THEME 2: DATA FOR SCIENCE

DR. ZHIMING ZHAO
UNIVERSITY OF AMSTERDAM
Challenge 1: support system level of sciences

Challenge 2: share solutions to common problems

Challenge 3: Interface virtual research environment(s)

Challenge 4: re-use technologies (e.g. from e-Infrastructures)

Public clouds
Objectives of the “Data for Science” theme

1. to optimize data processing and to develop common models, rules and guidance for research data workflow documentation;
2. to facilitate data discovery and use, and to provide integrated end-user information technology to access heterogeneous data sources;
3. to facilitate discovery of software services and their composition;
4. to make data citable by developing approaches with practical examples, exchange of expertise, and agreements with publishers;
5. to characterise users and build a community evolving from current RI communities;
6. to characterise ICT resources (including sensors and detectors) to allow virtualisation of the environment (for instance onto Grid- or Cloud-based platforms);
7. to facilitate the connection of users, composed software services, appropriate data and necessary resources in order to meet end-user requirements
Direct Benefits to environmental RIs

1. A **reference model** guided design and engineering approach
2. A **knowledge base** for reusable solutions to common challenges
3. A set of **recommended services** and a **reference architecture**
4. A set of **developed services deployed** on e-Infrastructures (e.g. EUDAT or EGI)

7. to facilitate the connection of users, composed software services, appropriate data and necessary resources in order to meet end-user requirements
Outline

- Overview of the *data for science* theme
- Activities and achievements
- Summary
Outline

- Overview of the *data for science* theme
- Activities and achievements
- Summary
Adopted approach: **identify common challenges**
Adopted approach: RI and e-Infrastructures

RI development: **Architecture design**

Meta information linking: **Linking model**

Common vocabulary: **Reference model**

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
Adopted approach: work package and tasks

WP 5: RM guided RI
   Task 5.1: review of existing RIs
   Task 5.2: RI characterisation and ENVRIPLUS reference model
   Task 5.3: semantic description and linking
   Task 5.4: development plan and reference architecture

WP 6: Identification & Citation
   Task 6.1: Identification/Citation

WP 7: Data processing and optimization
   Task 7.1: Processing
   Task 7.2: Optimization

WP 8: Curation, Cataloguing and Provenance
   Task 8.1: Curation
   Task 8.2: Cataloguing
   Task 8.3: Provenance

SIOS

WP 9: Validation & deployment
   Task 9.1: integration and validation
   Task 9.2: research to operation

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
Adopted approach: agile task forces

WP 5: RM guided RI

WP 6: Identification & Citation

WP 7: Data processing and optimization

WP 8: Curation, Cataloguing and Provenance

WP 9: Validation & deployment

Task 5.1: review of existing RIs

Task 5.2: RI characterisation and ENVRIPLUS reference model

Task 5.3: semantic description and linking

Agile activities (13 agile task forces)

Task 6.1: Identification/Citation

Task 6.2: Processing

Task 6.3: Curation

Task 7.1: Optimization

Task 7.2: Cataloguing

Task 7.3: Provenance

Task 8.1: Provenance

Task 8.3: Provenance

Task 9.1: Integration and validation

Task 9.2: Research to operation

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
<table>
<thead>
<tr>
<th>#</th>
<th>Use Case</th>
<th>Agile Group Leaders</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC_3</td>
<td><strong>How do mosquito born deseases emerge and what are trends?</strong></td>
<td>Matthias O. (UGOT)</td>
</tr>
<tr>
<td>TC_2</td>
<td><strong>Euro-ArgoData subscription service</strong></td>
<td>Thierry (IFREMER) &amp; Yin (EGI)</td>
</tr>
<tr>
<td>TC_4</td>
<td><strong>Sensor registry</strong></td>
<td>Thomas B. (IFREMER)</td>
</tr>
<tr>
<td>TC_16</td>
<td><strong>Description of a National Marine Biodiversity Data Archive Centre</strong></td>
<td>Dan L.(MBA) &amp; Abraham, Alex (CU)</td>
</tr>
<tr>
<td>IC_1</td>
<td><strong>Dynamic data citation, identification &amp; citation</strong></td>
<td>Alex V. (LU)</td>
</tr>
<tr>
<td>IC_2</td>
<td><strong>Provenance Implementation case</strong></td>
<td>Barbara Magagna (EAA)</td>
</tr>
<tr>
<td>IC_3</td>
<td><strong>User support to re-process data using their own algorithms (EISCAT 3D)</strong></td>
<td>Ingemar H. (EISCAT) &amp; Leonardo C.(CNR)</td>
</tr>
<tr>
<td>IC_8</td>
<td><strong>Cataloguing, curation, provenance across RIs WP8</strong></td>
<td>Keith J. (NERC)</td>
</tr>
<tr>
<td>IC_9</td>
<td><strong>Provenance - use of DOI for tracing of data re-use.</strong></td>
<td>Markus F. (NILU)</td>
</tr>
<tr>
<td>IC_10</td>
<td><strong>Domain extension of existing thesauri</strong></td>
<td>Barbara M. (EAA)</td>
</tr>
<tr>
<td>IC_11</td>
<td><strong>Semantic Linking Framework</strong></td>
<td>Zhiming, Z., Paul M. (UvA)</td>
</tr>
<tr>
<td>IC_12</td>
<td><strong>Implementation of ENVRI(plus) RM for EUFAR and LTER</strong></td>
<td>Barbara M. (EAA)</td>
</tr>
<tr>
<td>IC_13</td>
<td><strong>The eddy covariance fluxes of GHGs</strong></td>
<td>Dario P., Domenico V. (UNITUS)</td>
</tr>
<tr>
<td>IC_14</td>
<td><strong>SOS&amp;SSN ontology based data acquisition &amp; NRT data quality checking services</strong></td>
<td>Robert H., Markus S. (UniHB)</td>
</tr>
<tr>
<td>US#</td>
<td>Theme2 WPs</td>
<td>RIs</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>SC_3</td>
<td>WP7, WP8, 9</td>
<td>LifeWatch-SW, EGI</td>
</tr>
<tr>
<td>TC_2</td>
<td>WP9</td>
<td>EuroArgo, EMSO, ICOS-SOCAT, EGI, EUDAT</td>
</tr>
<tr>
<td>TC_4</td>
<td>WP8, WP5, 9</td>
<td>EMSO, EPOS, EuroArgo, Marine Gliders, ICOS, Drones, EGI</td>
</tr>
<tr>
<td>TC_6</td>
<td>WP5, WP9</td>
<td>EMSO, SeadataNet, JERICO, EMBRC, EMODNET, COPERNICUS, EGI</td>
</tr>
<tr>
<td>IC_1</td>
<td>WP6</td>
<td>ICOS, ANAEE, ACTRIS, LTER, IAGOS</td>
</tr>
<tr>
<td>IC_2</td>
<td>WP5, WP6, 8</td>
<td>LTER, ICOS, EUDAT</td>
</tr>
<tr>
<td>IC_3</td>
<td>WP7, WP9</td>
<td>EISCAT-3D, EGI</td>
</tr>
<tr>
<td>IC_8</td>
<td>WP8</td>
<td>EMSO, EuroArgo, EPOS, ICOS, LTER, EUDAT</td>
</tr>
<tr>
<td>IC_10</td>
<td>WP5, WP8</td>
<td>LTER, EMBEC, LifeWatch-ITA</td>
</tr>
<tr>
<td>IC_11</td>
<td>WP5</td>
<td>LTER, ICOS</td>
</tr>
<tr>
<td>IC_12</td>
<td>WP5</td>
<td>EUFAR, LTER</td>
</tr>
<tr>
<td>IC_13</td>
<td>WP7, WP9</td>
<td>ICOS, LTER, ANAEE</td>
</tr>
<tr>
<td>IC_14</td>
<td>WP4, WP9</td>
<td>EMSO, FIXO3, ANAEE, EGI</td>
</tr>
</tbody>
</table>
Theme 2 activities during last two years
Phase 1: requirement, technology and gap analysis

- **Kickoff**
- **Requirement collection**
- **Technology review Gap analysis**
- **Call for use case**
- **Establish Agile task forces**

M1 - M24

WP 5,6

D5.1
Phase 2: Reference model guided service design

- WP 5,6: Requirement collection
- WP 7-8: Technology review Gap analysis
- WP 9: Refine ENVRI RM
- Design services WP6-8
- M13: D5.1, D5.2, D5.4, D6.1, D7.1, D7.3, D8.1, D8.3
- M16: Use case mini workshop

Kickoff, Call for use case, Establish Agile task forces
Phase 3: service prototype and validation

- **Kickoff:** M1, M4, M7, M10
- **Requirement collection:** WP 5,6
- **Technology review Gap analysis:** WP 7-8
- **WP 9:** Refine ENVRI RM
- **Design services WP6-8:** ENVRI RM (Eng/Tech viewpoint)
- **Prototype services WP6-8:**

Use case mini workshop:
- D5.1
- D5.2, D5.4, D6.1, D7.1, D7.3, D8.1, D8.3

- **Call for use case:**
- **Establish Agile task forces:**

RI f2f meetings:
- D5.5

- **D5.2, D5.4, D6.1, D7.1, D7.3, D8.1, D8.3**

H2020 Project

Project Number: 654182
Milestones: timeline

- **M1** to **M4**: Kickoff, Requirement collection, Technology review, Gap analysis
- **M7** to **M10**: Establish Agile task forces
- **M13** to **M16**: Design services
- **M18** to **M20**: Refine ENVRI RM
- **M22** to **M24**: Design services

**WP5,6**
- Call for use case
- Establish Agile task forces

**WP7-8**
- Technology review
- Gap analysis

**WP9**
- Requirement collection

**WP 6-8**
- Prototype services

**WP 6-8**
- Design services

**ENVRI RM**
- (Eng/Tech viewpoint)

**WP 7-8**
- ENVRI RM
- (Eng/Tech viewpoint)
- Requirement collection

**WP5,6,7,8**
- Establish Agile task forces

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**WP 6-8**
- Design services

**MS19** (Reference model)
- **MS20** (development plan)
- **MS22** (curation and cataloguing)
- **MS25** (processing and optimization)
Milestones: deviations

- M1
- M4
- M7
- M10
- M13
- M16
- M18
- M20
- M21
- M24

RM task

WP 5, 6
WP 7-8
WP 9

Prototype services WP6-8

Kickoff
Requirement collection
Technology review Gap analysis
Refine ENVRI RM
Design services WP6
ENVRI RM (Eng/Tech viewpoint)

WP 7-8
WP 5, 6

Establish Agile task forces

Call for use case

M1
M4
M7
M10
M13
M16
M18
M20
M21
M24

RM task

MS19, MS20, MS22, MS25
D5.2, D5.4, D6.1, D7.1, D7.2, D8.1, D8.3

D5.1
D5.4
D5.5

Use case workshops

D5.1
D5.4
D5.5

Milestones:

- M1:
  - RM task
- M4:
  - Requirement collection
- M7:
  - Technology review Gap analysis
- M10:
  - Refine ENVRI RM
- M13:
  - Design services WP6
- M16:
  - ENVRI RM (Eng/Tech viewpoint)
- M18:
  - WP 5, 6
- M20:
  - WP 7-8
- M21:
  - WP 9
- M24:
  - Prototype services WP6-8

H2020 Project

Project Number: 654182
Outline

- Overview of the *data for science* theme
- **Activities and achievements**
- Summary
WP 5: **Reference model guided RI design**

**WP 5:** RM guided RI

- **WP 6:** Identification & Citation
  - Task 6.1: Identification/Citation
  - Task 6.2: Processing
  - Task 6.3: Optimization

- **WP 7:** Data processing and optimization
  - Task 7.1: Processing
  - Task 7.2: Optimization

- **WP 8:** Curation, Cataloguing and Provenance
  - Task 8.1: Curation
  - Task 8.2: Cataloguing
  - Task 8.3: Provenance

- **WP 9:** Validation & deployment
  - Task 9.1: Integration and validation
  - Task 9.2: Research to operation

**WP 5:**

- **Task 5.1:** Review of existing RIs
- **Task 5.2:** RI characterisation and ENVRIPLUS reference model
- **Task 5.3:** Semantic description and linking
- **Task 5.4:** Development plan and reference architecture

*Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)*
WP 5: Objectives

1. **Update the requirement analysis and technology review** performed in the ENVRI project to reflect the progress many RIs have made in their architecture design and system implementation.

2. **To promote interoperability across** RIs by providing a *novel reference model* (RM) that should be developed based on the existing ENVRI RM, with consideration of the latest development insights from successful RIs.

3. **To provide an ontological framework** based on the ENVRI RM to describe the requirements and system architecture of RIs, and to link these descriptions with the technologies provided by data and computing infrastructures.

4. **To provide a design and implementation plan and a reference architecture** for an abstract yet customisable research environment by engaging RI developers from different domains, in particular via work packages 6–9, and using the updated ENVRI RM.
## WP 5: Deliverable and deviations

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>D5.1</td>
<td>A consistent characterisation of existing and planned RIs</td>
<td>M12</td>
<td>Submitted in time</td>
</tr>
<tr>
<td>D5.2</td>
<td>A definition of ENVRIPLUS Reference Model</td>
<td>M18</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D5.4</td>
<td>A development plan for common operations and cross-cutting services</td>
<td>M18</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D5.5</td>
<td>A model architecture for new RIs to adopt and to act as a guide for existing RIs in their development</td>
<td>M24</td>
<td>Submitted in time</td>
</tr>
<tr>
<td>D5.3</td>
<td>A definition of the ENVRIPLUS Semantic linking framework at conceptual and formal levels</td>
<td>M36</td>
<td>On track</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS19</td>
<td>Reference model defined</td>
<td>M18</td>
<td>Online available in time, but D5.2 delayed</td>
</tr>
<tr>
<td>MS20</td>
<td>A development plan for common operations and cross-cutting services</td>
<td>M18</td>
<td>Delayed to M21, due to the delay of RM</td>
</tr>
</tbody>
</table>
WP 5: Overall Status

1. Successes
   a. Reference model, v2.1: envri.eu/rm
   b. OIL-e: v2.1 (http://www.oil-e.net/ontology/rm-archetypes.owl)
   c. Mapping between OIL-e and CERIF
   d. RM has been applied by ICOS, EISCAT-3D, EUFAR, LTER etc. (Highlight: reference model guided design)
   e. Publications and other achievements
      • Organized 1st IT4RIs, in eScience 2015: https://staff.fnwi.uva.nl/z.zhao/workshop/it4ris/
      • Two book chapters
      • 4 workshop paper in e-Science 15, AINA 16 and RTSS 16
      • Talks and special session in DI4R 16
      • Article in newsletter #2, blog article, and also T15.1 training materials outputs.

2. Issues and corrective actions
   a. Practical guidelines and tools of RM will be highlighted in the next phase

3. Plans for the next period
   a. Proceed with engineering/technology view of RM
   b. Build tools for RM guided design, semantic linking and engineering
WP 6: Inter RI Data Identification and Citation Services

WP 5: RM guided RI  
Task 5.1: review of existing RIs  
Task 5.2: RI characterisation and ENVRIPLUS reference model  
Task 5.3: semantic description and linking  
Task 5.4: development plan and reference architecture

WP 6: Identification & Citation  
Task 6.1: Data Identification/Citation

WP 7: Data processing and optimization  
Task 7.1: Processing  
Task 7.2: Optimization

WP 8: Curation, Cataloguing and Provenance  
Task 8.1: Curation  
Task 8.2: Cataloguing  
Task 8.3: Provenance

WP 9: Validation & deployment  
Task 9.1: integration and validation  
Task 9.2: research to operation

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
WP 6: **Objectives**

To **improve the efficiency of data identification and citation** by providing convenient, effective and interoperable identifier management and citation services.
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>D6.1</td>
<td>A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations</td>
<td>M20</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D6.2</td>
<td>A report on negotiations with publishers, providers of existing data citation systems and other scientific organisations on implementing a global data citation system.</td>
<td>M36</td>
<td>On track</td>
</tr>
<tr>
<td>D6.3</td>
<td>Use-case study reports: (a) an online, standards based publication mechanism for marine biological data; (b) workflow and guidance for tested citation tracking models</td>
<td>M46</td>
<td>On track</td>
</tr>
</tbody>
</table>
1. Success
   a. Publications:
      • Klump, J.et al., (2016). *20 Years of persistent identifiers – Which systems are here to stay?* Geophysical Research Abstracts
   b. Software:
      • The ICOS meta data service software repository(https://github.com/ICOS-Carbon-Portal/meta)

2. Issues and corrective actions
   a. Alignment with the reference architecture will be improved

3. Plans for the next period
   a. Further prototype identification and citation services
   b. Negotiation with publishers
   c. Use cases
WP 7: Objectives

- **Improving the efficiency of data processing** by providing system-level optimisation to select, provision and use interoperable infrastructure services for data integration, processing and storage.

- **Improving the performance of the research infrastructures** by providing system-level optimisation among data, processing, and underlying services, including their digital networks.
## WP 7: Deliverable and deviations

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>D7.1</td>
<td>Interoperable data processing for environmental RI projects: system design</td>
<td>M20</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D7.3</td>
<td>Performance optimisation for environmental RI projects: system design</td>
<td>M20</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D7.2</td>
<td>Interoperable data processing for environmental RI projects: system design: prototype</td>
<td>M42</td>
<td>On track</td>
</tr>
<tr>
<td>D7.3</td>
<td>Performance optimisation for environmental RI projects: prototype</td>
<td>M42</td>
<td>On track</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS25</td>
<td>System design for data processing and optimization services</td>
<td>M20</td>
<td>Delayed to M21, when D7.1 and D7.3 were submitted</td>
</tr>
</tbody>
</table>
WP 7: Overall Status

1. Success
   a. Publications
      • M. Assante, et al. (2016) Virtual Research Environments as-a-Service by gCube. 8th IWSG
      • Zhao, Z., et al. (2016) Time critical requirements and technical considerations, IT4RIs, RTSS 16
      • Zhou, H., et al.(2016) Inter-locale Virtual Cloud For Nearly Real-time Big Data Applications, IT4RIs, RTSS
   b. Organized 2nd IT4RIs in RTSS 2016: on nearly real time data processing
   c. Use cases
      • SC_3, TC_2, TC_3 and TC_13 (Highlight: data subscription service)
   c. Software services:
      • Data Miner and
      • Infrastructure optimizer

2. Issues and corrective actions
   a. Engage more RIs in the development loop

3. Plan for next step
   a. Refine software services based on existing use cases and new requirements
   b. Apply the services to new use cases
WP 8: **Data Curation and Cataloguing**

WP 5: RM guided RI

Task 5.1: review of existing RIs

Task 5.2: RI characterisation and ENVRIPLUS reference model

Task 5.3: semantic description and linking

Task 5.4: development plan and reference architecture

WP 6: Identification & Citation

Task 6.1: Identification

WP 7: Data processing and optimization

Task 7.1: Processing

Task 7.2: Optimization

WP 8: Curation, Cataloguing and Provenance

Task 8.1: Curation

Task 8.2: Cataloguing

Task 8.3: Provenance

WP 9: Validation & deployment

Task 9.1: integration and validation

Task 9.2: research to operation

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
WP 8: Objectives

- **Improving the efficiency and quality of user experience** for data curation by providing automatic and self-adaptable curation and semantic annotation services,
- **Improving catalogue interoperability** among different Research Infrastructures,
- **Improving the efficiency of data provenance and tracing** by providing convenient, effective and interoperable standards-based data provenance services.
### WP 8: Deliverable and deviation

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8.1</td>
<td>Data curation in system level of sciences: system design</td>
<td>M18</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D8.3</td>
<td>Interoperable cataloguing and harmonization: system design</td>
<td>M18</td>
<td>Delayed to M21</td>
</tr>
<tr>
<td>D8.2</td>
<td>Data curation in system level of sciences: prototype and deployment</td>
<td>M42</td>
<td>On track</td>
</tr>
<tr>
<td>D8.4</td>
<td>Interoperable cataloguing and harmonization: prototype and deployment</td>
<td>M42</td>
<td>On track</td>
</tr>
<tr>
<td>D8.5</td>
<td>Data provenance and tracing for Environmental sciences: system design</td>
<td>M36</td>
<td>On track</td>
</tr>
<tr>
<td>D8.6</td>
<td>Data provenance and tracing for Environmental sciences: prototype and deployment</td>
<td>M42</td>
<td>On track</td>
</tr>
<tr>
<td>MS22</td>
<td>System design for data curation and cataloguing</td>
<td>M18</td>
<td>Delayed to M21, when D8.1 and D8.3 were submitted</td>
</tr>
</tbody>
</table>
WP 8: Overall Status

1. Achievements
   a. Publications:
   b. Use cases:
      • TC_4, IC_2 and IC_8 (Highlight: flagship catalogue)

2. Issues and corrective actions
   a. Metadata mapping will be highlighted in the next phase jointly with T5.3

3. Plans for the next period
   a. Start provenance service
   b. Continue with the service prototype of catalogue and curation
   c. More use cases
WP 9: Service Validation And Deployment

WP 5: RM guided RI

Task 5.1: review of existing RIs
Task 5.2: RI characterisation and ENVRIPLUS reference model
Task 5.3: semantic description and linking
Task 5.4: development plan and reference architecture

WP 6: Identification & Citation

Task 6.1: Identification/Citation

WP 7: Data processing and optimization

Task 7.1: Processing
Task 7.2: Optimization

WP 8: Curation, Cataloguing and Provenance

Task 8.1: Curation
Task 8.2: Cataloguing
Task 8.3: Provenance

WP 9: Validation & deployment

Task 9.1: integration and validation
Task 9.2: research to operation

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
WP 9: Objectives

- **Improve the usability of developed services** in their intended environment. The results will be demonstrated, validated and refined via use cases.

- **Improve the deployment of the developed results**. Services and environment will be deployed in the research infrastructure, possibly via resources of the e-Infrastructure and data infrastructures.
## WP 9: Deliverables

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>D9.1</td>
<td>Service deployment in computing and internal e-Infrastructures Version1</td>
<td>M28</td>
<td>On track</td>
</tr>
<tr>
<td>D9.2</td>
<td>Serving key data service stakeholders and policy initiatives version 1</td>
<td>M30</td>
<td>On track</td>
</tr>
<tr>
<td>D9.3</td>
<td>Service deployment in computing and internal e-Infrastructures Version2</td>
<td>M40</td>
<td>On track</td>
</tr>
<tr>
<td>D9.4</td>
<td>Serving key data service stakeholders and policy initiatives version 2</td>
<td>M46</td>
<td>On track</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Deadline</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS29</td>
<td>Service deployment in computing and internal e-Infrastructures</td>
<td>M28</td>
<td>On track</td>
</tr>
</tbody>
</table>
WP 9: Overall Status

1. Success
   a. Use cases
      • E-Infrastructure support for 13 use case agile teams
   b. Deployment guidelines
      • https://confluence.egi.eu/display/EC/Guideline+for+Service+Deployment
   c. Service portfolio
      • https://confluence.egi.eu/display/EC/ENVRIplus+Service+Portfolios

2. Issues and corrective actions
   a. Training and support for e-infrastructure services will be enhanced

3. Plans for the next period
   a. Use case support, refine deployment guidelines and service portfolio
   b. Enhance the connection among communities of developers and e-Infrastructures
Outline

- Overview of the *data for science* theme
- Activities and achievements
- **Summary**
SUMMARY 1: **ENVRI RM and development plan**

- **D5.1** requirements, technology review, gap analysis
- **RM 2.x** Envri.eu/rm
- **OIL-E 2.1** Semantic linking framework
  - http://www.oil-e.net/ontology/rm-core.owl
- **D5.2** Reference model
- **D5.4** Development plan
- Use cases
- Interview notes
SUMMARY 2: Reference model guided design

- **D5.1**: Requirements, technology review, gap analysis
- **D5.2 RM**: Use cases
- **D5.4 Development plan**: Input from agile use case teams
- **D6.1**: Identification and citation service architecture and recommendations
- **D7.1**: Data processing service architecture and recommendations
- **D7.3**: Optimization service architecture and recommendations
- **D8.1**: Curation service architecture and recommendations
- **D8.3**: Cataloguing service architecture and recommendations

Project Number: 654182
SUMMARY 3: Reference architecture

D5.1: Requirements, technology review, gap analysis

D6.1: Identification and citation service architecture and recommendations

D7.1: Data processing service architecture and recommendations

D7.3: Optimization service architecture and recommendations

D8.1: Curation service architecture and recommendations

D8.3: Cataloguing service architecture and recommendations

Use cases

D5.2 Reference model

D5.4 Development plan

D5.5 Reference architecture
SUMMARY 4: RI engagement

D5.1: Requirements, technology review, gap analysis

D5.2, D5.4, and D5.5: Reference model, development plan and reference architecture

Design of Citation/identification, processing, optimization, curation and cataloguing (D6.1, D7.1, D7.3, D8.1 and D8.3)

Deployment and validation on E-infrastructure
SUMMARY 5: meet WP 5 objectives

Objectives

1. Update the requirements analysis and technology review

2. To promote interoperability across RIs by providing a novel reference model (RM)

3. To provide an ontological framework

4. To provide a design and implementation plan and a reference

D5.1 Executive summary
RM 2.x Envri.eu/rm
OIL-E 2.1
D5.4 D5.5
SUMMARY 6: next steps in WP 5

Objectives

1. Update the requirements analysis and technology review
   - D5.1
   - Executive summary

2. To promote interoperability across RIs by providing a novel reference model (RM)
   - RM 2.x
   - Envri.eu/rm
   - D5.2

3. To provide an ontological framework
   - OIL-E 2.1

4. To provide a design and implementation plan and a reference
   - D5.4
   - D5.5

Next step

1. Refinement and revisiting in the context of WP6-9

2. Continue with engineering/technology viewpoints

3. Annotation, search and other tools

4. Refinement in the context of WP6-8
SUMMARY 7: meet WP 6-8 objectives

Objectives

1. (WP6) to improve the efficiency of data identification and citation

2. (WP7 -1) to improve the efficiency of data processing

3. (WP7 -2) to improve the performance of the research infrastructures

4. (WP8 -1) to improve the efficiency and quality of user experience for data curation

5. (WP8 -2) to improve catalogue interoperability among different research infrastructures

6. (WP8 -3) to improve the efficiency of data provenance and tracing
SUMMARY 8: next steps in WP 6-8

Objectives

1. **(WP6)** to improve the efficiency of data identification and citation
2. **(WP7 -1)** to improve the efficiency of data processing
3. **(WP7 -2)** to improve the performance of the research infrastructures
4. **(WP8 -1)** to improve the efficiency and quality of user experience for data curation
5. **(WP8 -2)** to improve catalogue interoperability among different research infrastructures
6. **(WP8 -3)** to improve the efficiency of data provenance and tracing

Next step

1. Refine services based on roadmap and recommended architecture (defined in D5.3, D5.4)
2. Validate the prototypes in more use cases
3. Start provenance tasks
SUMMARY 9: meet WP 9 objectives

1. Improve the usability of developed services in their intended environment.

2. Improve the deployment of the developed results.

Use cases

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Theme</th>
<th>WP 9</th>
<th>WP 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC 1</td>
<td>WP 9</td>
<td>WP 8</td>
<td>WP 11</td>
</tr>
<tr>
<td>TC 2</td>
<td>WP 9</td>
<td>WP 10</td>
<td>WP 12</td>
</tr>
<tr>
<td>TC 3</td>
<td>WP 9</td>
<td>WP 10</td>
<td>WP 12</td>
</tr>
<tr>
<td>TC 4</td>
<td>WP 9</td>
<td>WP 10</td>
<td>WP 12</td>
</tr>
</tbody>
</table>

Deployment guidelines & Service portfolio

ENVRIplus Service Portfolios

New Service
To add a new service, click the button below.

Overview

<table>
<thead>
<tr>
<th>Service Name</th>
<th>Service Area</th>
<th>Service Phase</th>
<th>Topline Description</th>
<th>Service Vendor</th>
<th>Action Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Real-time Infrastructure Planner (DRIP)</td>
<td>compute</td>
<td>alpha</td>
<td>Optimised infrastructure planning and processing for time-critical applications.</td>
<td>UNIV (University of Amsterdam)</td>
<td></td>
</tr>
<tr>
<td>CEUR/ERB</td>
<td>Data processing and analytics</td>
<td>Production</td>
<td>Open, user-friendly and extensible data analytics platform ready for Open Science and VREs.</td>
<td>CEUR/ERB</td>
<td></td>
</tr>
</tbody>
</table>
1. Improve the usability of developed services in their intended environment.

2. Improve the deployment of the developed results.

Next steps

1. More demonstrators of use cases
2. More concrete services in portfolio
3. More deployments on e-infrastructures
HIGHLIGHTS IN THEME 2
Highlights

Common vocabulary:
- Reference model (T5.2)

Processing:
- (T7.1)

Curation:
- (T8.1)

Optimization:
- (T7.2)

Provenance:
- (T8.3)

Meta information linking:
- Linking model (T5.3)

RI development:
- Architecture design (T5.4)

Deployment and operation:
- (T9.1)

Validation:

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
Highlights

Common vocabulary: Reference model (T5.2)

Identification/Citation (T6.1)

Processing (T7.1)

Curation (T8.1)

Optimization (T7.2)

Provenance (T8.3)

Meta information linking: Linking model (T5.3)

RI development: Architecture design (T5.4)

Deployment and operation

Validation

Cataloguing (T8.2)

Storage, computing, networking and other technologies provided by underlying e-Infrastructures (EGI, EUDAT, etc.)
THEME 2 HIGHLIGHT 1
Reference model guided approach

FROM WP5
T5.2, Cardiff University (Alex Hardisty)
T5.3, University of Amsterdam (Paul Martin, Zhiming Zhao)
Abraham Nieva, Malcolm Atkinson, Aurora Constantin, Barbara Magagna, Markus Stocker
The ENVRI Reference Model (RM), started in the former ENVRI project, is an ongoing development of ENVRIplus that can be applied to the informatics engineering design challenges faced by RIs.

Like the design of buildings, the RM decomposes the modelling procedure for a complex distributed system into different viewpoints.

The RM promotes a standard vocabulary for describing environmental research infrastructures and for modelling their components and architecture from the scientific, informational, computational, engineering and technology viewpoints.

Current version: V2.1

http://envri.eu/rm
WHO IS USING THE ENVRI RM?
PRESENTLY APPLIED TO DESIGN WORK

By RIs, assisting research infrastructure engineers with systems design:

- **DASSH**—archive for marine species and habitats data; deliver metadata and access capabilities into partner infrastructures (MEDIN, NBN, EMBRC, etc.).
- **EISCAT-3D**—designing and investigating big data strategies via outsourcing to e-Infrastructure providers such as EGI.eu and EUDAT.
- **EMSO**—Data Management Architecture Design; EMSOdev D6.2, 08/16.
- **EUFAR**—identifying/understanding missing business processes.
- **ICOS**—consolidation of thinking regarding different design sub-groups.
- **LTER, Euro-Argo, SeaDataCloud and EPOS**—recently started.

*Having a “personal trainer” helped a lot to start working with the RM. Gaps within EUFAR’s working procedures could already be identified.*

*EISCAT is developing a new portal within EGI and EUDAT projects. The RM helps to define the structure and content of the new data model the portal is dependent on.*

*The RM represents an incredibly useful tool to help crystallise one’s thinking, once an initial familiarity has been achieved.*

H2020 Project

Project Number: 654182
**EUFAR** brings together operators of instrumented research aircraft and remote-sensing instruments with the scientific user community.

Using RM to model their **processes**:

- **Process** of applying for Trans-National Access to aircraft.
- **Process** of managing a complete campaign to collect data—from initial contact with the aircraft operator, through flight planning and action, instrument installation and test, data acquisition and checking/logging to final reporting on success and costs of the campaign.
**ORCID integration in RIs** enables automated linking of submitted and curated datasets to contributors’ ORCID identifiers.

As demonstrated by Argo, RIs can credit contributors by automatically updating their ORCID record on their behalf whenever new data is published.

The ENVRI Reference Model **guides ORCID integration** in RIs by:

- specifying ORCID's role in relation to RIs.
- suggesting computational components required for RIs to integrate with ORCID.
- Showing that the ORCID iD is an information object manipulated by RIs.
WHY IS THE RM IMPORTANT?

FOR KNOWLEDGE SHARING

- **Discover** reusable components among research infrastructures.
- **Design** new research infrastructures.
- **Optimise** the evolutionary path.

**WHY IS THE RM IMPORTANT FOR KNOWLEDGE SHARING?**

- Discover reusable components among research infrastructures.
- Design new research infrastructures.
- Optimise the evolutionary path.

**Semantic description and tools (T5.3)**

- RI: how did other RIs implement my missing functionality?
- RI: how should I upgrade my services?
- New RI: What are the best practices for meeting my requirements?

**INTEGRATION OF EXISTING INFRASTRUCTURES**

- ICOS
- LTER
- Euro-Argo
- DASSH
- EUFAR

**NEW RIS**

- RI knowledge base

**H2020 Project**

Project Number: 654182
RIs will last much longer than ENVRIPLUS. ICT provision transforms radically over time. Technologies come and go but logicalities are more stable.

Those responsible for RIs should know what the important RI elements are and where and what are the interfaces to them—so that interoperability can be achieved, and so that third-party components/services can be widely exploited.

e.g., in the cataloguing highlight that you will hear about next.

The goal is not to have all RI functions implemented in the same or similar ways (although there are ‘critical mass’ benefits arising from that as well) but to standardise external and inter-RI interfaces to simplify the task of using or building tools for multiple RIs.

RM provides the framework within which medium to long-term ICT options can be considered and proposed, for example by a strategic “ENV RI Information Systems Strategy and Engineering Group”.

H2020 Project

Project Number: 654182
THEME 2 HIGHLIGHT 2

Flagship product catalogue

FROM WP 8

Thomas Loubrieu, Keith Jeffery

Euro-ARGO, SeaDataNet (IFREMER), EPOS (NERC, INGV, GFZ), ICOS (LU), IAGOS (CNRS), LTER (UBA-GmbH), ANAEE (INRA), EMBRC (MBA)
Any resources provided by an RI that are open to users and should be discoverable, easy understandable, advertised… think of it as a marketing tool for RIs.

**Data services:**
- Global Ocean physics properties from 0 to 2000 meters deep by Euro-ARGO profiling floats (ARGO).
- Atmosphere Carbon concentrations measured by IAGOS planes.
- …

**Acquisition services:**
- Ecological Research Station in Tropical forest, French Guyana (ANAE).
- Marine observatory of mid-ocean ridge processes, from the subsea floor to the water column (EMSO).
- …

Not always trivial to synthesize: some RIs provide a collection of heterogeneous datasets.
FLAGSHIP PRODUCT CATALOGUE

The main challenge: get **sustainable involvement and support from RIs** for the catalogue activity at transdisciplinary level.

**Strategy:**
- Benefit from the **transdisciplinary** nature of ENVRIplus RIs.
- **Get RIs involved** by offering them advantage **by advertising their outcomes**.
- Focus on an immediate, low risk development.

**Use Case:** a **flagship data and acquisition services catalogue** to expose products of RI beyond their traditional communities:
- The products from different RIs should be presented as homogeneously, intuitively and with good looking descriptions.
- The products from different RIs will be balanced.

**Success will be measured by:**
- How users can actually retrieve resources in the catalogue, of course, but also…
- The satisfaction of RIs regarding how they are represented in the catalogue.
THE DEMONSTRATOR WILL HAVE:

- Products from:
  - Euro-ARGO, SeaDataNet (IFREMER)
  - EPOS (NERC, INGV, GFZ)
  - ICOS (LU)
  - IAGOS (CNRS)
  - LTER (UBA-GmbH)
  - ANAEE (INRA)
  - EMBRC (MBA)

- Show them in EUDAT/B2FIND infrastructure implemented by DKRZ with CKAN software (up and running system, flexible, open-source and popular, quick win, and an RI-neutral service).
H2020 Project

Project Number: 654182

ARCHITECTURE

RI components

Flagship product Catalogue (EUDAT)

ISO19115-2, DC, DCAT

https://ckan.org/

Conceptual Canonical Metadata Scheme

Conversion to the canonical metadata superset recommendation.
ACHIEVEMENTS

 envriplus
 ENVRiplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist... read more

 13 datasets found

North Atlantic Ocean - Temperature and salinity observation collection V1.1
SeeDataNet Temperature and Salinity historical data collection, including revised quality flags after quality control with ODV. For data access please register at...

North Atlantic Ocean - Temperature and salinity observation collection V1.1
SeeDataNet Temperature and Salinity historical data collection, including revised quality flags after quality control with ODV. For data access please register at...

Mediterranean Sea - Temperature and Salinity Climatology V1.1
Mediterranean Sea Climatology computed from the SeeDataNet V1.1 aggregated dataset. The version used for the DNA software is the 4.6.9. The period covers 1500-2013. For data...

c02 bliscrosee

c02 measurement at B8 station

EUDAT D7.1 EUDAT/EGI Pilot Activity
This document describes the work undertaken by the EUDAT2020 project to strengthen the interoperability between EUDAT and EGI, and to prepare the ground for establishing a...
ACHIEVEMENTS

H2020 Project

ACHIEVEMENTS

North Atlantic Ocean - Temperature and salinity observation collection V1.1

SeaDataNet Temperature and Salinity historical data collection, including revised quality flags after quality control with ODV.

For data access please register at [http://www.marine-id.org](http://www.marine-id.org)

The dataset format is ODV binary collections. You can read, analyse and export from the ODV application provided by Alfred Wegener institute at [http://odv.awi.de](http://odv.awi.de).

Additional Info

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline</td>
<td>Not stated</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Publication Year</td>
<td>2014</td>
</tr>
<tr>
<td>SpatialCoverage</td>
<td>(10°-60°, 65°-10°)</td>
</tr>
</tbody>
</table>
OUTCOMES

- A technical platform with a collection of RI metadata in B2FIND, where both data managers / IT people in RI and ITC work together to share descriptions of data and acquisition services.

- Fundamental questions on how the RI products should be made visible at a trandisciplinary level are raised:
  - Which user questions the RI answers?
  - How to preserve visibility of PIs?
NEXT STEPS

- **EUDAT/B2FIND actions:**
  - Automate metadata collection from RI (harvesting in OAI-PMH and CSW protocols).
  - add attributes: quicklooks, related documents (bibliography, user manual, etc.).

- **Refine metadata profile mapping for all demonstration RIs:**
  - Syntax, semantics: detailed format, agreement on shared keywords (e.g. RI labels and data service vs acquisition service).
  - Measure metadata quality.
  - Refine metadata granularity for all demonstration RIs (balance outcomes of each).

- **Link** data services in flagship product catalogue with datasets.
THEME 2 HIGHLIGHT 3
Data subscription service

FROM WP 7 & 9
Thierry Carval, Baptiste Grenier, Jani Heikkinen,
Glenn Judeau, Spiros Koulouzis
Broadcast Argo RI observations via EU E-infrastructures

- RIs products such as the Argo dataset are registered in the ENVRIPLUS catalogue.

- In the catalogue, an inter-thematic data subscription service derived from RIs’ data is proposed, starting with Argo dataset, then with additional RIs (such as EMSO and ICOS).

- The data subscription use case is implemented under the guidance of the ENVRI Reference Model:
  - Data curation: managing continuous updates
  - Data citation scheme
  - Accounting of data distribution
Broadcast Argo observations over EU E-infrastructures
A data subscription use case

To provide regularly specific data to scientists

The user provides his criteria:
- time, spatial area, parameters, data types, update period for delivery (e.g. daily, monthly, etc.).
- Data are extracted from E-infrastructures and formatted on the cloud.
- The user’s cloud account is updated regularly with new data.
- An accounting of data delivery is performed:
  - A citation scheme is attached to the delivered data (ID):
    - bibliographic surveys can track the use of these data in publications
    - reproducibility is possible
  - A user’s identification scheme could be implemented (e.g. Marine-ID, Shibboleth, OpenID, etc.).
USE CASE: DATA SUBSCRIPTION

notify

new data

DSS

B2SAFE

Push data

data access

register

GDAC

ARGO Web UI

DRIP Manager

DRIP Planner

DRIP Provisioner

DRIP Deployer

initiate action

plan

provision

deploy

execute

 instantiate

Worker VM
Generalization with other Research Infrastructures

Feeds VREs for trans-disciplinary science: ocean, space, atmosphere, earth, etc.

- Environmental monitoring and forecasting: EU ocean-atmosphere models
- Calibration and validation with in-situ data: SMOS, Sentinel satellite missions, etc.
- Efficient workplace for advanced services
END OF THE THEME 2 REPORT
THE FOLLOWING SLIDES ARE BACKUP
WP 5: REFERENCE MODEL GUIDED RI DESIGN

WP LEADER: DR. PAOLA GROSSO (UVA)

DURATION: M1-M36

PARTNERS: CU, CNR_SIOS, CNRS_IAGOS, IS-ENES, FIXOS, EAA_LTER, EGI.EU, EISCAT, EMBL_ELIXIER, ETHZ_EPOS, EURO-ARGO, FZJ_IAGOS, IFREMER_JERICO, INGV_EPOS, INRA_ANAEE, LU_ICOS, MBA_EMBRC, NERC_EPOS, UNIHIB_EMSO, USTAN_EMBRC, UCPH_INTERACT, UEDIN, UHEL_ICOS, UVA,

EFFORT: 187PM

Supporting environmental research with integrated solutions - the Earth is our lab
Task 5.1: Requirement Analysis, Technology review and Gap analysis

**Leader:** UEDIN (prof. dr. M. Atkinson)

**Duration:** M1 - M12

**Achievements**
- D5.1 of requirement analysis, technology review and gap analysis.
- An executive summary of the report has been composed (led by Alex V.)
- The need to achieve data harmonisation,
- The need to learn from one another and pool efforts
- Sustainably delivering data services immediately to meet current RI priorities while taking into account longer-term issues and technology trends.
**Task 5.2: Reference model for ENVRI Plus**

- **Leader:** CU (mr. A. Hardisty & dr. A. Nieva)
- **Duration:** M7-M30
- Refine the ENVRI RM based on updated requirements
- **EV and TV design** in progress
- **Key achievements**
  - **M15:** ENVRI RM V2
  - **M18:** ENVRI RM V2.1
  - **D 5.2**
  - Aligned D6.1, 7.1, 7.3, 8.1 and 8.3
  - Driving specification of common and cross-cutting services in D5.5 (Ident & Cit, Cataloguing, etc.)

**Align the design of the Theme2 services**

**M24**

**Engineering/technology viewpoints**

**With RM views:**

- RM 1
- RM 2
- RM 2.1
- RM 2.x

**Driving requirements:**

- T5.1 activities
- TMF 16
- IC_12
- RM 1
- RM 2
- RM 2.1
- RM 2.x
- Theme2 tasks
- Data fabric interest group
Task 5.3: Semantic description and linking

- **Leader:** UvA (dr. P. Martin and dr. Z. Zhao)
- **Duration:** M7-M36
  - Open information linking for environmental science (OIL-e)
  - Align OIL-e with ENVRI RM
  - Mapping OIL-e with other metadata models, e.g., CERIF
  - Describe more RIs via f2f meetings
- **Key achievements:**
  - Mapped OIL-e (semantic linking ontology) with CERIF.
  - Agile team IC_11
**Task 5.4: Interoperation based architecture**

- **Leader:** NERC (prof. dr. K. Jeffery)
- **Duration:** M7-M36
  - Design an architecture which meets the requirements from T5.1, takes account of the reference elements in T5.2 (i.e. which satisfies the RIs);
  - Design an architecture that is effective, efficient and sustainable;
  - Design an architecture based on best practice in ICT for research;
  - Design an architecture that is evolvable based on changing RI requirements and changing ICT opportunities;
- **Key achievements:**
  - M18, D 5.4: theme2 roadmap
  - M24, D 5.5: reference architecture

![Diagram](image-url)
WP 6: INTER RI DATA IDENTIFICATION AND CITATION SERVICES

WP LEADER: ALEX VERMEULEN (LU)

DURATION: M1-M48

PARTNERS: CINECA_EUDAT, CNR_SIOS, CNRS_IS-ENES, DKRZ_IS-ENES,
EAA_LTER, EISCAT, INRA_ANAEE, LU_icos, MBA_EMBRC,
IFREMER_SEADATANET, NILU_ACTRIS, UHB_EMSO,

EFFORT: 65PM
**Task 6.1: Identification and Citation**

- **Leader:** LU (dr. A. Vermeulen and dr. M. Hellström)
- **Duration:** M1-M48

- Data uptake tool for ICOS: data PID’ing, store at trusted repository, enrich metadata in versionable metadata store, dynamic landing pages
- Libraries, publishers on data citation systems
- Publishing subsets, collections with adequate citation based on PID’ed DO
- Progress in ICOS data lifecycle & GBIF marine biological data
- Participation and contributing to relevant RDA working groups and activities:
  - Data Citation
  - Data Type Registries
  - Research Data Collections

**M7**

- ICOS
- DASSH

**M20**

- WP 6 services, tools, and recommendations.

**M24**

- Case study
WP 7: DATA PROCESSING AND ANALYSIS

WP LEADER: DR. LEONARDO CANDELA (CNR)
DURATION: M7-M42
PARTNERS: CINECA, EUDAT, CNR, CSC, DKRZ, IS-ENES, EISCAT, ETHZ, EPOS, EURO-ARGO, INGV, EMSO, UNIHB, UNITUS, ICOS, UVA,
EFFORT: 92PM

Supporting environmental research with integrated solutions
- the Earth is our lab
Task 7.1 Data processing

- Leader: CNR (dr. L. Candela)
- Duration: M7-M42
  - Investigated the feasibility of using D4Science in a number of agile case studies: SC_3, TC_2, IC_3, IC_13
  - Re-engineered the entire system and released DataMiner framework
  - Performance enhancement
  - Reinforced algorithm importing facility (input typologies enlarged)

Key achievements:
- D7.1
- Publications
- Software deployment
- Use case prototype

Data processing platform
- Data miner
  - Processing technologies: gcube workflow, etc.

Case study
- M24 study
- H2020 Project
- ENVIRO

T5.1 activities
- D4Science
- D7.1
- RM
- Architecture options
Task 7.2 Performance optimization

- **Leader:** UvA (dr. P. Martin and dr. Z. Zhao)
- **Duration:** M7-M42
- Investigated the feasibility of using D4Science in a number of agile case studies: SC_3, TC_2, IC_3, IC_13
- QoS based virtual infrastructure planning and provisioning (experiments on EGI FedCloud)
- RI provisioning of assets: datasets but also software, access to resources, persons...
- End-user discovery/contextualization/action: workflow construction including datasets and software and access to resources

**Key achievements:**
- Software prototype
- Use case
- Publications

**Technology review**
- Optimization for data transfer, processing, virtual infrastructure
- Architecture options
- Service prototype deployment
- Case study

**M7**
- **T5.1 activities**

**M18**
- **D 7.3**

**H2020 Project**
- **Project Number:** 654182
WP 8: DATA CURATION AND CATALOGUING

WP LEADER: PROF. DR. KEITH JEFFERY (NERC)
DURATION: M7-M42
PARTNERS: CEA, ICOS, CINECA, EUDAT, CNR, EXP, SIOS, CNRS, IAGOS, CSC, EUDAT, DKRZ, EPOS, EAA, LTER, EISCAT, IFREMER, SEADATANET, EUROFLEET, INGV, EPOS, EMSO, INRA, ANAEE, LU, IOCS, MBA, EMBRC, NERC, EPOS, NILU, ACTRIS, USTAN, EMBRC, UVA
EFFORT: 159PM
**Task 8.1: Curation**

- **NERC** (prof. dr. K. Jeffery)
- **Duration:** M7-M42
  - Process of curation related to data collection and quality control (includes annotation)
  - Discovery, contextualization, action using curated datasets
  - Data Management Plan (DCC)
  - Data Lifecycle (DCC)
  - OAIS (architectural recommendation – very flexible)

**Key achievements:**
- **D8.1**
- T5.1 activities
- OIL-E 1
- RM 1
- RM 2.x
- Relevant wp/tasks (6,7,8)
- Curation recommendations
- Air mass flagging case
- DMPonline
- IC_8
- TC_2
- Case study
- Service prototype deployment
- H2020 Project
- Project Number: 654182
Task 8.2: Cataloguing

Leader: IFREMER (dr. T. Loubrieu)

Duration: M7-M42

Initial agreement on: 1) A superset of attributes to describe the products (RDA, metadata), and 2) A set of support standards to describe products: ISO 19119, 19139, DC and DCAT.

Collection of inputs from the RI catalogues

Integration in EUDAT/B2FIND catalogue system.

CKAN and CERIF are recommended

Key achievements:

D8.3 Flagship catalogue demonstrator
Task 8.3: **provenance**

- **Leader:** EAA (dr. B. Magagna)
- **Duration:** M25-M42
  - Join the requirement analysis (5.1) and RM review (5.2)
  - Contributed to the **D5.1 and D5.2**
- **Key achievements:**
  - An internal discussion wiki
WP 9: SERVICE VALIDATION AND DEPLOYMENT

WP LEADER: DR. YANNICK LEGRE (EGI.EU)

DURATION: M13-M48

PARTNERS: CEA ICOS, CINECA EUDAT, CNR ACTRIS, CNRS IAGOS, CSC EUDAT, EAA LTER, EGI.EU, EISCAT, EMSC EPOS, ETHZ EPOS, EURO-ARGO, FZJ IAGOS, IFREMER JERICO, INGV EPOS, EMSO, INRA ANAEE, LU IOCS, NERC EPOS, NILU ACTRIS, UCPH INTERACT, UHEL ICOS, UITE SONET-VI, UNIHB EMSO, UNITUS ICOS, USTAN EMBRC

EFFORT: 118PM
Task 9.1: Validation and integration

- **Leader:** EGI.eu (Y. Legre & Y. Chen)
- **Duration:** M13-M48
- **Case study teams**
- **Deployment technologies in** SC_3 (Mosquito), TC_2 (Euro-Argo), TC_13 (ICOS), IC_3 (EISCAT-3D)
- **Reviewed a number of technologies**
- **Key achievements:**
  - Deployment guideline
  - Initial service portfolio of theme2
Task 9.2: From research to operation

- **Leader:** LU (A. Vermeulen)
- **Duration:** M13-M48
  - Track the usability of the results of 9.1, promote the integration between RI and initiatives: COPERNICUS, EEA, etc.
  - Review the designed mechanisms with the stakeholder together
  - Provide operation support for RIs
- **Key achievements:**
  - Depend on 9.1 and other tasks,
  - More to be expected in coming phases