**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:**

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

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

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

**A.1.1** Could you describe a basic use-case involving interaction with the RI?

SIOS is a distributed system for earth science, including biological, marine, etc. Students and scientists access observing services through data manage system which is under development.

**A.1.2.** Could you describe how data is acquired, curated and made available to users?

Data is made available from each data management system in each organisation. data is accessed through data portal. users can access different observations stream from different organisations. each organisation manages own data, in future users will be able to access integrated data sets and services.

**A.1.3** Could you describe the software and computational environments involved?

[refer to data manage plan]

**A.1.4** What are the responsibilities of the users who are involved in this use case?

Different kind of uses, report back publications, results of use. due to data manage policy, use of SIOS’s data need to cited.

**A.1.5** Do you have any use case involving interactions with other RIs (in the same or different domains?)

There is overlap between SIOS and ICOS and EMSO, signed agreement with EMSO and ICOS to share data and coordinate investment in RIs. in the next 3 years, SIOS data management will be implemented will address the problem

**A.2** What datasets are available for sharing with other RIs as part of ENVRIplus? Under what conditions are they available?

SIOS is under development, for time being, data is owned by different organisations, a number of data will be available, eg. dataset for marine will be relevant for EMSO, others include with INTERACT and ICOS.

**A.3** Apart from datasets, does your RI also bring to ENVRIplus:

**A.3.1** Software? In this case, is it open source?

It should be possible but will be decided in the frame of SIOS implementation phase

**A.3.2** Computing resources (for running datasets through your software or software on your datasets)?

This has to be evaluated in the frama of implementation phase

**A.3.3** Access to instrumentation/detectors or lab equipment? If so, what are the open-access conditions? Are there any bilateral agreements?

access program could be relevant. SIOS has agreement with INTERACT and EMSO. has discussed agreement with ICOS and EM (greenland ecological monetary)

**A.3.4** Users/expertise to provide advice on various topics?

[based on the SIOS implementation phase and on organizing aspects ]

**A.3.5** Access to related scholarly publications?

the SIOS RI doesn’t contain a common publication repository, but each individual organization has its own.

**A.3.6** Access to related grey literature (e.g. technical reports)?

[yes ]

**A.4** What plans does your RI already have for data, its management and exploitation?

**A.4.1** Are you using any particular standard(s)?

 i. Strengths and weaknesses

[each partner of the SIOS has its own infrastructure, but not all use the same standards. ]

 **A.4.2** Are you using any particular software(s)?

 i. Strengths and weaknesses

[can’t answer]

 **A.4.3** 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.

[SIOS do not consider to change standards and software, it will provide interoperability ]

**A.5** What part of your RI needs to be improved in order:

**A.5.1** For the RI to achieve its operational goals?

[can ‘t answer]

**A.5.2** For you to be able to do your work?

[easy access to discover visualize and download data]

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

 **A.6.1** If so please provide the documentation/links

[not at the moment ]

**A.7** 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

[can’t answer

**A.8** Do you have an overall approach to security and access?

[yes

**A.9** Are your data, software and computational environment subject to an open-access policy?

[yes]

**A.10** What are the big open problems for your RI pertinent to handling and exploiting your data?

[can’t answer]

**A.11.** Are you interested in any particular topic [1-6] to discuss in more detail?

**A.11.1** If so, would you like us to arrange a follow up interview with more detail questions about any particular topic to be discussed?

[can’t answer]

**A.12**. 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.

[can’t answer ]

**B. Specific questions per topic**

**B.1 Community support**

We define a Community Support as a subsystem 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.

**B.1.1.** **Requirements for the Community Support Subsystem:**

 **i.** How many communities do you support: users, developers or others? These communities may require different support mechanisms.

This was never checked.

**ii.** What are the required functionalities of your Community Support Subsystem?

[can’t answer ]

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

[can’t answer ]

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

[ISO 19115 ISO 19139]

**v.** What community software/services/applications do you use?

Can’t answer]

**B.1.2.**  **Training Requirements**

i. What is your community training plan?

SIOS communities is very divers, many organisation has their own training activities. Can be for students or scientists. An example: The University Centre in Svalbard (UNIS) has its own high quality training program for new students related to field security, i.e. how to operate safe and in accordance with environmental regulations, for all students and scientists.

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

*WP15 can develop and deliver training about methodologies, infrastructures, tools and services for those who want to build environments for big data. The content and training can benefit those tool developers who want to create new environments for scientists, for the users of ENVRIPLUS Research Infrastructures. The topics covered in this area include:*

* *􏰀 Methodologies, tools and e-infrastructures for high-throughput, high-performance and cloud computing*
* *Application porting and integration approaches to clouds and grids*
* *Approaches, tools and online services for data storage, organisation, transfer and processing*
* *Workflows and pipelines - organising and sharing multi-stage simulations at community level*
* *Scientific gateways - integrate applications, data and services into web-based portals*
* *Developing PaaS systems for application developers*
* *Developing SaaS systems for scientific end users*

***o*** *WP15 plans to develop and deliver training about building e-infrastructures, federated infrastructures for scientific communities. The content and training can benefit the IT operators of RIs, those who need to build and operate IT infrastructures to support environmental sciences data, applications, tools and environments. The topics covered in this area include:*

* *􏰀 Deploying clusters and desktop infrastructures for high-throughput or high-performance computing*
* *Deploying virtualisation and hypervisor technologies to build IaaS clouds*
* *Federating cloud systems into multi-organisational and multi-national Virtual Organisations (including connecting clouds to monitoring, accounting, user management and resource allocation systems)*

*o Facilitate harmonisation of e-infrastructure training content and events among the European e-infrastructures to maximise benefits for ENVRIPLUS RIs.*

[It is possible]

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

[ probably yes ]

**B.2.** **Identification and Citation**

**B.2.1.** **Identification**

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

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

[ sensor ,, measurements category , time ]

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

[yearly, montly, daily, possibly hourly ]

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

[measurement area/region, measurement station, ]

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

[they can be stored separately as well as in a database ]

 **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?

[can’t answer]

 **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?

[yes]

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

[yes]

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

[yes]

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

* “Raw” sensor data?

[this will be decided/discussed during the implementation phase]

* Physical samples?

[as above ]

* Data undergoing processing (QA/QC etc.)?

[as above]

* Finalized “publishable” data?

[as above]

**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?

[can’ t answer ]

 **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)?

[can’ t answer]

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

[can’ t answer]

**B2.2 Citation**

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

[testing theory, comparison model result, process 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?

[yes]

* By DOIs, if available, in the References section?

[yes]

* By using other information?

[can’t answer]

**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

[yes ]

* Geographic selection

[yes]

* Specific parameters or observables

[yes]

i**v.** 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?

[can’t answer]

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

* Downloads/access
* Visualization at your own data portal
* Visualization at other data portals
* References in scientific literature
* References in non-scientific literature
* Scientific “impact”

[they could all used ]

**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)
* Experts in the RI’s organization (named individuals)
* “Principal investigators” in charge of measurements or data processing (named individuals)
* Staff (scientists, research engineers etc.) performing the measurements or data processing (named individuals)

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?

 All of the point listed can have credits, a standard way should be defined when metadata are provided. ]

**B.3. Curation**

**B.3.1** Will the responsibility for your RI’s curation be shared with other organisations?

[can’t answer]

**B.3.2** Does the curation cover datasets or also:

**i.** Software?

 [can’t answer]

**ii**. Operating environment?

 [Answers here]

**iii.** Specifications/documentation?

 [can’t answer]

**B.3.3.**  What is your curation policy on discarding

 **i**. Datasets?

 **ii**. Software?

 **iii**. Operating environments?

 **iv**. Documents?

[can’t answer]

**B.3.4**. How will data accessibility be maintained for the long term? E.g. What is your curation policy regarding media migration?

[can’t answer]

**B.3.5**. Do you track with a logging system all curation activities?

[can’t answer]

**B.3.6** What metadata standards do you use for providing

 i. Discovery,

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

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

Please supply documentation.

[ISO 19115]

**B.3.7**. If you curate software how do you do it? Preserving the software or a software specification?

i. What provisions will you make for curating workflows?

[can’t answer]

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

[can’t answer]

iii. 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?

[can’t answer]

**B.4. Cataloguing**

**B.4.1**. Do you use catalogues or require using catalogues for the following items?

i. Observation system

ii. Data processing system

iii. Observation event and collected sample

iv. Data processing event

v. Data product

vi. Paper or report product

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

[it is planned to use catalog]

**B.4.2**. For each used or required catalogue, consider the following questions:

**i.** Item descriptions:

* Which fields do you use for describing items?

[alfanumerical]

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

[ISO19115]

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

[can’t answer]

* Do you maintain a cross-link or inter-links between:
	+ Catalogue items (e.g. between observation system - observation event - result dataset)?

[can’t answer]

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

[can’t answer]

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

[interoperable local repository]

 **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.

[can’t answer]

* ***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)?

[can’t answer]

* 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?

[can’t answer]

**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)?

[ multicriteria, keywords, map ,calendar ]

* 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?

If so, please describe them shortly.

[can’t answer]

* 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.

[can’t answer]

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

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

[can’t answer]

* 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.

 [can’t answer]

**B.5. Processing**

**B.5.1** **Data processing desiderata: input**

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

* Typologies
* Volume
* Velocity
* Variety

 [can’t answer

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

 [can’t answer]

iii. Please provide concrete examples of data.

 [can’t answer]

**B.5.2.** **Data processing desiderata: analytics**

i. Computing needs quantification:

* How many processes do you need to execute?
* How much time does each process take/should take?

 [can’t answer]

 ii. Process implementation:

* What do you use in terms of:
	+ Programming languages?
	+ Platform?
	+ Specific software requirements?

[can’t answer]

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

[Answers here]

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

[Answers here]

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

[Answers here]

 iii. Do you use batch or interactive processing?

[Answers here]

iv. Do you use a monitoring console?

[Answers here]

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

* Do you reuse sub-processes across processes?

[Answers here]

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

[Answers here]

**B.5.3**. **Data processing desiderata: output**

i. What data are produced? Please provide:

* Typologies
* Volume
* Velocity
* Variety

[Answers here]

ii. How are analytics outcomes made available?

[Answers here]

**B.5.4**. **Statistical questions**

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

[Answers here]

ii. 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)

[Answers here]

**B.5.5.** **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)?

[Answers here]

ii. Is the data continuous or discrete?

[Answers here]

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

iv. Typically how many datums approximately are there?

[Answers here]

**B5.6.** **Statistical data analysis**

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]

**B.6. Provenance**

**B.6.1**. Where/when do you need provenance in the RI, e.g., in the data processing workflows, data collection/curation procedures, versioning control in the repositories etc.?

 [Answers here]

**B.6.2.** Which information should be retained if your RI needs provenance tracking?

 [Answers here]

**B.6.3**. What information do you need to record regarding the following:

i. Scientific question and working hypothesis?

 [Answers here]

ii. Investigation design?

 [Answers here]

iii. Observation and/or measurement methods?

 [Answers here]

iv. Observation and/or measurement devices?

 [Answers here]

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

 [Answers here]

**B.6.4**. Do you know/use controlled vocabularies, e.g. ontologies, taxonomies and other formally specified terms, for the description of the steps for data provenance?

 [Answers here]

**B.6.5**. What support, e.g. software, tools, and operational procedures, do you think is needed for provenance tracking?

 [Answers here]

**B.7. Optimization**

Related to answer (part x) from general question 2:

**B.7.1.** Do you have use case/scenarios to show potential bottlenecks in 1) using RI, such as storing, accessing and delivering data, doing processing, handling the workflow complexity etc. 2) managing RIs, such as load balance in resource usage etc[ZZ1] .

 [Answers here]

**B.7.2**. To understand those bottlenecks,

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

ii. what complexity might the data processing workflow have?

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

iv. …

 [Answers here]

**B.7.3.** What does it mean for x to be optimal in your opinion?

 [Answers here]

**B.7.4.** How do you measure optimality in this case?

 i. Are there any existing metrics being applied?

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

 [Answers here]

**B.7.5.** Do you already know what needs to be done to make *x* optimal?

i. 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]

**B.7.6.** 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]