The Data for Science theme: software and solutions to address common challenges facing environmental research infrastructures

Dr. Zhiming Zhao leads the Data for Science theme (Theme 2). He is currently a senior researcher in University of Amsterdam. He led the scientific linking model research in the FP7 ENVRI project, and is currently also the scientific coordinator of the H2020 project SWITCH.

Theme 2 will establish an information technology (ICT) approach for handling the lifecycle of scientific data. This approach will inspire interoperability and common solutions that can benefit research infrastructures (RI) and facilitate interdisciplinary research.

A sustainable natural environment is essential for the future of human society. However, challenges such as climate change, natural disasters, pollution, and loss of biodiversity threaten the natural balance of our world. To understand these challenges and more importantly their impact on our local environment, scientists not only need to measure and observe the environment on a large scale over a prolonged period of time, but also need to understand the interactions between different environmental systems that exist on multiple levels. However, the complexity of these environmental systems, and the diversity of the scientific domains in which they are studied make this task extremely difficult. This despite ICTs having already been extensively harnessed to support research within all the different sub-domains of the environmental and earth sciences.

Since historically these RIs have been constructed to address specific environmental domains, gaps between infrastructures exist that make those scientific activities that require data and software distributed across different RIs particularly difficult. Recently, infrastructures have been strongly encouraged to support interdisciplinary research and to contribute to global cross-domain initiatives such as societal challenges, Copernicus and the Global Earth Observation System of Systems (GEOSS). Moreover, many research infrastructures face the same challenges for managing data and for supporting the research activities of scientists throughout the experiment lifecycle. Sharing solutions to those common problems will not only reduce development costs but also promote interoperable solutions among different infrastructures.

Many research infrastructures face the same challenges for managing data and for supporting the research activities of scientists throughout the experiment lifecycle.

Typical common problems include: 1) how to identify and cite data from different sites or infrastructures; 2) how to control the quality of nearly real-time data from sensors and annotate them; 3) how to catalogue the data and to allow users to search and access data from different sites or infrastructures; 4) how to support scientists to perform experiments using data, software tools and resources from different remote infrastructures; 5) how to effectively manage the infrastructure resources in the scientific experiments and allow scientists to achieve their goals more quickly; and 6) how to effectively record the events and results generated during experiments so that scientists can reproduce them independently.

The *Data for Science* theme in the ENVRIplus project has been proposed to break the barriers between RIs and to provide shared solutions for many common challenges and issues. The challenges that the Data for Science theme face include: 1) different RIs prioritise problems differently in their own development agenda; 2) RIs do not share a common view on the architecture and constituent components; and 3) there are few standards in consistent and pervasive use.

To tackle those issues, a common reference model, which can help vocabulary and concepts converge between different RIs and user communities, plays an extremely important role. A reference model guided engineering approach has been proposed for the Data for Science theme, as shown in Fig 1.

A reference model for providing a common understanding for data solutions and components within an RI will be provided based on a requirement analysis and review of existing technologies and RI projects. This will be used to guide the development of new solutions for common problems facing all environmental science RIs. Finally, the developed software will be validated and disseminated to the final user communities.

The Data for Science theme in the ENVRIplus project has been proposed to break the barriers between RIs and to provide shared solutions for many common challenges and issues.

The "Data for Science" theme has a duration of four years. In the first six months, the main activities focus on requirement analysis, technology review and gap identification. Then, the Environmental Research Infrastructure Reference Model (ENVRI-RM) and the semantic linking framework developed in the previous ENVRI project will be refined and extended based on the evolving requirements and developed facilities of the RIs, and the evolving ICT capabilities and services available. The Data for Science team is very pleased with the demonstrated involvement of specialists from all RIs and e-Infrastructures such as EUDAT (European Data Infrastructure) and EGI (European Grid Infrastructure) in efforts to refine and execute the development plan.

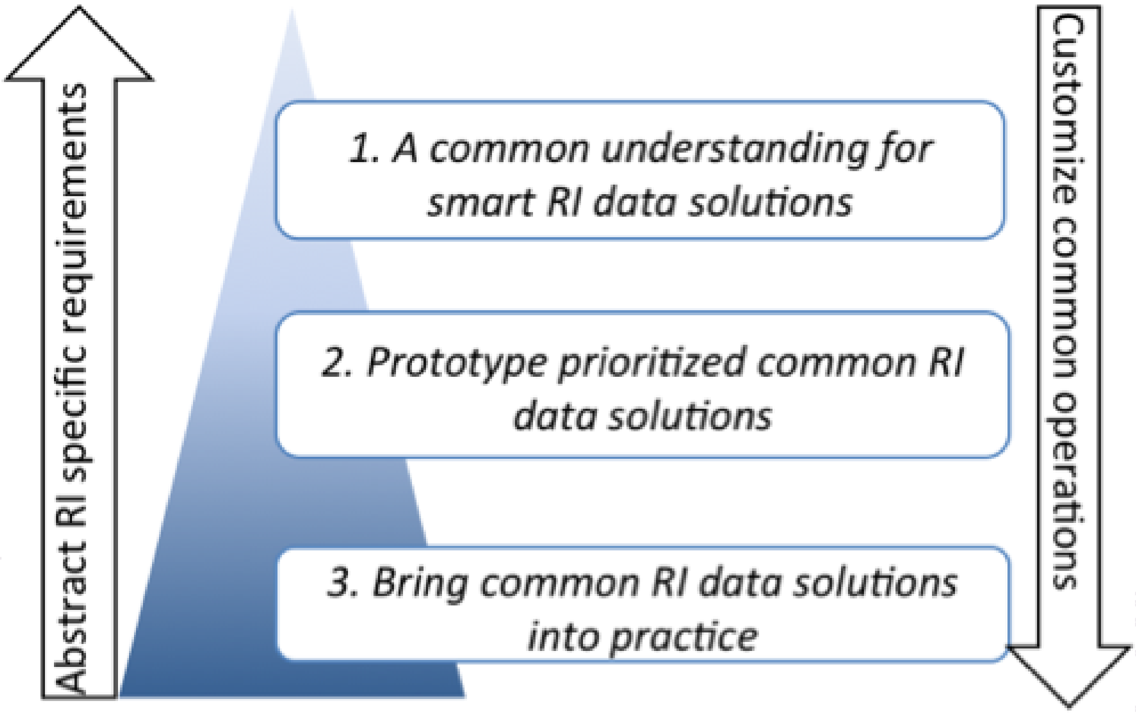
Abstract RI specific requirements
1. A common understanding for smart RI data solutions
2. Prototype prioritized common RI data solutions
3. Bring common RI data solutions into practice
Customize common operations

Figure 1—The basic idea of the reference model guided approach.