CF) convention are specified. The WIS standard is mature and in operational use between WIS nodes.
2. **WMO Integrated Global Observing System (WIGOS) standard:** The WIGOS standard is intended to contain use metadata on data quality, uncertainties, and operating protocols in addition to discovery metadata, but is not backwards compatible with the WIS standard. The WIGOS standard is an emerging standard with its first implementation specified, but lacks vocabulary for many items. The standard is not yet in operational use. The transport layer for metadata in WIGOS is yet to be defined.

Due to maturity considerations, the atmospheric RIs decided to use the WIS metadata standard as goal for establishing basic interoperability. Other solutions can be used where more detailed metadata needs to be served. In ACTRIS, one node (surface in situ) has already implemented a metadata server following the WIS standard. The other nodes have chosen server technologies allowing the same solution.

## 2.4. Format for data

The probably most widely used data format for data exchange in the atmospheric domain is the NetCDF format, often with the profile of the CF convention (Climate and Forecast metadata convention, <http://cfconventions.org/>) in addition. This doesn't rule out that other data formats are used in earlier stages of the data production chain. The OPeNDAP data service specification is closely related to the NetCDF format in terms of its data model, but also allows data streaming. The atmospheric RIs have therefore concluded to apply the NetCDF format as default in their inter-operability applications. In ACTRIS, all DC nodes have either already implemented data services based on the NetCDF format, or are in the process of doing so.

## 2.5. Vocabulary

Vocabulary deployed in atmospheric domain metadata goes well beyond names for observed variables. For many metadata items, WMO has defined code lists, e.g. for station location and settings. For other types of metadata, e.g. instrument categories and types, especially at a more detailed level required for use metadata, such lists are lacking. Vocabulary may also include the hierarchy of terms naming metadata items, i.e. the ontology to describe the data. Also here, no common approach exists for the atmospheric domain. Implicitly by aiming to use the WIS standard, the atmospheric RIs decided to use CF convention vocabulary for variable names, and WMO vocabulary wherever required in the WIS standard. This is also the guideline within ACTRIS.

## 2.6. Access conditions

In specifying conditions for data access, data policies and licenses are or will be employed. Both serve a similar purpose, but may reside in different realms. A data policy often only makes an ethical statement, in contrast to a license, which sets legally binding conditions. In some cases, data policy and license can even be combined, e.g. when data is legally open without constraints and liability as in a Creative Commons 0 (CC0) license, but the data policy may include an ethical statement that the author should be acknowledged or quoted. By adopting the FAIR criteria for data handling for its EOSC, the European Commission has made a clear statement that free, unrestricted, and easy data access should be the goal for European RIs.

Among the atmospheric RIs, both data policies and licenses are used:

- **ACTRIS:** currently fair-use data policy based on the policy of the WMO Global Atmosphere Watch (GAW) Programme. Requires contact with data originator and negotiation of acknowledgement in case of significant use of data. Data is free for non-commercial use. Two DC nodes require authentication for data access.
- **IAGOS:** data policy base on WMO Resolution 40 for the exchange of meteorological and related data and products, observations are provided with free and unrestricted access for scientific (non-commercial) use. Authentication required
- **ICOS:** Creative Commons Attribution 4.0 (CC BY 4.0) license for data, author has to be attributed with defined attribution.

All of these access conditions are free in the sense that no fees are required for data access, even though some require authentication.

Due to the complexity of the topic and various needs, it is unlikely that data access conditions will ever be homogeneous across RIs. The fact that data access is free and somewhat open across the atmospheric RIs leaves room for collaboration and development of applications across RIs. In this case, the need for a license broker application arises. Such an application determines data policy / license of a data product based on the combination of data policies / licenses of the input products.

## 2.7. Authentication

Authentication is required to access all or parts of the data served by the atmospheric RIs. When using data across RIs, the use experience increases significantly when the user only has to authenticate once to obtain access to data across RIs (single-sign-on). Numerous authentication services provide this functionality, but only a few offer underlying protocols that feature business level security (SAML OAuth2 authentication protocols). In addition, the authentication providers need to be relevant for the atmospheric data domain. From the available authentication services, the atmospheric RIs decided to put priority on the EUDAT B2ACCESS and the Google Identity Platform services that meet these requirements.

The named authentication services will be further explored and implemented across the ACTRIS DC. For test purposes, the ACTRIS aerosol remote sensing topical repository has implemented a CAS (Central Authentication Server) server integrated with their THREDDS server.

## 2.8. Citation services

A data citation service identifies a dataset with a unique persistent identifier (PID) that ideally can be resolved globally. The purposes of such a service include:

- **Traceability:** Data used in an application can be uniquely and persistently identified, making the result transparent.
- **Attribution:** The entities involved in data generation can be credited for their work, given that sufficient metadata are provided with the PID.
- **Use accounting:** data use events can be counted, quantifying attribution and relevance of data

The most commonly used PID for data products are Digital Object Identifiers (DOIs). DOIs for data can be issued at different granularity. Common use cases for DOIs include:

1. Identify every single dataset contained in a data archive at fixed granularity (primary DOI). The granularity for issuing these primary DOIs needs to be determined by the primary archive considering aspects such as data reporting intervals, common use cases, and structure of archive.
2. Secondary DOIs, used to identify a user defined collection of data that may reside in several distributed primary archives.

Primary DOIs are a prerequisite for quantitative accounting of data use at the primary archive level. Therefore, the atmospheric RIs decided to implement these given the criteria above.

For ACTRIS, this service is on the roadmap for being implemented.

## 2.9. Data discovery

To maximise data use, an essential success criterion for a RI observing the atmosphere, the data has to be readily findable for an interested potential new user. To this end, a number of relevant data portals exist for the atmospheric domain:

- WMO Information System, data discovery system administered by the WMO, with synchronised central registration points across the globe, e.g. <https://gisc.dwd.de/wisportal/#>
- WMO Integrated Global Observing System (WIGOS), currently no portal yet.
- Global Earth Observation System of Systems (GEOSS), registry for all types of Earth observation data, European node: <http://www.geoportal.org/>. GEOSS harvests the WIS, registering data in WIS registers data in GEOSS.

The atmospheric RIs have set the goal of having their data holdings searchable in WIS and GEOSS. For ACTRIS, surface in situ data are already searchable in both, WIS and GEOSS.

### 3. Outlook: Atmospheric RIs in ENVRI-FAIR

Data interoperability between the atmospheric ENVRIplus RIs ACTRIS, IAGOS, and ICOS has reached a maturity level where a first inter-RI data access demonstrator, i.e. where a data portal of one or several of the participating RIs offers access to data of the other RIs. Such a demonstrator is currently being discussed in the ENVRIplus project.

In the remainder of the ACTRIS-2 project, focus will be placed on:

- Implementation of WIS compliant metadata servers for all ACTRIS DC topical repositories.
- Implementation of standard compliant data servers for all ACTRIS DC topical repositories.
- Implementation of primary DOIs throughout ACTRIS DC.

Even though the basic technical functionality for data interoperability between the atmospheric RIs has been established, significant work remains. This work, to be tackled e.g. in the ENVRI-FAIR project, includes:

- Implementing standard profiles for serving metadata
- Metadata vocabulary, standardised or mapped.
- Metadata ontologies, standardised or mapped.
- Implementing common authentication.
- Implementing data citation and data use accounting services
- Data license broker service

These and further aspects will be covered in the upcoming ENVRI-FAIR project, where the collaboration between the European geoscientific RIs, including the atmospheric RIs, will be continued.