**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | ENVRI, Common Operations of Environmental Research Infrastructure | |
| **Vertical (area)** | | Environmental Science | |
| **Author/Company/Email** | | Yin Chen/ Cardiff University/ [ChenY58@cardiff.ac.uk](mailto:ChenY58@cardiff.ac.uk) | |
| **Actors/Stakeholders and their roles and responsibilities** | | The ENVRI project is a collaboration conducted within the European Strategy Forum on Research Infrastructures (ESFRI) Environmental Cluster. The ESFRI Environmental research infrastructures involved in ENVRI including:   * ICOS is a European distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks. * EURO-Argo is the European contribution to Argo, which is a global ocean observing system. * EISCAT-3D is a European new-generation incoherent-scatter research radar for upper atmospheric science. * LifeWatch is an e-science Infrastructure for biodiversity and ecosystem research. * EPOS is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics. * EMSO is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards.   ENVRI also maintains close contact with the other not-directly involved ESFRI Environmental research infrastructures by inviting them for joint meetings. These projects are:   * IAGOS Aircraft for global observing system * SIOS Svalbard arctic Earth observing system   ENVRI IT community provides common policies and technical solutions for the research infrastructures, which involves a number of organization partners including, Cardiff University, CNR-ISTI, CNRS (Centre National de la Recherche Scientifique), CSC, EAA (Umweltbundesamt Gmbh), EGI, ESA-ESRIN, University of Amsterdam, and University of Edinburgh. | |
| **Goals** | | The ENVRI project gathers 6 EU ESFRI environmental science infra-structures (ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS, and EMSO) in order to develop common data and software services. The results will accelerate the construction of these infrastructures and improve interoperability among them.  The primary goal of ENVRI is to agree on a reference model for joint operations. The ENVRI Reference Model (ENVRI RM) is a common ontological framework and standard for the description and characterisation of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures. The ENVRI RM serves as a common language for community communication, providing a uniform framework into which the infrastructure’s components can be classified and compared, also serving to identify common solutions to common problems. This may enable reuse, share of resources and experiences, and avoid duplication of efforts. | |
| **Use Case Description** | | ENVRI project implements harmonised solutions and draws up guidelines for the common needs of the environmental ESFRI projects, with a special focus on issues as architectures, metadata frameworks, data discovery in scattered repositories, visualisation and data curation. This will empower the users of the collaborating environmental research infrastructures and enable multidisciplinary scientists to access, study and correlate data from multiple domains for "system level" research.  ENVRI investigates a collection of representative research infrastructures for environmental sciences, and provides a projection of Europe-wide requirements they have; identifying in particular, requirements they have in common. Based on the [analysis evidence](http://confluence.envri.eu:8090/download/attachments/327687/D3.3%20Analysis%20of%20Requirements%20V1.0.pdf?version=1&modificationDate=1366965933706&api=v2), the ENVRI Reference Model ([www.envri.eu/rm](http://www.envri.eu/rm)) is developed using ISO standard Open Distributed Processing. Fundamentally the model serves to provide a universal reference framework for discussing many common technical challenges facing all of the ESFRI-environmental research infrastructures. By drawing analogies between the reference components of the model and the actual elements of the infrastructures (or their proposed designs) as they exist now, various gaps and points of overlap can be identified. | |
| **Current**  **Solutions** | **Compute(System)** | |  |
| **Storage** | | File systems and relational databases |
| **Networking** | |  |
| **Software** | | Own |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Most of the ENVRI Research Infrastructures (ENV RIs) are *distributed, long-term, remote controlled observational networks* focused on understanding processes, trends, thresholds, interactions and feedbacks and increasing the predictive power to address future environmental challenges. They are spanning from the Arctic areas to the European Southernmost areas and from Atlantic on west to the Black Sea on east. More precisely:   * ***EMSO*,** network of fixed-point, deep-seafloor and water column observatories, is geographically distributed in key sites of European waters, presently consisting of thirteen sites. * ***EPOS*** aims at integrating the existing European facilities in solid Earth science into one coherent multidisciplinary RI, and to increase the accessibility and usability of multidisciplinary data from seismic and geodetic monitoring networks, volcano observatories, laboratory experiments and computational simulations enhancing worldwide interoperability in Earth Science. * ***ICOS*** dedicates to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks. The ICOS network includes more than 30 atmospheric and more than 30 ecosystem primary long term sites located across Europe, and additional secondary sites. It also includes three Thematic Centres to process the data from all the stations from each network, and provide access to these data. * ***LifeWatch*** is a “virtual” infrastructure for biodiversity and ecosystem research with services mainly provided through the Internet. Its Common Facilities is coordinated and managed at a central European level; and the *LifeWatch Centres* serve as specialized facilities from member countries (regional partner facilities) or research communities. * ***Euro-Argo*** provides, deploys and operates an array of around 800 floats contributing to the global array (3,000 floats) and thus provide enhanced coverage in the European regional seas*.* * ***EISCAT- 3D***, makes continuous measurements of the geospace environment and its coupling to the Earth's atmosphere from its location in the auroral zone at the southern edge of the northern polar vortex, and is a distributed infrastructure. |
| **Volume (size)** | | Variable data size. e.g.,   * The amount of data within the ***EMSO*** is depending on the instrumentation and configuration of the observatory between several MBs to several GB per data set. * Within ***EPOS***, the EIDA network is currently providing access to continuous raw data coming from approximately more than 1000 stations recording about 40GB per day, so over 15 TB per year. EMSC stores a Database of 1.85 GB of earthquake parameters, which is constantly growing and updated with refined information.   + 222705 – events   + 632327 – origins   + 642555 – magnitudes * Within ***EISCAT 3D*** raw voltage data will reach 40PB/year in 2023. |
| **Velocity**  **(e.g. real time)** | | Real-time data handling is a common request of the environmental research infrastructures |
| **Variety**  **(multiple datasets, mashup)** | | Highly complex and heterogeneous |
| **Variability (rate of change)** | | Relative low rate of change |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Normal |
| **Visualization** | | Most of the projects have not yet developed the visualization technique to be fully operational.   * ***EMSO*** is not yet fully operational, currently only simple graph plotting tools. * Visualization techniques are not yet defined for ***EPOS***. * Within ***ICOS*** Level-1.b data products such as near real time GHG measurements are available to users via ATC web portal. Based on Google Chart Tools, an interactive time series line chart with optional annotations allows user to scroll and zoom inside a time series of CO2 or CH4 measurement at an ICOS Atmospheric station. The chart is rendered within the browser using Flash. Some Level-2 products are also available to ensure instrument monitoring to PIs. It is mainly instrumental and comparison data plots automatically generated (R language & Python Matplotlib 2D plotting library) and daily pushed on ICOS web server. Level-3 data products such as gridded GHG fluxes derived from ICOS observations increase the scientific impact of ICOS. For this purpose ICOS supports its community of users. The Carbon portal is expected to act as a platform that will offer visualization of the flux products that incorporate ICOS data. Example of candidate Level-3 products from future ICOS GHG concentration data are for instance maps of European high-resolution CO2 or CH4 fluxes obtained by atmospheric inversion modelers in Europe. Visual tools for comparisons between products will be developed by the Carbon Portal. Contributions will be open to any product of high scientific quality. * ***LifeWatch*** will provide common visualization techniques, such as the plotting of species on maps. New techniques will allow visualizing the effect of changing data and/or parameters in models. |
| **Data Quality (syntax)** | | Highly important |
| **Data Types** | | * Measurements (often in file formats), * Metadata, * Ontology, * Annotations |
| **Data Analytics** | | * Data assimilation, * (Statistical) analysis, * Data mining, * Data extraction, * Scientific modeling and simulation, * Scientific workflow |
| **Big Data Specific Challenges (Gaps)** | | * Real-time handling of extreme high volume of data * Data staging to mirror archives * Integrated Data access and discovery * Data processing and analysis | |
| **Big Data Specific Challenges in Mobility** | | The need for efficient and high performance mobile detectors and instrumentation is common:   * In **ICOS**, various mobile instruments are used to collect data from marine observations, atmospheric observations, and ecosystem monitoring. * In **Euro-Argo**, thousands of submersible robots to obtain observations of all of the oceans * In **Lifewatch**, biologists use mobile instruments for observations and measurements. | |
| **Security & Privacy**  **Requirements** | | Most of the projects follow the open data sharing policy. E.g.,   * The vision of **EMSO** is to allow scientists all over the world to access observatories data following an open access model. * Within **EPOS**, EIDA data and Earthquake parameters are generally open and free to use. Few restrictions are applied on few seismic networks and the access is regulated depending on email based authentication/authorization. * The **ICOS** data will be accessible through a license with full and open access. No particular restriction in the access and eventual use of the data is anticipated, expected the inability to redistribute the data. Acknowledgement of ICOS and traceability of the data will be sought in a specific, way (e.g. DOI of dataset). A large part of relevant data and resources are generated using public funding from national and international sources. * **LifeWatch** is following the appropriate European policies, such as: the European Research Council (ERC) requirement; the European Commission’s open access pilot mandate in 2008. For publications, initiatives such as Dryad instigated by publishers and the Open Access Infrastructure for Research in Europe (OpenAIRE). The private sector may deploy their data in the LifeWatch infrastructure. A special company will be established to manage such commercial contracts. * In **EISCAT 3D**, lower level of data has restrictions for 1 year within the associate countries. All data open after 3 years. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Different research infrastructures are designed for different purposes and evolve over time. The designers describe their approaches from different points of view, in different levels of detail and using different typologies. The documentation provided is often incomplete and inconsistent. What is needed is a uniform platform for interpretation and discussion, which helps to unify understanding.  In ENVRI, we choose to use a standard model, Open Distributed Processing (ODP), to interpret the design of the research infrastructures, and place their requirements into the ODP framework for further analysis and comparison. | |
| **More Information (URLs)** | | * ENVRI Project website: [www.envri.eu](http://www.envri.eu) * ENVRI Reference Model [www.envri.eu/rm](http://www.envri.eu/rm) * [ENVRI deliverable D3.2](http://confluence.envri.eu:8090/download/attachments/327687/D3.3%20Analysis%20of%20Requirements%20V1.0.pdf?version=1&modificationDate=1366965933706&api=v2) : Analysis of common requirements of Environmental Research Infrastructures * ICOS: <http://www.icos-infrastructure.eu/> * Euro-Argo: <http://www.euro-argo.eu/> * EISCAT 3D: <http://www.eiscat3d.se/> * LifeWatch: <http://www.lifewatch.com/> * EPOS: <http://www.epos-eu.org/> * EMSO <http://www.emso-eu.org/management/> | |
| **Note:** <additional comments> | | | |