**QUESTIONS**

Version 2: 18 August 2015

Cristina-Adriana Alexandru, Rosa Filgueira Vincente, Alex Vermeulen, Keith Jeffery, Thomas Loubrieu, Leonardo Candela, Paul Martin, Barbara Magagna, Yin Chen and Malcolm Atkinson

**ENVRIplus topics:** *(ENVRIplus development areas address the following topics):*

1. Identification and citation
2. Curation
3. Cataloguing
4. Processing
5. Provenance
6. Optimization
7. Community support

**A. Generic questions (for all topics)**

1. What is the basic purpose of your RI, technically speaking?

a. Could you describe one or several basic use-cases involving interaction with the RI that cover topics 1-7 ?

A user can go IAGOS website search for metadata then download data with PID which can be cited in publication.

b. Could you describe how data is acquired, curated and made available to users?

 

After an aircraft is landed, data is automatically transferred into the reception server as LOA data. LOA data will be validated. Validation process can be automatically or manually validated by PIs (within 3 days). Validated data (L1 data) and calibrated data (L2 data) will be stored in a centralised database, from where, end-users can access via a web-based data portal. Data levels are as follows:

* LOA: raw data
* LOB: automatic validation
* L1: validated data (by PI)
* L2: Calibrated data
* L3: process/analysed data
* L4: added value data, eg. correlation with satellite data

Raw data are in text file (ASCII), after validation, PIs will provide text files in NASA Ames or netCDF format. There is on-going work on metadata standard, the plan is to use ISO 19115 and align with INSPIRE.

Since 1994, already 2TB data in store, including >45,000 flights. ~ 500 users.

Currently IAGOS involve 6 institutions (from France and Germany), has 6 aircrafts, and aims to have 20 by 2025.

c. Could you describe the software and computational environments involved?

Self-developed software, such as FLEXPART, a tool for lagrangian transportation model. Database was postgresql but changed now to be MONGO DB. sometime use Matlab, but prefer to use open-source software.

d. What are the responsibilities of the users who are involved in this use case?

User will use data, they do not have particular responsibilities. When they register to use IAGOS data they need to agree to cite the project in any publications.

e. Do you have any use case involving interactions with other RIs (in the same or different domains?)

Share data with ACTRIS and ICOS, for example, in a map, users can see ACTRIS and ICOS platform.

2. What datasets are available for sharing with other RIs ? Under what conditions are they available?

All IAGOS datasets are able to be shared.

3. Apart from datasets, does your RI also bring to ENVRIplus and/or other RIs:

a. Software? In this case, is it open source?

b. Computing resources (for running datasets through your software or software on your datasets)?

c. Access to instrumentation/detectors or lab equipment? If so, what are the open-access conditions? Are there any bilateral agreements?

d. Users/expertise to provide advice on various topics?

e. Access to related scholarly publications?

f. Access to related grey literature (e.g. technical reports)?

No

4. What objectives would you like to achieve through participation to ENVRIplus?

* Improve data discovery
* Metadata standardisation
* Interoperability
* Citation and DOI management

5. What services do you expect ENVRIplus technology to provide?

* Citation
* Cataloguing
* Provenance

6. What plans does your RI already have for data, its management and exploitation?

a. Are you using any particular standard(s)?

 i. Strengths and weaknesses

Data is in NASA Ames and netCDF format, metadata use ISO 19115 standard

b. Are you using any particular software(s)?

 i. Strengths and weaknesses

FLEXPART--tool for lagrangian transportation model

c. Are you considering changing the current:

 i. standard(s)

 ii. software

 iii. working practices

 as part of a future plan?

 Please provide documentation/links for all the above which apply.

No

7. What part of your RI needs to be improved in order:

a. For the RI to achieve its operational goals?

b. For you to be able to do your work?

Manage data provenance and citation

8. Do topics [1-6] cross-link with your data management plan?

a. If so please provide the documentation/links

provenance, curation, citation

9. Does your RI have non-functional constraints for data handling and exploitation? For example:

a. Capital costs

b. Maintenance costs

c. Operational costs

d. Security

e. Privacy

f. Computational environment in which your software runs

g. Access for scrutiny and public review

If so please provide the documentation/links

Mainly the maintenance costs, supported by AERIS (the French Atmospheric Data Center) which is a collaboration among many French organisation (CNRS, CNES, Meteo-France, etc.)

10. Do you have an overall approach to security and access?

User access is under password control. However this needs to be improved, for example to use certificate-based approach.

11. Are your data, software and computational environment subject to an open-access policy?

yes, all IAGOS data open to access for research purposes. (but registration required)

12. What are the big open problems for your RI pertinent to handling and exploiting your data?

Although amount of IAGOS data is not large, to process data based on user given parameters can be complicated since it may involve many files open/read process. How to improve the performance of data processing/generation is at the moment a big challenge.

13. Are you interested in any particular topic [1-6] to discuss in more detail?

a. If so, would you like us to arrange a follow up interview with more detail questions about any particular topic to be discussed?

Data discovery, provenance, data identification & citation, data curation & cataloguing, data processing

14. Optional: If you are not the right person to reply to some questions from the above, please suggest the right person to contact from your RI for those questions.

Damien Boulanger is the Manager of the IAGOS Information system

**B. Specific questions per topic**

**1.** **Identification and Citation**

**a.** **Identification**

 i. What granularity do your RI’s data products have:

§ Content-wise (all parameters together, or separated e.g. by measurement category)?

all parameters together, download one or several parameters

§ Temporally (yearly, monthly, daily, or other)?

flight (at least daily)

§ Spatially (by measurement station, region, country or all together)?

global (flight trajectory)

 ii. How are the data products of your RI stored - as separate “static” files, in a database system, or a combination?

database system

 iii. How does your RI treat the “versioning” of data - are older datasets simply replaced by updates, or are several versions kept accessible in parallel? How do you identify different version of the same dataset ?

since IGAS we keep all versions for archive but only last available to users

 iv. Is it important to your data users that

§ Every digital data object is tagged with a unique & persistent digital identifier (PID)?

yes

§ The metadata for data files contains checksum information for the objects?

no

§ Metadata (including any documentation about the data object contents) is given its own persistent identifier?

file identifier (ISO 19115)

§ Metadata and data objects can be linked persistently by means of PIDs?

on going

 v. Is your RI currently using, or planning to use, a standardized system based on persistent digital identifiers (PIDs) for:

§ “Raw” sensor data? no

§ Physical samples? no

§ Data undergoing processing (QA/QC etc.)? no

§ Finalized “publishable” data? yes

 vi. Please indicate the kind of identifier system that are you using - e.g. Handle-based (EPIC or DOI), UUIDs or your own RI-specific system?

 DOI

 vii. If you are using Handle-based PIDs, are these handles pointing to “landing pages”? Are these pages maintained by your RI or an external organization (like the data centre used for archiving)?

by French Atmospheric Data Center AERIS

 viii. Are costs associated with PID allocation and maintenance (of landing pages etc.) specified in your RI’s operational cost budget?

In kind

**b.** **Citation**

 i. How does your “designated scientific community” (typical data users) primarily use your data products? As input for modelling, or for comparisons?

comparison with satellite or model for validation, atmospheric process studies and trends analysis

 ii. Do your primary user community traditionally refer to datasets they use in publications:

§ By providing information about producer, year, report number if available, title or short description in the running text (e.g. under Materials and Methods)? yes

§ By adding information about producer, year, report number if available, title or short description in the References section?

§ By DOIs, if available, in the References section? yes

§ By using other information?

§ By providing the data as supplementary information, either complete or via a link

 iii. Is it important to your data users to be able to refer to specific subsets of the data sets in their citation? Examples:

§ Date and time intervals

§ Geographic selection

§ Specific parameters or observables

yes to all

 iv. Is it important to be able to refer to many separate datasets in a collective way, e.g. having a collection of “all data” from your RI represented by one single DOI?

yes

 v. What strategy does your RI have for collecting information about the usage of your data products?

§ Downloads/access requests yes

§ Visualization at your own data portal yes

§ Visualization at other data portals no

§ References in scientific literature yes

§ References in non-scientific literature no

§ Scientific “impact” § yes (number of citation)

 vi. Who receives credit when a dataset from your RI is cited?

§ The RI itself

§ The RI’s institutional partners (all or in part, depending on the dataset contents) yes

§ Experts in the RI’s organization (named individuals) yes

§ “Principal investigators” in charge of measurements or data processing (named individuals) yes

§ Staff (scientists, research engineers etc.) performing the measurements or data processing (named individuals) yes

What steps in tooling, automation and presentation do you consider necessary to improve take up of identification and citation facilities and to reduce the effort required for supporting those activities?

ask to editors to check RI citation requirements for new papers

**2.** **Curation**

a. Will the responsibility for your RI’s curation activities be shared with other organisations?

no

b. Does the curation cover datasets only or also:

 i. Software? planned

 ii. Operating environment? no

 iii. Specifications/documentation?

Store software together with dataset, but not versions of data. In the past, only kept the last version of data, but in the new project, IGAS, will keep all versions of data.

c. What is your curation policy on retaining/discarding

 i. Datasets? everything kept

 ii. Software? git for versioning

 iii. Operating environments? no

 iv. Documents? all

d. How will data accessibility be maintained for the long term? E.g. What is your curation policy regarding media migration?

Institute’s policy

e. Do you track all curation activities with a logging system?

no

f. What metadata standards do you use for providing

 i. Discovery, ISO 19115, INSPIRE

 ii. Contextualisation (including rights, privacy, security, quality, suitability...) ISO 19115

 iii. Detailed access-level (i.e. connecting software to data within an operating environment)? General link to portal

Please supply documentation.

g. If you curate software how do you do it? Preserving the software or a software specification? Unit test for no regression

h. What provisions will you make for curating workflows or other processing procedures/protocols?

[Answers here]

i. If you curate the operating environment how do you do it? Preserving the environment or an environment specification?

[Answers here]

j. What steps in tooling, automation and presentation do you consider necessary to improve take up of curation facilities and to reduce the effort required for curation?

[Answers here]

**3.** **Cataloguing**

a. Do you use catalogues or require using catalogues for the following items?

 i. Observation system require

 ii. Data processing system require

 iii. Observation event and collected sample maybe

 iv. Data processing event no

 v. Data product yes

 vi. Paper or report product no

 vii. Research objects or feature interest (e.g. site, taxa, ...) yes (airport)

 viii. Services (processing, discovery, access, retrieval, publishing, visualization, etc.) no

b. For each used or required catalogue, consider the following questions:

 **i.** **Item descriptions:** ISO 19115

§ Which fields do you use for describing items?

ISO 19115 fields

§ Which standards do you apply for these fields (format or standard model)?

ISO 19115

§ Do you use controlled vocabularies for these fields? If so, please cite the vocabulary service providers.

NetCDF-CF for parameters, WMO on going task for controlled vocabulary for atmosphere

§ Do you maintain a cross-link or inter-links between:

o Catalogue items (e.g. between observation system - observation event - result dataset)?

ISO19115 to sensorml

o Fields for item description and actual items (e.g. between dataset fields - dataset access services or between sample fields - label on the sample)?

[Answers here]

§ Which repositories/software do you use to manage your metadata?

Managed by AERIS (geonetwork)

 **ii.** **Inputs:**

§ ***Human inputs:***Do you provide/need facilities for editors/reviewers to maintain the metadata in the catalogues (e.g. forms, validation workflow, etc.)? If so, please describe them briefly.

NA (machine input)

§ ***Machine inputs:*** Do you use/ need automated harvesting to populate your catalogues? If so, which protocol do you use (e.g. csw, oai-pmh, other, specific)?

no

§ How do you manage duplicates? i.e. Do you apply governance rules in a network of catalogues, do you use unique identifiers, or take other actions?

yes, unique identifiers

 **iii.** **Human outputs:**

§ What specific feature is provided/required in your web discovery function (multi-criteria search, graphical selection components (e.g. map, calendar), facets, keyword or natural language)?

Not natural language

§ Do you evaluate the accessibility, quality, and usage of your catalogue by using a dashboard or value-added products? If so, do you provide/need:

o Popularity or Usage feedback?

o Any other synthesis, indicators or statistics?

no (but planned in AERIS portal)

If so, please describe them shortly.

§ Is the catalogue freely readable or do you apply a specific authorization scheme? If you are applying a specific authorization scheme, please cite the authentication system (SSO) and the authorization rules.

free

 **iv.** **Machine outputs:**

§ Do you provide/need machine interfaces for accessing your catalogues? If so, which protocols are implemented?

CSW (on going)

§ Do you need to fit in Applicable Regulations requirements (e.g. INSPIRE) or embedding frameworks (GEOSS, OBIS)? If so, please cite the regulation, applicable interface requirements, impacts (format, performance, access policy, etc.) on your catalogues.

INSPIRE

**4.** **Processing**

**a.** **Data processing desiderata: input**

 i. What data are to be processed? What are their:

§ Typologies tabular data

§ Volume per flight 1Mo

§ Velocity human operation (3 days for validation / 2-6 months for calibration)

§ Variety homogeneous

 ii. How is the data made available to the analytics phase? By file, by web (stream/protocol), etc.

file

 iii. Please provide concrete examples of data.

example of NASA Ames provided

**b.** **Data processing desiderata: analytics**

 i. Computing needs quantification:

§ How many processes do you need to execute?

One per level of data

§ How much time does each process take/should take?

For a flight : Few seconds for levels 0 to 2. Level 4: 3 hours

§ To what extent processing is or can be done in parallel?

Flight granularity

 ii. Process implementation:

§ What do you use in terms of:

o Programming languages?

Java, Python, Fortran

o Platform (hardware, software)?

Linux, opensource softwares

o Specific software requirements?

[Answers here]

§ What standards need to be supported (e.g. WPS) for each of the above?

none

§ Is there a possibility to inject proprietary/user defined algorithms/processes for each of the above?

no

§ Do you use a sandbox to test and tune the algorithm/process for each of the above?

no

 iii. Do you use batch or interactive processing?

yes

 iv. Do you use a monitoring console?

Nagios for hardware management. We plan to use nifi for dataflow management

 v. Do you use a black box or a workflow for processing?

§ If you use a workflow for processing, could you indicate which one (e.g. Taverna, Kepler, proprietary, etc.)

§ Do you reuse sub-processes across processes?

Black box so far

 vi. **Please provide concrete examples** of processes to be supported/currently in use;

see schema above

c. **Data processing desiderata: output**

 i. What data are produced? Please provide:

§ Typologies tabular

§ Volume L2+L4 = 10Mo per flight

§ Velocity L2: 2-6 months for calibration (L4 produced automatically when data level is changed)

§ Variety homogeneous

 ii. How are analytics outcomes made available?

Available on download but no web-based workspace

d. **Statistical questions**

 i. Is the data collected with a distinct question/hypothesis in mind? Or is simply something being measured?

measured

e. Will questions/hypotheses be generated or refined (broadened or narrowed in scope) after the data has been collected? (N.B. Such activity would not be good statistical practice)

no

**f.** **Statistical data**

 i. Does the question involve analysing the responses of a single set of data (univariate) to other predictor variables or are there multiple response data (bi or multivariate data)? no

no

 ii. Is the data continuous or discrete? discrete

 iii. Is the data bounded in some form (i.e. what is the possible range of the data)? aircraft data (flight granularity)

 iv. Typically how many datums approximately are there? One each 4 seconds during flight

**g.** **Statistical data analysis NA**

 i. Is it desired to work within a statistics or data mining paradigm? (N.B. the two can and indeed should overlap!)

[Answers here]

 ii. Is it desired that there is some sort of outlier/anomaly assessment?

[Answers here]

 iii. Are you interested in a statistical approach which rejects null hypotheses (frequentist) or generates probable belief in a hypothesis (Bayesian approach) or do you have no real preference?

[Answers here]

**5.** **Provenance**

a. Do you already have data provenance recording in your RI? yes

If so:

b. Where/when do you need it, e.g., in the data processing workflows, data collection/curation procedures, versioning control in the repositories etc.?

processing workflow, collection

c. What systems are you using?

GIT for code, no standard for data

d. What standards are you using?

 i. Advantages/disadvantages

 ii. Have you ever heard about the PROV-O standard?

no standard

e. Do you need provenance tracking? yes

 i. If so, which information should be contained? version of the data, version of the data processing

f. What information do you need to record regarding the following:

 i. Scientific question and working hypothesis? no

 ii. Investigation design? no

 iii. Observation and/or measurement methods? yes

 iv. Observation and/or measurement devices? yes

 v. Observation/measurement context (who, what, when, where, etc.)? yes

 vi. Processing methods, queries? yes

 vii. Quality assurance? yes

g. Do you know/use controlled vocabularies, e.g. ontologies, taxonomies and other formally specified terms, for the description of the steps for data provenance? no

h. What support, e.g. software, tools, and operational procedures (workflows), do you think is needed for provenance tracking?

[Answers here]

i. How does your community use/plan to use the provenance information?

 i. Do you have any tools or services in place/planned for this purpose?

Produce this information in the frame of the IGAS Work package 4 (IAGOS for Copernicus Atmospheric Service, FP7)

no tools planned yet

**6.** **Optimization NA**

15. Related to your answer to the generic question 7 (What part of your RI needs to be improved):

 i. What does it mean for this to be optimal in your opinion?

[Answers here]

 ii. How do you measure optimality in this case?

§ Are there any existing metrics being applied?

§ Are there any standard metrics applied by domain scientists in this discipline?

[Answers here]

 iii. Do you already know what needs to be done to make this optimal?

§ Is it simply a matter of more resources, better machines, or does it require a rethink about how the infrastructure should be designed?

[Answers here]

 iv. What would you *not* want from an 'optimal' solution? For example, maximizing one attribute of a component or process (e.g. execution time) might come at the cost of another attribute (e.g. ease-of-use), which ultimately may prove undesirable.

[Answers here]

**b.** **Follow-up questions to answers from other sections which suggest the need for the optimization of certain RI components.**

[Answers here]

c. Do you have any use case/scenarios to show potential bottlenecks in 1) the functionality of your RI, for example the storage, access and delivery of data, doing processing, handling the workflow complexity etc. 2) ensuring the non-functional requirements of your RI, for example ensuring load balance in resource usage etc.

[Answers here]

d. To understand those bottlenecks:

 i. what might be the peak volume in accessing, storing, and delivering data?

[Answers here]

 ii. what complexity might the data processing workflow have?

[Answers here]

 iii. Are there any specific quality requirements for accessing, delivering or storing data, in order to handle the data in nearly real time?

[Answers here]

7. **Community support**

We define Community Support as being concerned with managing, controlling and tracking users' activities within an RI and with supporting all users to conduct their roles in their communities. It includes many miscellaneous aspects of RI operations, including for example (non-exhaustively) authentication, authorization and accounting, the use of virtual organizations, training and helpdesk activities.

a. **Training Requirements**

 i. Do you use or plan to use e-Infrastructure technology?

Not at the moment.

 ii. What is your community training plan?

Do not have specific training plan. Dataset is simple for user to use.

 iii. Does your community consider adopting e-Infrastructure solutions (e.g., Cloud, Grid, HPC, cluster computing).

[Answers here]

 iv. Is your community interested in training courses that introduce state-of-the-art e-Infrastructure technology?

Probably useful for the ICT team

 v. What topics (related to e-Infrastructure solutions) would your community be interested in?

[Answers here]

 vi. Who would be audience?

§ *Please describe their knowledge background of e-Infrastructure technology*

[Answers here]

 vii. What are appropriate methods to deliver this training?

[Answers here]

b. **Requirements for the Community Support Subsystem:**

 i. What are the required functionalities of your Community Support capability?

[Answers here]

 ii. What are the non-functional requirements, e.g., privacy, licensing, performance?

[Answers here]

 iii. What standards do you use, e.g., related to data, metadata, web services?

[Answers here]

 iv. What community software/services/applications do you use?

[Answers here]