ENVRI Reference Model

Yin Chen

ISR Roadmap Workshop, Kiruna, 10-11 March, 2015
Why we need it?
ENVRI collaborates with ESFRI Environmental Science Research Infrastructures

- EISCAT 3D: Upper space observing
- EURO Argo: Open sea observing
- ICOS: Greenhouse observing
- COPAL: Tropospheric research aircraft
- LifeWatch: Biodiversity observing
- EMSO: Deep sea observing
- IAGOS: Aircraft for global observing
- EPOS: Volcano, earthquake observing
- SIOS: Svalbard arctic Earth observing
Why we need it?

- To help the community reach a common vision
- To provide a common language for communication
- To provide a uniform framework into which RIs’ components can be placed/compared
- To provide common solutions to common problems
- To secure interoperability
- To enable reuse, share of resource/experiences, avoid duplication efforts
What is ENVRI Reference Model?

Top 500 most frequently used words in the Reference Model
What is ENVRI Reference Model?

- A standard for description/characterisation of data, computation of environmental RIs (Chen 2013)

- An abstract conceptual Model
  - captures common requirements
  - captures state-of-the-art design experiences
  - With a view of informing future implementation

- A common framework

- A taxonomy of terms, concepts and definitions
How did we build it?

- Based on the analysis of common requirements of a collection of representative Environmental Research Infrastructures
Subsystems with points of references between them
ENVRI Common Subsystems

- **Acquisition** -- brings the measures/data streams into the system (non-reproducible data)
- **Curation** -- manages/maintains quality data (reproducible data)
- **Access** -- facilities discovery, access (published data)
- **Processing** -- facilities analysis/mining/experiments (combined/derived data)
- **Community Support** -- supports users to conduct their roles in communities (user generated data)
Analysis: EMSO Architecture
Analysis: LifeWatch Architecture

User Community Support
Data Processing
Data Access
Data Curation
Analysis: EURO-Argo Architecture

- **Marine governmental agencies and research institutes (including NODCs)**

- **Observing systems**

- **Other observations (e.g. scientific cruises...)**

- **Data Acquisition**

- **Data Curation**

- **Data Access**

- **Community Support**

- **MyOcean in-situ TAC**

- **SeaDataNet**

- **Data Processing**

- **International organisations (IOC/ICDE, ICES, ICES World Data Systems)**

- **Community Support**

- **Operational users**

- **Re-analysis users**

- **Vocabularies**

- **Users**

1/11/15

Project number: 283465
Microsoft: The Data-Intensive Research Lifecycle

- Acquisition
- Acquisition & Modelling
- Collaboration & Visualisation
- Analysis & Data Mining
- Dissemination & Sharing
- Archiving & Preserving
- Curation
- Community Support

Data
# Common Functions (Acquisition)

## Functions/Operations at Data Acquisition Sub-system

<table>
<thead>
<tr>
<th>Functions/Embedded Services</th>
<th>ICOS</th>
<th>EPOS</th>
<th>EMSO</th>
<th>EISCAT-3D</th>
<th>LifeWatch</th>
<th>EURO-Argo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument Integration</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Instrument Configuration</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Instrument Calibration</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Instrument Access</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Configuration Logging</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Instrument Monitoring</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td><em>(Real-time) (Parameter/Data) Visualisation</em></td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Process Control</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>No</td>
</tr>
<tr>
<td>Data Collection</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>*(Real-time) Data Collection</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Sampling</td>
<td>No</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Noise Reduction</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td><em>(Real-time) Data Transmission</em></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Transmission Monitoring</td>
<td>Yes</td>
<td>Unknown</td>
<td>No</td>
<td>No</td>
<td>Not Applicable</td>
<td>No</td>
</tr>
</tbody>
</table>
### Functions/operations in the Data Curation Sub-system

<table>
<thead>
<tr>
<th>Functions/Embedded Services</th>
<th>ICOS</th>
<th>EPOS</th>
<th>EMSO</th>
<th>EISCAT-3D</th>
<th>LifeWatch</th>
<th>EURO-Argo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Quality Checking</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Quality Verification</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Identification</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Data Cataloguing</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Data Product Generation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Versioning</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Unknown</td>
</tr>
<tr>
<td>Workflow Enactment</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>No</td>
</tr>
<tr>
<td>Data Preservation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Replication</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Replication Synchronisation</td>
<td>No</td>
<td>Unknown</td>
<td>No</td>
<td>Unknown</td>
<td>Not Applicable</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Functions/operations at Data Access Sub system

<table>
<thead>
<tr>
<th>Functions/Embedded Services</th>
<th>ICOS</th>
<th>EPOS</th>
<th>EMSO</th>
<th>EISCAT-3D</th>
<th>LifeWatch</th>
<th>Euro-Argo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Control</td>
<td>Unknown</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>Data Conversion</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Compression</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Data Visualisation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Publication</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Citation</td>
<td>No</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
<td>Unknown</td>
<td>No</td>
</tr>
<tr>
<td>(Resources/Data) Annotation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Metadata Harvesting</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Yes</td>
<td>No</td>
<td>Unknown</td>
<td>No</td>
</tr>
<tr>
<td>Resource Registration</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Semantic Harmonisation</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Data Discovery and Access</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

A full function list is on ENVRI wiki

[http://envri.eu/group/envri/wiki/-/wiki/Main/Analyse%20Common%20Requirements%20for%20Data%20Processing](http://envri.eu/group/envri/wiki/-/wiki/Main/Analyse%20Common%20Requirements%20for%20Data%20Processing)
ENVRI Reference Model
Basis: ODP Approach

Using Open Distributed Processing (ODP) (ISO/IEC 10746)

- A framework for structuring design specification for large-scale complex distributed systems
  - An object modelling approach
  - A viewpoints-based approach to architecture
  - Allow collaborative design activities
ODP Viewpoints

Enterprise
Business Aspects
The purpose, scope and policies for the organization that will own the system

Information
Information System Aspects
Information handled by the system and constraints on the use and interpretation of that information
What is it about?

Computational
Application Design Aspects
Functional decomposition of the system into objects suitable for distribution
How does each bit work?

Technology
Implementation
System hardware & software and actual distribution
With what?

Engineering
Solution Types & Distribution
Infrastructure required to support distribution
How do the bits work together?

18/03/2014
Adapted from ISO/IEC 19793, 2009
Project number: 283465
ENVRRI RM: Science Viewpoint

- Derive use scenarios from common requirements, identifying *communities, roles, behaviours*

- Model defines:
  - 5 common *Communities* in according to 5-subsystem
    - Data Acquisition: who collects raw data
    - Data Curation: who manages, archives quality data
    - Data Publication: who assists publication, discovery & access
    - Data Service Provision: who provides services to derive knowledge
    - Data Usage: who makes use of data/services

- Community *roles & behaviours*
Data-oriented approach:

- Follow **data-lifecycle** in each subsystems
- Identify **information objects, actions, state changes** when events/actions occur

Model defines:

- A set of **information objects** handled by a subsystem
- A set of **action types** that cause the states changes
- A set of **constraints** on those objects
- **Dynamic schema** -- how information objects evolve as the system operates
- **Static schema** -- allowable state changes
Service-oriented, Brokered approach

- Core functionality is encapsulated in a set of service objects
- Access to such object via brokers which provides an interoperability layer between heterogeneous components

Model defines

- A set of computational objects
  - Each encapsulates specific functionalities
  - Each provides a set of interfaces to invoke functions
- A set of binding objects to coordinate multi-party interactions
Example of Using the Reference Model
Interactions with User Communities

- **EISCAT 3D**
  - Pilot study, Feb 2013 to date, dialogue continues
  - EISCAT International Symposium, Lancaster, 10 Aug 2013

- **EPOS, EMSO, GFBio, Helsinki Uni Ownership Determination**
  - All starting to use the language and model concepts

- **RDA Data Foundation & Terminology** (use case for evaluation)

- **DASISH**
  - Teleconference, 12 Feb 2014
  - ODP & Reference Model workshop, Colchester, 17 March 2014

- **ICOS**
  - Cardiff meeting, 27-28 Jan 2014
  - Meeting with data management team, 13 March 2014
  - Workshop for thematic centres (Jun, Sep 2014)
EISCAT 3D Research Infrastructure

- **EISCAT**: European incoherent scatter radar for atmospheric, geospace research
- **EISCAT 3D**: next generation 3D imaging radar
- Studies how Earth’s atmosphere is coupled to space, is uniquely located for studies into arctic ionosphere
Using Referent Model to Analyse EISCAT 3D Data Infrastructure

- EISCAT 3D focuses on radar design
- Its data infrastructure is embedded in overall infrastructure
- Difficult for computer technologists to understand
Results of Using the ENVRI RM

- Clarified the boundary for the data infrastructure
- Clarified the functionalities specified in the design
EISCAT-3D EGI, EUDAT Pilot

Data & Computing sites

(1) Migrate files

File catalogue

Metadata catalogue

(2) Register files and metadata

(3) Lookup data and metadata

(4) Read files and run applications

EISCAT science gateway

Authentication, Authorization, Single sign-on

Processing and mining applications

App. 1

App. 2
Example of using the ENVRI Reference Model in ICOS Research Infrastructure

Relevant data products from representative networks

higher level output
Workflow of Handling ICOS Data
Computational Viewpoint: ICOS Core Computational Objects
Benefits of Using the Reference Model
Benefits of Using the RM (Immediate 1-5 years)

- **Professional framework** enables clearly define roles, processes in RI operations
- Makes it far **easier to design RI** in the construction phase
- Helps to **evaluate current RIs for division of tasks**
- Helps to find missing or duplicated actions
- **Easier definition of requirements** of IT components
- **Enabling more modular approach** for the RI IT solutions
- **Making possible to use external suppliers** (e.g. international IT co-operation projects) for the component development.
Benefits of Using the RM (Intermediate 5-10 years)

- A common language ensures common understanding
- Avoiding duplications
- Enabling re-use of components, solutions & policies
- The use of planned standard modular approach enables scalable design solutions
- Better risk management of RI development, due possibility of changing individual modules and operations of the RIs, without needing to completely redesign the systems due some ad-hoc solutions.
- Improving the trustworthiness of the RI products due clearly defined and standardized ways to present workflows
Benefits of Using the RM (Long-term 10-20 years)

- **Greater level of interoperability** through the use of common standards, enabling data usage and communication between the RIs commonplace.

- **Support of cross-disciplinary perspectives** and products and enablement of systems science approach.

- **Larger potential user base** due easier use of the RI products, which increases the impact and return of investment of RIs.
• Resources Information
Resources


- **Training Videos**: [http://tinyurl.com/rmtrainingvideos](http://tinyurl.com/rmtrainingvideos)

- **Guidelines for use**: [http://tinyurl.com/o5l7bn9](http://tinyurl.com/o5l7bn9)

- **Help line**: envri-rm-request@list.uva
Publications & Presentations


4. Zhao, Z., et. al. (2012). "Planning data intensive workflows using the Network Service Interface (NSI)", the 7th Workshop on Workflows in Support of Large-Scale Science, in the context of Supercomputing, Salt Lake City, 2012;


