# Seeing the forest through the trees: A review of integrated environmental modelling tools

Carlos Granell (European Commission, Joint Research Centre, Ispra, Italy)

Sven Schade (European Commission, Joint Research Centre, Ispra, Italy)

Nicole Ostländer (European Commission, Joint Research Centre, Ispra, Italy)

**Abstract**

Today's interconnected socio-economic and environmental challenges require the combination and reuse of existing integrated modelling solutions. This paper contributes to this overall research area, by reviewing a wide range of currently available frameworks, systems and emerging technologies for integrated modelling in the environmental sciences. Based on a systematic review of the literature, we group related studies and papers into viewpoints and elaborate on shared and diverging characteristics. Our analysis shows that component-based modelling frameworks and scientific workflow systems have been traditionally used for solving technical integration challenges, but ultimately, the appropriate framework or system strongly depends on the particular environmental phenomenon under investigation. The study also shows that - in general - individual integrated modelling solutions do not benefit from components and models that are provided by others. It is this island (or silo) situation, which results in low levels of model reuse for multi-disciplinary settings. This seems mainly due to the fact that the field as such is highly complex and diverse. A unique integrated modelling solution, which is capable of dealing with any environmental scenario, seems to be unaffordable because of the great variety of data formats, models, environmental phenomena, stakeholder networks, user perspectives and social aspects. Nevertheless, we conclude that the combination of modelling tools, which address complementary viewpoints - such as service-based combined with scientific workflow systems, or resource-modelling on top of virtual research environments - could lead to sustainable information systems, which would advance model sharing, reuse and integration. Next steps for improving this form of multi-disciplinary interoperability are sketched.

**Keywords:** Integrated modelling; environmental modelling; reusability; sustainability

This paper can be cited as:

C. Granell, S. Schade, N. Ostländer. **Seeing the forest through the trees: A review of integrated environmental modelling tools**. *Computers, Environment and Urban Systems*, 41: 136-150, 2013, ISSN 0198-9715.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.06.001

# 1 Introduction

The super storm Sandy devastated the east coast of United States and Canada in November 2012, causing almost 100 fatalities and 50 billion US Dollars in economic damage[1]. It is just one recent event from an alarming list[2] of natural hazards, such as earthquakes, strong storms, hurricanes, and tsunamis that cause serious damages and injuries. Such phenomena are the result of complex interrelations between natural systems and accumulated interventions made by human beings. Understanding the interrelation among natural and socio-economic systems and anticipating their impacts is one of the main drivers for environmental sciences.

Modelling is a powerful tool to understand the Earth and our environment. A model mimics and simplifies a natural system (e.g. climate, ecosystems, watershed, and atmosphere) or a part of it (incl. air pollution and soil erosion). The analysis of complex natural phenomena, such as the examples above, requires the combination of several models that may even span over multiple disciplines (biodiversity, oceans, agriculture, etc.). The notion of *Integrated Modelling* (IM[3]) addresses exactly this situation. Rotmans and van Asselt (2001) defined IM as the process to structure and sharing new knowledge that emerges from the interrelation of different constituent models in complex socio-economic-environmental scenarios. One of the objectives in IM is to support simulation, anticipating potential impacts, policy and decision-making processes (McIntosh et al., 2008), since it is assumed that an individual model cannot be sufficient to represent the complexity, parameters and variables needed in multi-dimensional scenarios.

In geospatial and environmental sciences, IM[4] is widely recognized as a project-oriented activity to explore a given problem or scenario, i.e., the combination of models depends eventually on the phenomenon being investigated (Parker et al., 2002; Jakeman & Letcher, 2003). Voinov and Shugart (2013) describe "integral modelling" as a modelling strategy that assembles ad-hoc models for a particular scenario without considering the sharing and reuse of its contained models. Still, reuse, sharing, replication and reproducibility are driving principles in science (Mesirov, 2010; Jasny, Chin, Chong, & Vignieri, 2011) to promote new knowledge and ensure scientific advancements. These principles are equally desirable in environmental modelling activities. Accordingly, Voinov and Shugart (2013) suggest "integrated modelling" as the modelling strategy that assimilates models which are designed to be shared and reused.

---

[1] http://www.the-scientist.com/?articles.view/articleNo/33084/title/Opinion--Super-Storm-Sandy/

[2] http://www.guardian.co.uk/global-development/datablog/2011/mar/18/world-disasters-earthquake-data

[3] We use throughout this paper two additional terms that accompany the concept of IM. We call *IM product* the resulting integrated model produced by the combination and integration of components and models. Any system, framework and technology that support the creation of IM products is called *IM tool*

[4] From now onwards, we assume the term IM in the realm of the environmental field, that is, integrated environmental modelling.

Although reuse should be a driving factor in assembling different models together in IM activities, the common practice differs (Voinov & Shugart, 2013) (Laniak et al., 2013). The high diversity of (environmental) models and the problem-specific design of system components still prevent reuse on a technical level. In this paper, we examine this scenery in detail and assess the conditions under which existing models and components can be assembled together, and shared among IM frameworks and tools. The remainder of the paper is organized as follows. Section 2 introduces required engineering concepts widely used throughout this paper. Section 3 describes the methodology and data employed in the review. Section 4 provides an in-depth analysis of reusability in a wide range of IM tools, accompanied by an overall discussion in Section 5. Concluding remarks are outlined in Section 6.

## 2 Engineering concepts of reuse

It is challenging to import and use existing models or components into IM tools. Modellers face for example difficulties to determine under which circumstances a particular component is compatible with a given IM tool, i.e. is that component potentially reusable? This paper explores this sort of questions and analyses a wide range of IM tools in order to assess the level of reusability in a qualitative manner. We assume that reusability is mostly determined by the interplay of (i) the execution environment of the target IM tools; and (ii) the nature of the third-party component or model.

The capability to easily import a third-party component or model into an execution environment is a fundamental aspect. We utilize the terms *white-box* and *black-box* to indicate the kind of techniques used to extend execution environments with new components and models (Fayad & Schmidt, 1997). The fact that a particular execution environment is designed following a white-box or black-box approach will reflect on the needed steps to import a third-party component into the environment. White-box approaches rely heavily on object-oriented features like inheritance. A new model and component is extended by inheriting from base classes defined in the target IM tool. It becomes a "closed" solution, which makes it difficult to be reused or transferred to other environments. In contrast to white-box approach, black-box environments are structured using object composition and delegation rather than inheritance. Furthermore, black-box environments emphasize dynamic object relationships rather than static class relationships. This leads to loosely-coupled components driven by design patterns such as strategy or mediator (Gamma, Helm, Johnson, & Vlissides, 1995).

For the third-party components, we distinguish well-known integration strategies such as *full integration*, *encapsulation* and *mediation*. First, full integration means that third-party models and components are fully embedded into target execution environments. This necessarily implies modifications of the source code of the component to be adapted into a target environment. On the positive side, it becomes a "native" component that takes for granted all of the characteristics supported by the IM tool. On the negative side, though, such modified components may be rarely reused in other IM tools. Second, encapsulation means third-party models and components interact with native components in the IM tool through a specific communication interface. Encapsulated components do not require invasive modifications of their source code since they are not regarded as "native" components, that is, are limited to other intrinsic characteristics of the target IM tools. Third, mediation (Wiederhold, 1992) is concerned with the addition of a middleware layer that implements the

required adaptation logic to potentially handle communication between various IM tools on one side, and third-party components on the other.

The relationship between these two factors will help us to analyse the easiness or difficulty of the integration of third-party components and models into IM tools, and thereby to determine the qualitative degree of reusability of such tools.

# 3 Methodology and data

The literature search for relevant papers was performed against well-known scientific data bases such as Web of Science [5]and Scopus[6], complemented with searches in individual journals that are relevant for the topic of study, i.e. reusability and integrated modelling in environmental fields. The combined results of these searches summed more than 200 references over the last decade (from year 2000 onwards) published mostly in scientific journals and high impact scientific conferences. Next we filtered manually these search results to avoid duplications and selected a potential group of relevant references (about 100). Next, we carefully reviewed each of these papers to finally establish the set of eligible papers for our study (n=77). Table 1 lists these papers per IM tool reviewed (See supplementary material).

INSERT TABLE 1 OVER HERE

We apply *viewpoints* to classify the eligible papers because many modelling paradigms and strategies have been proposed in IM during the last two decades. A viewpoint groups IM tools according to common strategies and building blocks to put pieces together into larger IM products. The environmental community have already paid much attention to *component-based modelling* paradigm. Nevertheless, we follow other approaches used mostly in other domains such as *scientific workflow systems, service-based modelling, virtual research environments, and resource-based modelling* to approach from different perspectives to the issue of reuse on IM. All of these viewpoints make special emphasis on characteristics such as modularity and decomposition to build IM products (Harris, 2002), which in principle are needed ingredients to support reusability. Indeed, from the technological perspective, the use of software engineering techniques greatly simplifies the integration and programming efforts required by scientists and modellers in IM activities, because the adaptation of well-known software engineering methodologies and design patterns ensures that IM tools are developed based on interoperable abstractions such as component, service, and resource (Verweij et al., 2010). Table 1 also shows IM tools grouped by viewpoint.

Our study has two unique characteristics. First, previous reviews are often constrained to a specific viewpoint or IM tool. For example Oinn et al. (2006) discussed the application of one scientific workflow system to life sciences, Lee and Percivall (2008) reviewed standards-based geospatial services in the realm of service-based strategies, and Argent (2006) compared three component-based modelling frameworks applied to an oil erosion scenario. The current review however considers various

[5] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

[6] http://www.info.sciverse.com/scopus

viewpoints, i.e., it covers an ample range of engineering techniques and tools. As a result, the IM tools and viewpoints studied here are not restricted to a particular field in the environmental sciences but to various disciplines such as ecology, agriculture, water resources, biodiversity, forestry, plant sciences, and life sciences (See supplementary material).

Second, the aim of the study is neither to propose a categorization of the array of IM tools analysed here nor to analyse them across different viewpoints. For the former, the diversity of conceptual approaches and theoretical backgrounds made it difficult to classify each IM tool under a single viewpoint. For the latter, rather than comparing for example component-based modelling frameworks with service-based modelling approaches, we compare and discuss IM tools within a viewpoint in order to understand how reusability is addressed by each viewpoint as a whole.

# 4 Viewpoint-based analysis of IM tools

In this section we analyse the selected IM tools by viewpoint. For each subsection, we briefly introduce the main concepts of the viewpoint and then compare and discuss the IM tools within that viewpoint, and ultimately reflect on reusability and integration. Based on the engineering concepts introduced earlier we selected a set of evaluation criteria to qualitatively characterize the level of reusability (See Table 2).

INSERT TABLE 2 OVER HERE

## 4.1 Component-based modelling frameworks

A component represents a coarse-grained functional view of various real-world entities, that is, it publishes only those functional aspects that are relevant to other components and systems. In this sense, a component provides a simplified view by hiding several dependences between real-world entities. Component-based modelling frameworks integrate components (and models) to produce IM products. Surveys and overviews on component-based modelling approaches for IM are well documented in the literature (e.g., Argent, 2004; Argent et al., 2006; Jagers, 2010). The list of component-based modelling frameworks in Table 3 shares the underlying ideas of modularity and decomposition. This way, the reuse of existing components is favoured against creating components completely new. Given the advantages of reuse models and components to create modular IM products, our assumption here is that some level of compatibility between the frameworks analysed in Table 3 must exist in order to share and integrate third-party components and models.

INSERT TABLE 3 OVER HERE

The OpenMI standard (Moore & Tindall, 2005; Gregersen, Gijsbers, & Westen, 2007) is aimed to couple diverse models implemented as local components through OpenMI frameworks, which contain (i) a set of application programming interfaces (API) to turn components and models into OpenMI-complaint components, (ii) a dedicated front-end editor to build OpenMI-compliant IM products, and (iii) a dedicated simulation/execution engine to run them (OATC, 2010). CCA (Common Component Architecture; Bernholdt et al., 2006) closely follows the basic tenet of encapsulating software functionality into components, specifically tailored to the needs of high-performance scientific

computing. Like CCA, The Earth System Modeling Framework (ESMF; Hill, DeLuca, Balaji, Suarez, & Da Silva, 2004) is also targeted to high-performance and parallel computing used mainly in global climate simulation and weather predictions (ESMF Joint Specification Team, 2011). Moore, Holzworth, Herrmann, Huth, and Robertson (2007) described the Common Modelling Protocol (CMP), a framework for building simulation models in a modular manner. CMP defines a transport protocol and describes a message based mechanism for packing and unpacking data, executable entry points, and a set of defined messages to transfer variables and events from one model and/or component to others involved in a simulation. The focus of BioMA is to run IM products against spatial databases. It is a direct result from the previous component-based framework called APES[7], which is aimed to estimate the biophysical behaviour of agricultural production systems in response to the interaction of weather, soil and agro-technical management options (Donatelli et al., 2010). Finally, the Invisible Modelling Environment (TIME; Murray et al., 2007) is targeted to hydrological catchment modelling applications (Stenson, Littleboy, & Gilfedde, 2011).

The overall level of heterogeneity of the above IM tools deserves further attention. IM tools in Table 3 are different in their scope (specific vs. generic) and in their technological basis (e.g. Java vs. .NET), which may pose some technical challenges to sharing and reusing components across them. In practice, a component or model developed for a specific IM tool is rarely compatible with other tools (Rizzoli et al., 2008; Donatelli et al., 2010) because of it is usually designed and developed for a particular component-based modelling framework. Take the Open MI approach as example. Existing third party components become OpenMI-compliant by following an inheritance strategy. Inheritance here means to create a new class that implements the based OpenMI interfaces. The integration of a new component is far enough to be an automated process because programing knowledge (C# or Java) and software developers (not modellers) are required to modify the source code of such a component. This to some extent limits reusability because accessing to the source code of a third-party component is needed (full integration). Other IM tools such as CCA and ESMF follow a similar inheritance strategy. The full integration approach is commonplace for the IM tools in this viewpoint.

From the standpoint of the execution environment, black-box frameworks are favoured against white-box frameworks to foster component reuse. This statement is easily mapped to Table 3 where all of the component-based modelling frameworks but BioMA/APES are white-box frameworks. This explains – combined with the extensive use of the full integration approach – why the integration and reuse of components across these IM tools is hardly realizable. Each component is especially modified to be aligned with the target IM tool. The emerging ESMF-CCA joint interoperability effort provides one exception (Bernholdt et al., 2006).

In summary, component–based modelling frameworks follow a full integration strategy to embed third-party components. In some cases the integration process is not so aggressive (e.g., BioMA), but in general source code of third-party components must be modified to inherit from base interfaces of IM tools. Reusing components among each other is then limited. Exceptions are rare but exist because of explicit interoperability agreements in place (e.g., CCA and ESMF). As Voinov and Shugart (2013) argued,

[7] APES stands for Agricultural Production and Externalities Simulator, http://www.apesimulator.org/

the lack of component reuse across component-based modelling frameworks may be partially explained by the extensive use of statistical models in this viewpoint. These models are tied to calibration data that in turn needs to be recalibrated when environmental conditions change. Distributed computing and web services technology are not poorly supported in component-based modelling frameworks.

## 4.2 Scientific workflow systems

The concept of workflow is seen as a set of analytical tasks such as data access, analysis, processing, and visualization (Deelman, Gannon, Shields, & Taylor, 2009). Workflow designers specify a control flow (e.g., sequence, forks, switch, joins, etc.) and data flow (how outputs of preceding tasks connect to the inputs of subsequent tasks) in order to structure the flow of required tasks. In a scientific context, most workflow tasks consist of the acquisition, manipulation, documentation, and processing of large amounts of scientific data, as well as the execution of computationally intensive analysis and simulations (Ludäscher & Goble, 2005). Scientific workflows may be then seen as a *description* of the combination of the previous tasks to meet scientific requirements. Such descriptions are then managed by scientific workflow systems, which are able to interpret and execute every single task contained within a scientific workflow.

Scientific workflow systems and component-based modelling frameworks share some commonalities and goals. Both typically support data-centric, dataflow-oriented workflows that can be computationally expensive and even require high-performance computation (e.g., CCA and ESMF). Furthermore, data-intensive use-case scenarios characterize both viewpoints. Ludäscher and Goble (2005), on the contrary, highlight annotation as the differentiator element. Scientific workflows are often more annotation-intensive due to the support of *reproducible* workflows in other scenarios, which requires detailed context metadata and data provenance information (Cohen-Boulakia & Leser, 2011). For instance, a common pattern in life sciences is to run the same workflow description (i.e., the same set of analytical tasks) over time but changing slightly the input data sets for each run. Control flow is not as relevant and critical as data sets used in each execution. In these cases, annotations (e.g., context metadata, provenance) become essential to document properly scientific workflows along with data sets used and produced, allowing scientists to create accurate records of scientific workflows in order to be reproduced or analysed later. Recent reviews on scientific workflow systems (Yu & Buyya, 2005; Rahman Ranjan, Buyya, & Benatallah, 2011) include most of the scientific workflow systems listed in Table 4. We here attempt to complement these studies from the perspective of workflow sharing and reuse.

INSERT TABLE 4 OVER HERE

The common characteristic of the above IM tools is their evolution from built-in functionalities towards the explicit support for remote web services and distributed computing. Most IM tools in Table 4 support web services through the Web Services Description Language (WSDL; Christensen, Curbera, Meredith, & Weerawarana, 2001) as a means to encapsulate third-party components to be part of scientific workflows. In addition Kepler (Ludäscher et al., 2006), Taverna (Oinn et al., 2004; Oinn et al., 2006), and VisTrails (Callahan et al., 2006; Santos, Lins, Ahrens, Freire, & Silva, 2009) support to some extent some OGC[8] services. This is a notable difference with respect to component-based workflow

[8] Open Geospatial Consortium, http://www.opengeospatial.org

frameworks, where the support of web services (and OGC services in particular) is rare. We further analyse this point in Section 4.4.

Taverna provides an extension for discovery and access to a wide range of web service repositories (Hull et al., 2006) such as the BioCatalogue (Bhagat et al., 2010), an online catalogue of web services for life sciences, which is part of the BioSharing[9] community. In addition, Taverna is closely connected to myExperiment, a social networking and web-based repository (Section 4.3), which foster the reuse and sharing of workflows by the scientific community.

Taverna and VisTrails provide good support for data provenance from different perspectives. The former focuses on the basic entities in scientific workflows such as workflow descriptions, individual services, and service inputs and outputs. However, the management of datasets used in such workflows are out of its scope (which may be addressed by the complementary SysMO project, see Section 4.3). This means that not all of the entities needed to support reproducibility are tracked in terms of data provenance. For the latter case, VisTrails is centred on the workflow output as the main entity for tracking data provenance. Any change in the workflow that may produce a distinct output is tracked as a different workflow version. The ability to ensure different workflow versions makes it easy to reproduce past versions of the same workflow. For this reason, VisTrails scores high in data visualization compared with the other scientific workflows systems. Indeed, this is a distinctive feature in VisTrails, since it is used in visualization-centric applications such medical imaging and dynamic simulations (e.g., earthquakes) in environmental and geospatial sciences (Silva, Anderson, Santos, & Freire, 2011).

In summary, scientific workflow systems largely exploit the reuse of remote web services to build scientific workflows. Compared with component-based modelling frameworks, this is an important step to promote reuse of third-party components, models and services in IM products. The interplay of local built-in functionalities and remote web services allows scientific workflow systems to follow a mixed approach between full integration and encapsulation in terms of reusability. On one hand, these IM tools come with a bunch of built-in local components that cover basic functionalities such as visualization, transformation, and management of input-output data. For instance, Triana (Taylor, Shields, Wang, & Rana, 2003; Taylor, Shields, Wang, & Harrison, 2006; Churches et al., 2006), Kepler and VisTrails provide sophisticated graphical interfaces and toolboxes for assisting modellers in the creation of workflows. On the other hand, these IM tools also support the addition of WSDL-based services into scientific workflows (e.g., Taverna, Kepler). In this sense, scientific workflow systems take the benefits of component-based modelling frameworks (execution environment, user control, etc.) and distributed web services (flexibility, additional functionalities via remote web services) to produce “hybrid” IM tools capable to work in varied scenarios such as life sciences, medical imaging, simulation, and environmental sciences. In addition, some of these systems (e.g., VisTrails and Taverna) provide other remarkable characteristics such as good support for data provenance. An added value of scientific workflow systems in general and Taverna in particular is their closeness to associated virtual research environments as we explain next.

[9] http://www.biosharing.org/

## 4.3 Virtual research environments

Virtual research environments (VREs) support multidisciplinary, collaborative research by managing numerous tasks involved in carrying out research at both small and large scales on the Web. Voss and Procter (2009) defined VREs as information infrastructures, collaborative tools and technologies needed by researchers to do their daily research activities, interact with other researchers, and to enable vertical (specific) and horizontal (generic) integration of resources on the Web. Apart from supporting generic research tasks such as data management (horizontal view), VREs should allow scientists to use their tools, technologies, and needs of particular disciplines (vertical view). Table 5 lists some promising implementations of VREs that are mainly web-based developments to improve social and collaborative aspects and enable the sharing of scientific workflows (Section 4.2).

INSERT TABLE 5 OVER HERE

VREs do not should thought of as a unique research platform but multiple infrastructures interconnected addressed to different purposes, similarly to the vision of multiple Digital Earth applications (Craglia et al., 2012). However, incipient VREs in Table 5 are still isolated and come mainly from particular fields such as molecular biology and bioinformatics. As each VRE targets a particular domain, their implementations follow different approaches to support much better specific requirements and needs. For this reason, some VRE implementations may even present opposed architectural designs: Galaxy (Giardine et al., 2005) is in turn a workflow framework like Taverna and an application server for workflow sharing like myExperiment (De Roure, Goble, & Stevens, 2009).

In general, VREs in Table 5 support collaborative research activities at different degrees. MyExperiment combined with Taverna provide the most complete pack to share scientific workflows. This is demonstrable by the wide use and proven projects based on these two complementary tools. Taverna allows the creation of scientific workflows and myExperiment facilitates the discovery and sharing of so-called "research objects" (De Roure, Goble, & Stevens, 2009). Rather than treating workflows as mere process-based descriptions, the notion of "research objects" considers also input and output data, provenance information, metadata, and any related piece of information relevant to fully understand a given workflow – so that experiments can be reproduced at any time. For instance a collection of research objects may include Taverna-based workflows, metadata, input and output data, execution logs, PDFs of papers, presentations, and other related objects that altogether refer to the same experiment. Experiments as aggregations of research objects are described using the OAI-ORE specification[10] that allows exporting to RDF (Resource Description Framework) format. This enables for instance semantic queries via a SPARQL[11] endpoint to find out about connections between different experiments and research objects (Bechhofer et al., 2013).

MyExperiment also exposes web services and APIs to ease the integration of tools, services, and aggregated objects into customized applications (Goble et al., 2010), facilitating thus some kind of vertical integration (Voss & Procter, 2009). On the negative side, in myExperiment data sets must be

[10] Open Archives Initiative Object Reuse and Exchange, http://www.openarchives.org/ore/

[11] A query language for RDF sources, http://www.w3.org/TR/rdf-sparql-query/

managed by the institution that created a workflow. This lack of support for data sets that are used and produced in myExperiment may pose limitations to replication. The data management issue is being mitigated by the SysMO-DB tool (Globe et al., 2009). The notion of *asset* in SysMO-DB is built on the above notion of research object to support reproducibility and replication by other users. Indeed, SysMO-DB acts as a mediator with a centralized data model to allow modellers to use their own workflows and dataset formats. So, the pair Taverna-MyExperiment extended with the data management features from SysMO-DB – a centralized data model and asset catalogues (Wolstencroft et al., 2011) – seems to help researchers and scientists to tackle with some challenges in collaborative scientific workflows.

Reproducibility is one of main goals in VREs. In this context, Galaxy Pages is a data provenance mechanism to support research reproducibility (Goecks, Nekrutenko, & Taylor, 2010). Galaxy Pages are like mash-ups web pages that enable users to document an entire experiment and may include a mix of media resources (e.g., text, graphs) to describe and annotate the steps of experiments, as well as embed other Galaxy items, such as datasets, histories and workflows. Thereby, the web pages become more of aggregated research objects, which address human beings instead of being machine-processable.

DataONE proposed a federated approach where different infrastructure nodes collaborate to enable data sharing (Michener et al., 2012). One key value of DataONE is the strong support for long-term preservation of data (Bach et al., 2012). This is a value asset since scientists likely make use of previous datasets for calibration purpose in new experiments. DataONE then ensures data quality, persistence and curation by means of established policies for data sharing and reuse. In this sense, DataONE and SySMO-DB put more emphasis on data sharing and preservation, whereas myExperiment, Crowdlabs and Galaxy pursue sharing scientific workflows in terms of descriptions of scientific tasks. Both aspects are primordial for enabling reuse and reproducibly in IM (Michener & Jones, 2012).

In summary, VREs are much like enabling infrastructures composed of set of tools, technologies, and services needed to do daily research activities. As middleware infrastructures, VREs may be seen as a collection of mediators to deal with particular research tasks: from the interaction and collaboration with other researchers, to other analytical tasks, including sharing, discovery and access of research resources. The notion of research resources (or assets) is also central in VREs. What is a first-class asset? How are similar assets combined? A common approach across VREs is to consider aggregations of related resources as a common data model to manage scientific resources (e.g., myExperiment Research Objects, Galaxy Pages, SysMO-DB assets). For instance, aggregation of research objects in myExperiment may group distinct kinds of resources together, such as an IM product, input datasets, calibration data, and related technical publications and presentations.

In general, current VREs focus on particular domains: Galaxy on genomic research, myExperiment on life sciences and bioinformatics. VREs devoted to environmental sciences are still rare. DataONE may be one exception but it is only pursuing data sharing and persistence so far (indeed, very valuable) and not in supporting other activities such as sharing and reusing models and workflows.

## 4.4 Service-based modelling

Like a task in scientific workflows, a service is an abstraction unit that allows modellers to share and encapsulate any given functionality. In service-based modelling, building blocks such as data, processes and models are exposed as services available and accessible via standards interfaces on the Web (Alonso, Casati, Kuno & Machiraju, 2004; Papazoglou, 2008). In this context, service-oriented architectures (SOA) are usually adopted in the development of collaborative, distributed web applications based on reusable and standardized components and services (Friis-Christiensen, Lucchi, Lutz, & Ostländer, 2009; Yang, Raskin, Goodchild, & Gahegan, 2010). SOA relies on service composition or service chaining that has been widely recognised as an approach to piece together two or more services together so as to support the development of rapid and interoperable distributed applications (Papazoglou, Traverso, Dustdar, & Leymann, 2007).

We discuss relevant works on (geospatial and environmental) service composition restricted to web services described by WSDL (Christensen, Curbera, Meredith, & Weerawarana, 2001) and WPS (Schut, 2007) specifications, which are widely used for service interface description in mainstream web services (Papazoglou, 2008) and geospatial services respectively (Lee & Percivall, 2008). Their use leads to WSDL-based or WPS-based services. This dualism is demonstrated in many examples that have assembled web services, either WSDL- or WPS-based, into geospatial service compositions. On one hand, Li, Di, Han, Zhao, and Dadi (2010) describe how GRASS commands can be exposed as WSDL-based web services to be orchestrated by WS-BPEL[12] engines. On the other, Kiehle (2006), Fook, Monteiro, Câmara, Casanova, and Amaral (2009), Granell, Díaz, and Gould (2010), Foerster, Lehto, Sarjakoski, Sarjakoski, and Stoter (2010), and Maué, Stasch, Athanasopoulos, and Gerharz (2011) are just a few examples aiming at integrating and sharing geospatial data and environmental models in terms of WPS-based web services.

These solutions are still isolated in terms of using either WSDL-based or WPS-based services. To address the lack of interoperability between WSDL- and WPS-based services, and then promoting reusability, the WPS specification permits linking to associated WSDL documents to describe both an entire WPS service and each single contained process. In the first case, the WPS capabilities document can include a descriptor pointing to the corresponding WSDL description file (URI). The WSDL document describes the interface of the entire WPS service instance. In the second case, each contained process description can have its own descriptor that points to a specific WSDL description file. In this case, the WSDL document provides the functional signature of a WPS process. The explicit support for WSDL descriptions is a way to make OGC services in the "closed" geospatial domain available to mainstream, mass-market web services – which are mainly based on WSDL – so that hybrid solutions might be suitable. Schade et al. (2012) recently discussed this WSDL-WPS integration.

Nevertheless, Lopez-Pellicer, Rentería-Agualimpia, Béjar, Muro-Medrano, and Zarazaga-Soria (2012) have recently evaluated the availability of WPS-based services on the Web. The results showed that none of the WPS services found in their survey (only 0.6% out of public OCG service instances found) used WSDL documents for describing processes individually. Hence, current WPS service instances still

[12] Web Services – Business Process Description Language, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf

offer very limited support for WSDL standard, which unavoidable limits reusability and interoperability with mass-market web services and most importantly with scientific workflow systems that support WSDL-based services (Section 4.3).

Some incipient works may start to change this trend though. As WS-BPEL works mainly with WSDL-based web services, OGC services should be realigned to these standards in order to benefit from existing tools, workflow engines and mass-market web service community. Yu et al. (2012) provided WSDL-based descriptions for various kinds of OGC services including WPS. These OGC services were orchestrated and executed using a geospatially-aware WS-BPEL workflow engine and tested in some environmental and geospatial use cases. Similarly, scientific workflow systems are also becoming attractive for data-intensive geospatial and environmental service compositions. For example, de Jesus, Walker, Grant, and Groom (2012) have recently presented a proof-of-concept WPS server implementation that is able to produce WSDL descriptions for WPS services. In their tool, such WSDL descriptions are then used from Taverna workflows, exploiting then Taverna built-in functionalities and tools to manage and execute geospatial scientific workflows. It is worth to note here that Taverna for example does not manage some input parameters of geometry type like bounding box, so the execution of geospatial services that require a bounding box as input from Taverna fails. Yu et al. (2012) have reported a similar flaw in WS-BPEL workflow engines as they are not designed for managing geospatial data definitions.

From the above examples we can conclude that the current trend is to encapsulate OGC services as WSDL-based services. OGC services remain unchanged and can be (re-)used in target IM tools such as business workflow engines (e.g., WS-BPEL) and scientific workflow systems (e.g. Taverna). Nevertheless, the success of the encapsulation approach to geospatial and environmental services depends strongly on the nature of the services themselves. First, the level of complexity of some geospatial schema may be a limitation at runtime, because some IM tools (e.g., Taverna, WS-BPEL engines) may not properly deal with some geospatial singularities such as complicated relationships and recursive type definitions (Tamayo, Granell, & Huerta, 2012). Similarly, WSDL parsers still offer poor support for complex data type definitions as some geospatial elements schema have (Yu et al., 2012). Second, geospatial workflows are often data-intensive and may require long processing times in execution and retrieving inputs data sets, which is not a remarkable feature in current business workflow engines (Barga & Gannon, 2007). In this sense, the implementation of geospatial enhancements into workflow engines and systems has led to geospatially-aware execution environments, such as BPELPower (Yu et al., 2012) and EO4VisTrails (McFerren, van Zyl, & Vahed, 2012). The latter is a geospatial extension of VisTrails to manage such special requirements (complexity schemas, long processing times, etc.) of geospatial workflows to increase up reusability and interoperability between OGC services and these systems.

In summary, service-based modelling follow an encapsulation approach exemplified by the prominent role of OGC WPS services in varied scenarios: (i) its key role in OGC service chaining and composition; (ii) its usage in scientific workflow systems such as Kepler, Taverna, and VisTrails applied to environmental and geosciences fields; (iii) its use in business process workflows such as WS-BPEL; (iv)its potential relation to component-based modelling frameworks such as OpenMI 2[13]; and finally (v) its role as

[13] http://www.openmi.org/

mediator to distribute processing capabilities over different computation models such as cloud and grid (Giuliani, Nativi, Lehmann, & Ray, 2012). This reflects the importance and flexibility of OGC WPS services to be part of IM products. On the negative side, however, it seems that building IM products only upon WPS services (and other types of web services) seems not to meet basic requirements in IM. Among others, the lack of user control and interactivity during execution seems to limit genuine service-based approaches in IM solutions.

## 4.5 Resource-based modelling

In previous sections we have seen different viewpoints on existing IM solutions. *Component-based modelling frameworks* offer robust mechanisms and environments for constructing IM products. *Service-based modelling* focuses on reusing distributed web services to create flexible workflows. *Scientific workflows systems* provide annotation capabilities for interactive, data-intensive scientific workflows. *VREs* offer aggregations of related resources as build blocks for sharing and reused. Despite their benefits, these viewpoints still suffer some barriers in terms of reusability. For example components and models are still coupled to specific interfaces and communication protocols provided by target IM tools, which implies that developers and modellers have to continuously adapt IM solutions to new versions of these interfaces and communication protocols.

Resource-based modelling may be an alternative approach to avoid tight-coupled interfaces and hence to ensure reusability over time. Resource-based modelling relies on the REST (Representational State Transfer) principles (Fielding, 2000) to turn HTTP (Hypertext Transfer Protocol) into an application protocol capable of manipulating and accessing resources (e.g., models, services, components). It seems reasonable to use HTTP directly rather than using multiple APIs and interfaces. Think for example on the interfaces, protocols and service interfaces from the IM tools visited in earlier viewpoints to get an idea of the variety and diversity of APIs potentially available for IM. Resource-based modelling should not be understood as a solution *per se* for enhancing reusability in IM but a complementary approach in combination with the preceding viewpoints. The adoption of resource-oriented modelling may ease interoperability between IM tools from distinct viewpoints (See Section 5).

In the geospatial domain, most recent works under resource-based modelling have attempted to specify a set of resources for a variety of OGC services with well-known data models. Mazzetti, Nativi, and Caron (2009) discussed a RESTful migration for publishing raster datasets (coverage data model) in comparison to the counterpart OGC Web Coverage Service interface (WCS; Baumann, 2010). The authors concluded that resource-based modelling may fit well to particular geospatial and environmental scenarios, though, the selection of the modelling approach (service-based via WSDL vs. resource-based via REST) depends largely on the particular use case in hand. In this line Pautasso, Zimmermann, and Leymann (2008) even claimed that RESTful services are well suited for ad hoc integration scenarios, whereas WSDL-based services are more flexible to address advanced requirements such as quality of service and security commonplace in business and industry settings.

Foerster, Brühl, and Schäffer (2011) proposed a RESTful interface for OGC WPS services. They concluded that their REST implementation, based on resources, breaks with the current data model described in the WPS specification. The authors suggested that next releases of OGC WPS specifications should be designed in a more modularized way to reflect different architectural styles (e.g. SOA, REST) sharing a

common model. Granell, Díaz, Tamayo, and Huerta (in press) assessed the application of REST principles to WPS-based services from a theoretical perspective. The authors suggested that RESTful interfaces are flexible enough to ease reuse and adaptation of geo-processing services in varied compositions, but changes in the underlying data model are necessary to move from a service-oriented to a true resource-oriented perspective.

Several recent studies follow this path. Finney and Watts (2011) proposed to enhance feature catalogues for supporting cross-domain access to various geosciences and environmental communities. To doing so, the authors explored a REST-based approach for geographic features through an enhanced implementation of an ISO 19110-based Feature Catalogue. ISO 19110 (ISO 19110, 2005) defines a methodology for cataloguing feature types and specifies how the classification of feature types, feature attributes, and feature relationships are organized into a feature catalogue. The REST-based implementation allows users to interactively retrieve details of the individual resources (profiles, feature-type, relationships, etc.) as well as relations to other resources contained in the feature catalogue. In the second work, Janowicz et al. (2013) introduced a RESTful proxy for the OGC Sensor Observation Service (SOS; Bröring, Stasch, and Echterhoff, 2012) to assign meaningful identifiers to sensor data and to directly publish raw sensor data on the Web. Most interestingly, the authors extended the OGC Observations & Measurements (O&M; ISO 19156, 2011) standard data model with Linked Data features and proper semantic vocabularies and ontologies (Heath & Bizer, 2011).

From the examples above, the applicability of the resource-based approach to current OGC service interfaces and underlying data models may lead to contradicting data models depending on the implementation strategy selected. As described by Granell et al. (in press) the exact mapping of OGC service interfaces into a set of interrelated resources may lead to conceptual gaps between service-based and resource-based data models. Resource-based modelling thus requires a conceptual shift not only at interface level but also at data model level. Therefore, resource-based modelling approaches should build upon well-defined, abstract data models such as *feature* and *observation* data models to create suitable service implementations (either service-based or resource-based) with a core, shared data model. Following these guidelines, service interfaces may be incompatible (different architectural styles) but using a common data model (e.g., feature, observation) behind the scenes. This suggests that RESTful experiments based on proper data models (Finney & Watts, 2011; Janowicz et al., 2013) are promising strategies to leverage resource-based modelling in IM. In this regard Granell, Díaz, Schade, Ostländer, and Huerta (2013) and Nativi, Mazzetti, and Geller (2013) recently discussed on design practices and implementation recommendations to build resource-based interfaces for environmental models.

In summary, like service-based modelling, resource-based modelling may complement other modelling viewpoints. For instance, DataONE (Section 4.3) provides RESTful APIs to access resources exposed by any infrastructure node. Users are able to interact with public resources available from a node to request data and metadata as well as log and even status information, among other functionalities. The implementation for this API relies on the resource abstraction. Other IM tools in the VRE and scientific workflow systems viewpoints also expose RESTful APIs to access individual and aggregated resources

and workflows respectively, ensuring uniform access and manipulation regardless of the IM tools employed.

# 5 Cross-viewpoint discussion

In this section we examine potential relationships and connections between viewpoints in terms of reusability. Figure 1 illustrates a simple 2-axis plot to qualitatively represent the level of reuse of each viewpoint analysed in the previous section. On the vertical axis, we use the white-box and black-box approach and, on the horizontal axis, we focus on the third-party component reuse approaches, as commented earlier in Section 2.

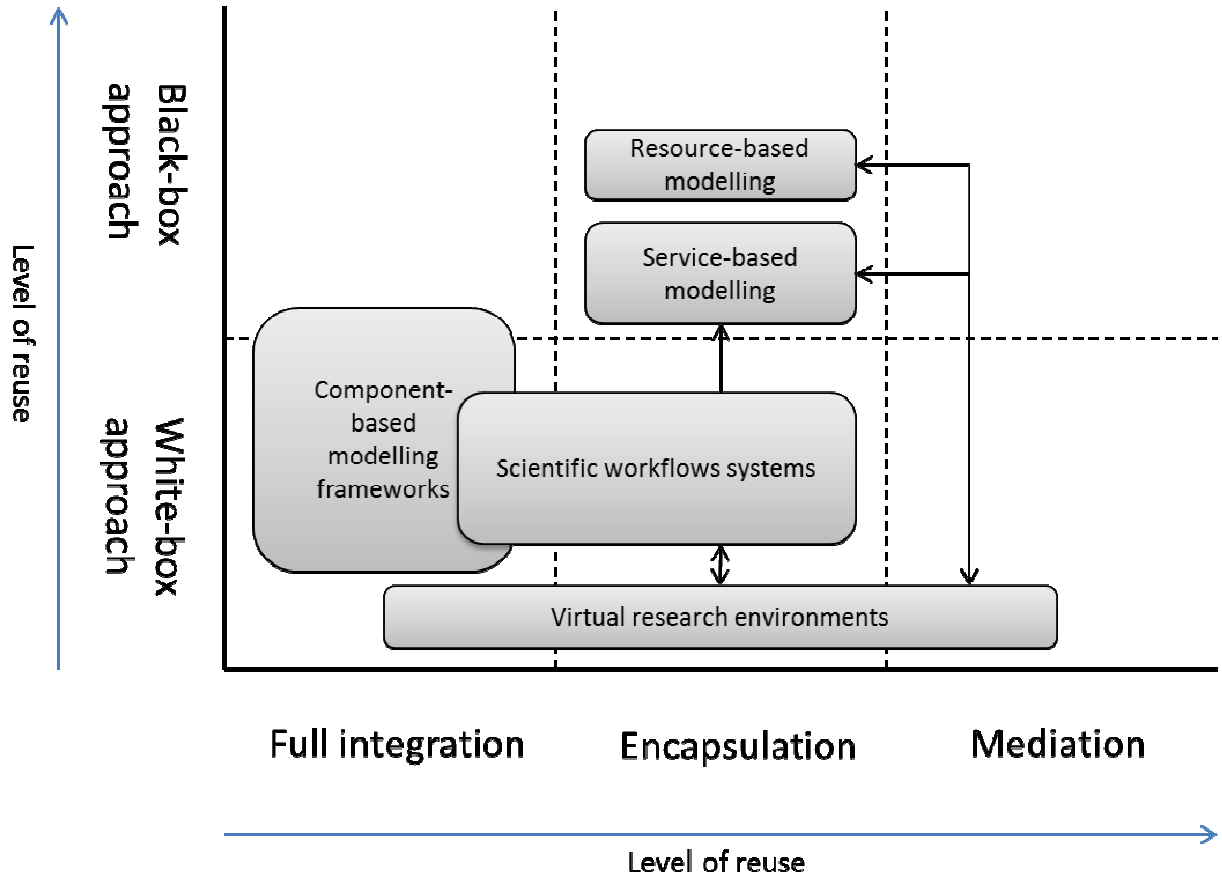


Figure 1. Cross-viewpoints interactions according to the engineering concepts described in Section 2

The advantages of adopting IM tools based on component-based modelling frameworks (Section 4.1) and scientific workflow systems (Section 4.2) are that these tools provide robust environments to execute and control IM products. Scientists may monitor each step of an IM product run, go back and forth, and may even manipulate state variables and parameters of the contained models. Full user control on the execution and simulation of IM products is a strong requirement from the environmental community. This also explains why full integration is the ”common”, widely supported way of integrating third-party components in these IM tools, mostly for component-based modelling frameworks (left-

bottom corner in Figure 1). Conversely, these IM tools face difficulties in the integration of other models and components that do not conform to their particular interfaces. As illustrated in Figure 1, black-box approach is in general poorly considered to develop IM tools despite it follows well-designed patterns to better support reuse of third-party components. Perhaps, the complexity, uncertainty, and need of skilled team-work imposed by black-box environment implementations are barriers to effective adaptation in favour of white-box approaches.

Component-based modelling frameworks (Section 4.1) are mostly domain-focused IM tools, whereas service-based modelling solutions (Section 4.4) are domain-independent. This is one of the reasons of the lack of reusability in component-based modelling frameworks (left-bottom corner in Figure 1). Components are mostly designed to talk to specific IM tools, which limits reuse between distinct IM tools even in the same viewpoint. However, being more specific, i.e., addressing specific environmental modelling requirements also has some benefits. For example, component-based modelling frameworks are more advanced in supporting IM requirements than service-based ones because the latter deals with generalities. Indeed, successful IM products reported in the literature (see Table 1) are mostly based on component-based modelling frameworks, which better "understand" the nature of environmental issues.

Scientific workflow systems (Section 4.2) are mostly targeted to life sciences and bioinformatics and support mostly remote web services as tasks within scientific workflows. The former aspect suggests that scientific workflow systems are oriented to concrete domains as in the case of component-based frameworks for environmental sciences. Particular requirements lead to specific tools and systems so as to successfully address the peculiarities of each discipline. Indeed, connections between component-based modelling frameworks and scientific workflow systems are still rare, since each one addresses distinct application domains. The latter aspect, support of web service technologies, represents an enhancement with respect to component-based modelling frameworks (See arrow from scientific workflow systems box to service-based modelling box in Figure 1). However, rather than assembling scientific workflows uniquely with remote services, the typical scenario is to combine local or built-in components (basic functionalities such as spatial commands, data flow management, data visualizations, etc.) with specialized remote services (e.g., bio-analysis, genomic analysis, data mining, etc.) to create a kind of "local-remote" workflows. This makes sense since not every task of a workflow needs to be a web service: some basic tasks such as input/output data management and results visualization may be locally supported by the IM tool itself, leaving other workflow tasks be covered by specialized web services. In this sense, scientific workflow systems are somewhere placed between pure desktop-based solutions (e.g., component-based modelling frameworks) and pure distributed, web-based solutions (e.g., service-based modelling), as illustrated in the central part of Figure 1.

VREs (Section 4.3) enable the sharing and (vertical and horizontal) integration of resources needed by different stakeholders (scientists, policy makers, modellers, etc.) throughout the lifecycle of an IM product. However, some open questions still remain such as how to address "customized" VREs, i.e., those VREs that are able to incorporate the specific characteristics of each discipline. For instance, genomic and environmental researchers have certainly different perspectives and therefore each group would need and interact with VREs in a different manner. The notion of aggregation is also connected to

this issue. Aggregation by definition implies reusability, i.e., an aggregation simply reuses the resources it contains. Nevertheless, a pending question is how to map the aggregation concept to IM and the singularities of environmental sciences (Laniak et al., 2013).

As commented in Section 4.3, the notion of *mediation* is widely used in VREs. Identifying generic mediators and components that are generic enough to be used across disciplines is a driving force in VREs. Generic mediators would lead to reuse on a large scale and avoid duplication by facilitating horizontal integration (same functionality used in many disciplines) and vertical integration (customized and specific mediators are built on top of generic ones). Nevertheless, further research is needed to delimit the boundaries of "similar disciplines or topics" so as to define the shared mediators in cross-cutting disciplines.

In summary, component-based modelling frameworks and scientific workflow systems have been traditionally used in IM but current trends are looking at service-based modelling, resource-based modelling, and virtual research environments. Service- and resource-based modelling viewpoints are better positioned in terms of reusability as they support the creation of loosely-coupled services driven by dynamic relations between service interfaces (black-box approach).

# 6 Concluding remarks

In this paper we have explored different sorts of viewpoints and IM tools from the reusability point of view. Based on the analyses and reflections in previous sections, we draw the following concluding remarks:

- The environmental issue at hand usually defines the viewpoint to choose, that is, there is not a clear winner viewpoint or IM tool for all cases and situations. Each environmental discipline presents requirements and needs that are better addressed by specific IM tools. Taking as example component-based modelling frameworks, OpenMI was designed for hydrological sciences, ESMF for global climate simulations, and CMP for agricultural simulation. These tools are thus better positioned for dealing with modelling tasks in such disciplines.
- Reuse of models across IM tools from different viewpoints is rare except for IM tools that support web service technologies, such as some scientific workflows systems and VREs. These tools are able to use the same WLSD-based web service. Exceptions to that are few and mostly come as *ad-hoc* implementations. For example Turuncoglu, Murphy, DeLuca, and Dalfes (2011) combined a component-based modelling framework (ESMF) and a scientific workflow system (Kepler) for a particular modelling scenario.
- Reuse of models across individual IM tools within the same viewpoint varies in function of the particular viewpoint. For component-based modelling frameworks, components are often developed and fully integrated for an individual framework. In scientific workflow systems, however, Taverna-based workflows are becoming *de-facto* format to share workflow descriptions among scientific workflows systems. In the case of VREs, the disparity of architectural styles and implementations employed make it extremely difficult to share models and resources between VRE implementations.

- Reuse of models and components is readily achievable between instances of the same IM tool. This means for example that CCA-compliant components and Kepler-based workflows can be shared and reused between running instances of CCA frameworks and Kepler systems respectively.
- In general, the level of reuse decreases dramatically when it comes to the sharing and reuse of an integrated model - an IM solution - as a whole, instead of a single component.

These remarks suggest that IM is seen as a project-oriented activity to explore an environmental problem or scenario. For this reason, some authors claim models or components used in a particular scenario cannot be realistically reused in others because of the large number of variables and parameters involved in models impede that such a model be exactly replicated (Parker et al., 2002). In this context, a given model may work for a specific environmental problem but other modellers and scientists could not use it in similar situations due to different environmental conditions.

Looking into the future we can envisage two hypothetical scenarios to improve reuse in IM. The first scenario would reduce the heterogeneity and variety of IM tools. Suppose that OpenMI and Taverna would become the "winning tools" in their respective viewpoints, and that interoperability between each other is achieved. Modellers would only have to create models, components and IM products compliant with these tools to ensure reusability. Although this hypothetical solution is technically feasible, it seems unrealistic that a few IM tools were capable to deal with any IM activities from different fields and disciplines. Simply the array of environmental problems, stakeholders, data formats, social aspects, and needs that define each discipline make this unaffordable.

The second hypothetical scenario would accept that integration will not happen generically, but on demand, i.e., in specific contexts. Instead of a few "omnibus" tools for IM as commented above, many IM tools would generate reusable IM products applicable to certain domains. One may argue that this happens today in the sense that models, components and services are already shared in some domains such as bioinformatics, genomics and ecology. In our opinion, though, the reuse of only models or components fails to achieve reusable and sustainable IM solutions. It is difficult if not impossible to reuse a model in isolation in other environmental scenarios. The current situation tends to "integral modelling" solutions (Voinov & Shugart, 2013), in which models are specifically designed not with the aim to be reused but to be uniquely part of an IM solution.

A conceptual shift based on "seeing the forest through the trees" is required to promote and foster reuse in IM. On one hand, IM activities are increasingly requiring joint research and collaboration towards a shared understanding of multi-disciplinary problems. However, this does not mean that every model may be exported and used to any other domain. Like a model, a tree also depends on its environment such as type of soil, climate, and meteorological conditions. A tree uniquely survives in those forests where its specific environmental conditions are met. On the other hand, new approaches for IM would account for not only models but also their context (data, documentation, provenance, metadata, results, etc.) as first-class citizens. A model is partly defined without its inputs data, calibration data, documentation, and other defining aspects that help others to correctly interpret the capabilities and limitations of a model. A model - like a tree - is just one part of an IM solution - the forest. An IM solution in terms of only contained models and components is a partial view of the forest.

By understanding that a forest is an ecosystem of interrelated trees, soil, climate conditions and other variables, we may better recognize and realized of the value of a forest as a whole.

As we move from left to right in Figure 1, a model is increasingly accompanied with related resources, from annotation capabilities in scientific workflow systems to the concept of research objects and assets (aggregation of resources) from the VRE viewpoint. Future research in IM should follow this direction to handle models and components as collections of aggregated and individual related resources such as needed datasets, metadata and documentation, results, execution environment used, and even involved people. The aggregation concept should be accompanied with service–based and resource-based modelling approaches to foster sustainability and reusability in IM. In addition, the next wave of VREs should be built upon a collection of inter-related mediators to co-operatively manage diverse tasks ranging from interaction and collaboration with other researchers to the sharing, annotation, and integration of aggregations to execution and simulation of IM products.

## Acknowledgements

This work has been partially supported by the ENVIROFI FP7 project (Grant Agreement No. 284898).

# References


Alonso, G., Casati, F., Kuno, H. & Machiraju, V. (2004). *Web Services: Concepts, Architectures and Applications*. Springer

Argent, R.M. (2004). An overview of model integration for environmental applications-components, frameworks and semantics. *Environmental Modelling and Software*, *19*(3), 219-234.

Argent, R.M., Voinov, A., Maxwell, T., Cuddy, S.M., Rahman, J.M., Seaton, S., Vertessy, R.A., & Braddock, R.D. (2006). Comparing modelling frameworks - A workshop approach. *Environmental Modelling and Software*, *21*(7), 895-910.

Bach, K., Schäfer, D., Enke, N., Seeger, B., Gemeinholzer, B., & Bendix, J. (2012) A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics*, *11*, 16–24.

Barga, R., & Gannon, D. (2007). Scientific versus business workflows. In I. Taylor, et al. (Eds.), *Workflows for e-Science* (pp. 9-18). Heidelberg: Springer.

Baumann, P. (Ed.) (2010). OGC WCS 2.0 Interface Standard – Core, Version 2.0. Open Geospatial Consortium Interface Standard. http://www.opengeospatial.org/standards/wcs

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., & Goble, C. (2013). Why linked Data is Not Enough for Scientists. *Future Generation Computer Systems*, 29(2), 599-611.

Bernholdt, D. E., Allan, B. A., Armstrong, R., Bertrand, F., Chiu, K. , Dahlgren, T. L., Damevski, K., Elwasif, W. R., Epperly, T. G. W., Govindaraju, M., Katz, D. S., Kohl, J. A., Krishnan, M., Kumfert, G., Larson, J.W., Lefantzi, S., Lewis, M. J., Malony, A. D., McInnes, L. C., Nieplocha, J., Norris, B., Parker, S. G., Ray, J., Shende, S., Windus, T. L., & Zhou, S. (2006). A component architecture for high-performance scientific computing. *International Journal of High Performance Computing Applications*, *20*, 163–202.

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R., & Goble, C.A. (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research*, *38*(suppl 2), W689-W694.

Bröring, A., Stasch, C., Echterhoff, J. (Eds.) (2012). OGC Sensor Observation Service Interface Standard, Version 2.0. Open Geopsatial Consortium Implementation Standard. http://www.opengeospatial.org/standards/sos

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, Huy T. (2006). VisTrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (pp. 745-747). New York, NY, USA: ACM.

Cohen-Boulakia, & S., Leser, U. (2011). Search, Adapt, and Reuse: The Future of Scientific Workflows. *SIGMOD Record*, *40*(2), 6-16.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). Web Services Description Language (WSDL) 1.1. The World Wide Web Consortium (W3C), http://www.w3.org/TR/wsdl.

Churches, D., Gombas, G., Harrison, A., Maassen, J., Robinson, C., Shields, M., Taylor, I., & Wang, I. (2006). Programming scientific and distributed workflow with Triana services. *Concurrency and Computation: Practice and Experience*, *18*(10), 1021-1037.

Common Modeling Protocol (CMP), 2008. Developing Managed Code Components for the Common Modeling Protocol. Version: 17 July 2008. http://www.grazplan.csiro.au/files/developing_man_components.pdf

Craglia, M., de Bie, K., Jackson, D., Pesaresi, M., Remetey-Fülöpp, G., Wang, C., Annoni, A., Bian, L., Campbell, F., Ehlers, M., van Genderen, J., Goodchild, M., Guo, H., Lewis, A., Simpson, R., Skidmore, A., & Woodgate, P. (2012). Digital Earth 2020: towards the vision for the next decade. *International Journal of Digital Earth*, *5*(1), 4-21.

de Jesus, J., Walker, P., Grant, M., & Groom, S. (2012). WPS orchestration using the Taverna workbench: The eScience approach. *Computers & Geosciences*, *47*, 75-86.

De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, *25*(5), 561-567.

Deelman, E., Gannon, D., Shields, M.S., & Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, *25*(5), 528-540.

Donatelli, M., Russell, G., Rizzoli, A.E., Acutis, M., Adam, M., Athanasiadis, I.N., et al. (2010). A component-based framework for simulating agricultural production and externalities. In: F. Brouwer, & M.K. van Ittersum (Eds.), *Environmental and agricultural modelling: integrated approaches for policy impact assessment* (pp. 63-108). Dordrecht: Springer.

Donchyts, D., Hummenl, S., Vaneçek, S., Gross, J., Harper, A., Knapen, R., Gregersen, J., Schade, P., Antonello, A., & Gijsbers, P. (2010). OpenMI 2.0 – What's new? In *International Environmental Congress on Environmental Modelling and Software Society (iEMS 2010)*, Otawa, Canada.

ESMF Joint Specification Team (2011). ESMF Reference Manual for Fortran, Version 5.2. Earth System Modeling Framework, http://www.earthsystemmodeling.org/esmf_releases/public/last/ESMF_refdoc/

Fayad, M., & Schmidt, D.C. (1997). Object-oriented application frameworks. *Communication of ACM*, *40*(10), 32-38.

Fielding, R.T. (2000). Architectural Styles and the Design of Network-based Software Architectures. PhD dissertation, University of California (Irvine), http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Finney, K.T., & Watts, D. (2011). REST-based semantic feature catalogue services. *International Journal of Geographical Information Science*, *25*(9), 1507-1524.

Foerster, T., Lehto, L., Sarjakoski, T., Sarjakoski, L.T., & Stoter, J. (2010). Map generalization and schema transformation of geospatial data combined in a Web Service context. *Computers, Environment and Urban Systems*, *34*(1), 79-88.

Foerster, T., Brühl, A., & Schäffer, B. (2011). RESTful Web Processing Service. In *Proceedings of the AGILE 2011 Conference on Geographic Information Science*, Utrecht, The Netherlands.

Fook, K.D., Monteiro, A.M.V., Câmara, G., Casanova, M.A., & Amaral, S. (2009). Geoweb Services for Sharing Modelling Results in Biodiversity Networks. *Transactions in GIS*, *13*(4), 379-399.

Friis-Christiensen, A., Lucchi, R., Lutz, M., & Ostländer, N. (2009). Service chaining architectures for applications implementing distributed geographic information processing. *International Journal of Geographical Information Science*, *23*(5), 561-580.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design Patterns: Elements of Reusable Software Architecture*. Reading: Addison-Wesley.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W.J., & Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, *15*(10), 1451-5.

Giuliani, G., Nativi, S., Lehmann, A., & Ray, N. (2012). WPS mediation: An approach to process geospatial data on different computing backends. *Computers & Geosciences*, *47*, 20-33.

Goble, C., Wolstencroft, K., Owen, S., Aleksejevs, S., Snoep, J., Krebs, O., Mueller, W., & Rojas, I. (2009). SysMO-DB: A pragmatic approach to sharing information amongst Systems Biology projects in Europe. In *Proceedings of the UK e-Science All Hand Meeting 2009*.

Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., & De Roure, D. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, *38*(suppl 2), W677-W682.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*, R86.

Granell, C., Díaz, L., & Gould, M. (2010). Service-oriented applications for environmental models: reusable geospatial services. *Environmental Modelling and Software*, *25*(2), 182-198.

Granell, C., Díaz, L., Tamayo, A., & Huerta, J. (in press). Assessment of OGC Web Processing Services for REST principles. *International Journal of Data Mining, Modelling and Management*, ArXiv preprint: http://arxiv.org/abs/1202.0723

Granell, C., Díaz, L., Schade, S., Ostländer, N., & Huerta, J. (2013). Enhancing Integrated Environmental Modelling by Designing Resource-Oriented Interfaces. *Environmental Modelling and Software*, *39*, 229-246

Gregersen, J.B., Gijsbers, P.J.A., & Westen, S.J.P. (2007). OpenMI: open modelling interface. *Journal of Hydroinformatics*, *9*(3), 175-191.

Harris, G. (2002). Integrated Assessment and Modelling – Science for Sustainability. In R. Costanza, & S.E., Jørgensen, (Eds.), *Understanding and Solving Environmental Problems in the 21st Century* (pp. 5-17). Amsterdam: Elsevier.

Heath, T, & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

Hill, C., DeLuca, C., Balaji, V., Suarez, M., & Da Silva, A. (2004). The architecture of the Earth System Modeling Framework. *Computing in Science and Engineering*, *6*(1), 18-28.

Holzworth, D.P., Huth, N. I., & de Voil, P. G. (2010). Simplifying environmental model reuse. *Environmental Modelling and Software*, *25*(2), 269-275.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, *34*, W729-W732.

ISO 19110 (2005). ISO 19110:2005, Geographic information -- Methodology for feature cataloguing. Technical Committee ISO/TC 211, International Organization for Standardization, Geneva, Switzerland, http://www.iso.org/iso/catalogue_detail.htm?csnumber=39965

ISO 19156 (2011). ISO 19156:2010, Geographic Information: Observations and Measurements. Open Geospatial Consortium and Technical Committee ISO/TC 211, International Organization for Standardization, Geneva, Switzerland, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=32574.

Jagers, H.R.A. (2010). Linking Data, Models and Tools: An Overview. In *Proceedings of the International Environmental Congress on Environmental Modelling and Software Society (iEMS 2010)*, Otawa, Canada.

Jakeman, A.J., & Letcher, R.A. (2003). Integrated assessment and modelling: features, principles and examples for catchment management. *Environmental Modelling and Software*, *18*(6), 491-501.

Janowicz, K., Broering, A., Stasch, C., Schade, S., Everding, T., & Llaves, A. (2013). A RESTful Proxy and Data Model for Linked Sensor Data. *International Journal of Digital Earth*, *6*(3), 233-254

Jasny, B.R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and Again, and Again. *Science*, *334*(6060), 1225.

Kepler Team (2010). Kepler 2.1 Getting Started Guide. Available at https://kepler-project.org/users/documentation

Knapen, M.J.R., Verweij, P., Wien, J.E, & Hummel, S. (2009). OpenMI – The universal glue for integrated modelling? In *Proceedings of the 18th World IMACS / MODSIM Congress*, Cairns, Australia.

Kiehle, C. (2006). Business logic for geoprocessing of distributed geodata. *Computers & Geosciences*, *32*(10), 1746-1757.

Laniak, G.F., Olchin. G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., & Hughes, A. (2013). Integrated environmental modeling: A vision and roadmap for the future. *Environmental Modelling and Software*, *39*, 3-23.

Lee, C., & Percivall, G. (2008). Standards-based computing capabilities for distributed geospatial applications. *Computer*, *41*(11), 50–57.

Li, X., Di, L., Han, W., Zhao, P., & Dadi, U. (2010). Sharing geoscience algorithms in a Web service-oriented environment (GRASS GIS example). *Computers & Geosciences*, *36*(8), 1060-1068

Lopez-Pellicer, F.J., Rentería-Agualimpia, W., Béjar, R., Muro-Medrano, P.R., & Zarazaga-Soria, F.J. (2012). Availability of the OGC geoprocessing standard: March 2011 reality check. *Computers & Geosciences*, *47*, 13-19

Ludäscher, B., & Goble, C. (2005). Guest Editors' Introduction to the Special Section on Scientific Workflows. *SIGMOD Record*, *34*(3), 3-4.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., & Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039-1065.

Mates, P., Santos, E., Freire, J., & Silva, C.T. (2011). CrowdLabs: Social Analysis and Visualization for the Sciences. In *Proceedings of 23rd International Conference on Scientific and Statistical Database Management (SSDBM)*, Portland, USA

Maué, P., Stasch, C., Athanasopoulos, G., & Gerharz, L.E. (2011). Geospatial Standards for Web-enabled Environmental Models. *International Journal of Spatial Data Infrastructures Research*, *6*, 145-167.

Mazzetti, P., Nativi, S., & Caron, J. (2009). RESTful implementation of geospatial services for Earth and Space Science applications. *International Journal of Digital Earth*, *2*(Supplement 1), 40-61.

McFerren, G., van Zyl. T., & Vahed, A. (2012). FOSS geospatial libraries in scientific workflow environments: experiences and directions. *Applied Geomatics*, *4*(2), 85-93.

McIntosh, B.S., Giupponi, C., Voinov, A.A., Smith, C., Matthews, K.B., Monticino, M., Kolkman, M.J., Crosman, N., van Ittersum, M., Haase, D., Mysiak, J., Groot, J.C.J., Sieber, S., Verweij, P., Quinn, N., Waeger, P., Gaber, N., Hepting, D., Scholten, H., Sulis, A., van Delden, H., Gaddis, E., & Assaf, H. (2008). Bridging the Gaps Between

Design and Use: Developing Tools to Support Environmental Management and Policy. In A.J. Jakeman et al. (Eds.), *Environmental Modelling, Software and Decision Support* (pp. 33-48). Amsterdam: Elsevier.

Mesirov, J.P. (2010). Accessible Reproducible Research. *Science*, *327*(5964), 415-416.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85-93.

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., & Vieglais, D.A. (2012). Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, *11*, 5-1.

Moore, R.V., & Tindall, C.I. (2005). An overview of the open modelling interface and environment (the OpenMI). *Environmental Science & Policy*, *8*(3), 279-286.

Moore, A.D., Holzworth, D.P., Herrmann, N.I., Huth, N.I., & Robertson, M.J. (2007). The Common Modelling Protocol: A hierarchical framework for simulation of agricultural and environmental systems. *Agricultural Systems*, *95*, 37-48.

Murray, N., Perraud, J.-M., Rahman, J., Seaton, S., Hotham, H., Watson, F., Bridgart, R., & Davis, G. (2007). TIME Reference Manual 4.3. Available at http://www.toolkit.net.au/time

Nativi, S., Mazzetti, P., & Geller, G.N. (2013). Environmental model access and interoperability: The GEO Web Model initiative. *Environmental Modelling and Software*, *39*, 214-246.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glove, K., Pocok, M.R., Wipat, A., & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*, 3045-3054.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K. Goble, C. Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., & Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, *18*, 1067–1100.

OpenMI Association Technical Committee (2010). OpenMI Documents Series: OpenMI Standard 2 Specification. The OpenMI Association, http://www.openmi.org/reloaded.

Papazoglou, M.P., Traverso, P., Dustdar, S., & Leymann, F. (2007). Service-Oriented Computing: State of the Art and Research Challenges. *Computer*, *40*(11), 38-45.

Papazoglou, M.P. (2008). *Web Services - Principles and Technology*. Prentice Hall

Parker, P., Letche, R., Jakeman, A., Beck, M.B., Harris, G., Argent, R.M., et al. (2002). Progress in integrated assessment and modelling. *Environmental Modelling and Software*, *17*(3), 209-217.

Pautasso, C., Zimmermann, O., & Leymann, F. (2008). RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision. In *Proceeding of the 17th International World Wide Web Conference (WWW2008)*, Bejing, China.

Rahman, M., Ranjan, R., Buyya, R., & Benatallah, B. (2011). A taxonomy and survey on autonomic management of applications in grid computing environments. *Concurrency and Computation: Practice and Experience*, *23*(16), 1990-2019.

Rizzoli, A.E., Leavesley, G., Ascough, J.C., Argent, R.M., Athanasiadis, I.N., Brilhante, V., Claeys, F.H.A., David, O., Donatelli, M., Gijsbers, P., Havlink, D., Kassahum, A., Krause, P., Quinn, N.W.T., Scholten, H., Sojda, R.S., &

Villa, F. (2008). Integrated Modelling Frameworks for Environmental Assessment and Decision Support. In A.J. Jakeman et al. (Eds), *Developments in Integrated Environmental Assessment* (pp. 101-118). Amsterdam: Elsevier.

Rotmans, J., & van Asselt, M. (2001). Uncertainty management in integrated assessment modelling: towards a pluralistic approach. *Environmental Monitoring and Assessment*, *69*(2), 101-130.

Santos, E., Lins, L., Ahrens, J., Freire, J., & Silva, C. (2009). VisMashup: Streamlining the Creation of Custom Visualization Applications. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1539-1546.

Schade, S., Ostländer, N., Granell, C., Schulz, M., McInerney, D., Dubois, G., et al. (2012). Which Service Interfaces fit the Model Web? In *Proceedings of the 4th International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2012)*, Valencia, Spain (pp. 1-6)

Schut, P. (2007). OpenGIS Web Processing Service 1.0.0. OpenGIS Standard 05-007r7. Open Geospatial Consortium

Silva, C.T., Anderson, E.W., Santos, E., & Freire, J. (2011). Using VisTrails and Provenance for Teaching Scientific Visualization. *Computer Graphics Forum*, *30*(1), 75-84.

Stenson, M.P., Littleboy, M., & Gilfedder, M. (2011). Estimation of water and salt generation from unregulated upland catchments. *Environmental Modelling and Software*, *26*(11), 1268-1278.

Tamayo, A., Granell, C., & Huerta, J. (2012). Measuring Complexity in OGC Web Services XML Schemas: Pragmatic Use and Solutions. International Journal of Geographical Information Science, 26(6): 1109-1130.

Taverna Team, 2009. Taverna 2.x Documentation. Available at http://www.taverna.org.uk/documentation/

Taylor, I., Shields, M., Wang, I., & Rana, O. (2003). Triana Applications within Grid Computing and Peer to Peer Environments. *Journal of Grid Computing*, *1*(2), 199-217.

Taylor, I., Shields, M., Wang, I., & Harrison, A. (2006). Visual Grid Workflow in Triana. *Journal of Grid Computing*, *3*(3-4), 153-169.

Triana Team, 2009. Triana User Guide. Available at: http://www.trianacode.org/documents.html.

Turuncoglu, U.U., Murphy, S., DeLuca, C., & Dalfes, N. (2011). A scientific workflow environment for Earth system related studies. *Computers & Geosciences*, *37*(7), 943-952.

Verweij, P.J.F.M., Knapen, M.J.R., de Winter, W.P., Wien, J.J.F., te Roller, J.A., Sieber, S., Jansen, J.M.L. (2010). An IT perspective on integrated environmental modelling: The SIAT case. *Ecological Modelling*, *221*(18), 2167-2176.

VisTrails Team (2011). VisTrails Documentation Release 2.0.0. Available at http://www.vistrails.org/index.php/Downloads

Voinov, A., & Shugart, H.H. (2013). 'Integronsters', integral and integrated modelling. *Environmental Modelling and Software*, *39*, 149-158

Voss, A., & Procter, R. (2009). Virtual research environments in scholarly work and communications. *Library Hi Tech*, *27*(2): 174-190.

Wiederhold, G. (1992). Mediators in the architecture of future information-systems. *Computer*, *25*(3), 38–49.

Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., & Snoep, J.L. (2011). The SEEK: A Platform for Sharing Data and Models in Systems Biology. In D. Jameson, M. Verma, & H.V. Westerhoff (Eds.), *Methods in Enzymology* (pp. 629-655). San Diego: Academic Press.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial Cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264-277.

Yu, J., & Buyya, R. (2005). A Taxonomy of Scientific Workflow Systems for Grid Computing. *SIGMOD Record*, *34*(3), 44-49.

Yu, G.E., Zhao, P., Di, L., Chen, A., Deng, M., & Bai, Y. (2012) BPELPower – A BPEL execution engine for geospatial web services. *Computers & Geosciences*, *47*, 87-101.

| # | Viewpoints / IM tools | References |
|---|---|---|
| **18** | **Component-based modelling frameworks** | |
| 4 | Background | Argent (2004) Argent et al. (2006) Jagers (2010) Rizzoli et al. (2008) |
| 5 | OpenMI | Moore and Tindall (2005) Gregersen et al. (2007) Knapen et al. (2009) OATC (2010) Donchyts et al. (2010) |
| 1 | CCA | Bernholdt et al. (2006) |
| 2 | ESMF | Hill et al. (2004) ESMF Joint Specification Team (2011) |
| 3 | CMP | Moore et al. (2007) CMP (2008) Holzworth et al. (2010) |
| 1 | BioMA/APES | Donatelli et al. (2010) |
| 2 | TIME | Murray et al. (2007) Stenson et al. (2011) |
| **20** | **Scientific Workflows Systems** | |
| 5 | Background | Deelman et al. (2009) Ludäscher and Goble (2005) Cohen-Boulakia and Lesser (2011) Yu and Buyya, (2005) Rahman et al. (2011) |
| 2 | Kepler | Ludäscher et al. (2006) Kepler Team (2010) |
| 5 | Taverna | Oinn et al. (2004) Oinn et al. (2006) Hull et al. (2006) Taverna Team (2009) Bhagat et al. (2010) |
| 4 | Triana | Taylor et al. (2003) Taylor et al. (2006) Churches et al. (2006) Triana Team (2009) |
| 4 | VisTrails | Callahan et al. (2006) Santos et al. (2009) Silva et al. (2011) VisTrails Team (2011) |
| **12** | **Virtual Research Environments** | |
| 1 | Background | Voss and Procter (2009) |
| 3 | myExperiment | De Roure et al. (2009) Goble et al. (2010) Bechhofer et al. (2013) |
| 1 | CrowdLabs | Mates et al. (2011) |
| 2 | SysMO | Goble et al. (2009) Wolstencroft et al. (2011) |
| 2 | Galaxy | Giardine et al (2005) Goecks et al. (2010) |
| 3 | DataONE | Michener and Jones (2012), Michener et al. (2012), Bach et al. (2012) |
| **18** | **Service-based modelling** | |
| 4 | Background | Papazoglou et al. (2007) Lee and Percivall (2008) Friis-Christiensen et al. (2009) Yang et al. (2010) |
| 4 | WSDL/WPS | Christensen et al. (2001) Schut (2007) Schade et al. (2012) Lopez-Pellicer et al. (2012) |
| 3 | WSDL-based chaining | Li et al. (2010) Yu et al. (2012) de Jesus et al. (2012) |
| 7 | WPS-based chaining | Kiehle (2006) Fook et al. (2009) Foerster et al. (2010) Granell et al. (2010) Maué et al. (2011) McFerren et al. (2012) Giuliani et al. (2012). |
| **9** | **Resource-based modelling** | |

| | | |
|---|---|---|
| 4 | Background | Fielding (2000) Pautasso et al. (2008) Granell et al. (2013) Nativi et al. (2013) |
| 5 | OGC RESTful services | Mazzetti et al. (2009) Foerster et al. (2011) Finney and Watts (2011) Janowicz et al. (in press) Granell et al. (in press) |
| **77** | **TOTAL** | |

Table 1. List of eligible papers on IM tools included in the review and grouped by viewpoint

| **Id** | **Criteria names** | **Criteria descriptions** |
|---|---|---|
| *C1* | *Integration strategy* | *Is it easy or difficult to import a third-party model or component into a model execution environment? Does it depend on how the third-party model has been designed? Or how model execution environments have been implemented? Or both?* |
| C1.1 | Environment integration strategy | Integration strategies from the target environment side, i.e., white-box and black-box strategies. |
| C1.2 | Third-party integration strategy | Integration strategies from the third-party model side, i.e., full-integration, encapsulation, and mediation strategies. |
| *C2* | *Integration process* | *What are the needed steps to make a third-party component compatible with a target execution environment?* |
| C2.1 | Data model | It indicates what kind of data models is considered first-class citizens in a target execution environment. |
| C2.2 | Integration method | It is concerned with the necessary steps or procedures to actually extend an execution environment with a third-party component or model. |
| C2.3 | Programming language | It simply indicates the set of programming languages supported, which is important in the migration process. |
| *C3* | *Re-usability* | *How much is a third-party model reusable in multiple model execution environments?* |
| C3.1 | Composition method | It refers to how two components are composed to create an IM product. |
| C3.2 | Provenance support | It refers to the ability to annotate, either automatic or manual, workflows or integrated models to support re-usability and reproducibility. It means whether a user may find out how a |

| | | |
|---|---|---|
| | | particular resource has been processed or manipulated. |
| C3.3 | Domain-specific or generic | It refers to the scope of the IM solution, i.e., whether it targets to a specific domain or it supports cross-domain scenarios. |
| C3.4 | Supporting tools and documentation | It simply indicates some extra pieces of information of the IM solution such as availability of source code, supporting tools, documentation, tutorial, etc. |

Table 2. List of evaluation criteria

| **IM tool** | **C1: Integration strategy** | | **C2: Integration process** | | | **C3: Reusability** | | |
|---|---|---|---|---|---|---|---|---|
| | **C1.1** | **C1.2** | **C2.1** | **C2.2** | **C2.3** | **C3.1** | **C3.3** | **C3.4** |
| OpenMI | White-box. | Full Integration. | Extendible base interfaces (Engine, ILinkableComponent). | Interface implementation. | Java, .NET (C#). | Base interfaces implementation (IExchangeItem, Link).<br><br>Linking by connecting and filtering IO data streams between components. | Generic, but originally designed for hydrology. | Source code available in SVN server maintained by OATC.<br><br>Front-end tools and simulation engine available.<br><br>Documentation, tutorials, etc. available online.<br><br>http://www.openmi.org/ |
| CCA | White-box. | Full Integration. | Extendible base interfaces (Component, Services). | Interface implementation + interface description language (SIDL). | Fortran 90,<br>C , C++ | Base interfaces implementation (Port).<br><br>Linking by using and providing *ports* between components. | Generic, but designed for high-performance computing. | Source code available.<br><br>Various CCA tools available in a single tar file.<br><br>Documentation, tutorials, etc. available online.<br><br>http://www.cca-forum.org/ |
| ESMF | White-box. | Full Integration. | Customized predefined model (Gridded Component, Coupler | Interface implementation. | Fortran 90, C++ | Customized data entities implementation (State, Coupler Component).<br><br>Exchanged data are | Specific on climate simulation. | Source code available in SVN server.<br><br>Documentation, tutorials, etc. available online. |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Component). | | | instances of State. Couplers enable aggregation of gridded components. | | http://www.earthsystemmodeling.org/ |
| CMP | White-box. | Full Integration. | Customized predefined model (System, Components, Property) | Interface implementation + XML-based document configuration (SMDL). | C++, .NET (VB, C#) | Customized data entities (Message, Event).<br><br>XML-based document description (SMDL). | Specific on agricultural simulation. | No code available<br><br>No dedicated web site, but some documentation available<br><br>http://www.grazplan.csiro.au |
| BioMA (APES) | Black-box. | Full Integration. | Extendible Base interfaces (IStrategyComponent, DomainClass) | Object composition + shared ontologies terms. | .NET (C#) | Extendible base interfaces (Simple, composites, and context Strategies).<br><br>Object composition: Aggregation of simple strategies into composite and context strategies + Semantic IOPE. | Specific on crop simulation. | Source code available as zip file (but link currently disabled)<br><br>Component utilities available (but link currently disabled)<br><br>Documentation, help, etc. available online.<br><br>http://agsys.cra-in.it/tools/bioma/help |
| TIME | White-box. | Full Integration. | N/A | N/A | .NET (VB, C#, Fortran 95) | N/A | Specific on hydrological catchment models. | Source code available in SVN server maintained by CSIRO.<br><br>Front-end tools available.<br><br>Documentation, tutorials, etc. available online<br><br>http://www.toolkit.net.au/tools/Time |

Table 3. Comparison matrix for component-based modelling frameworks based on the criteria in Table 2

| **IM tool** | **C1: Integration strategy** | | **C2: Integration process** | | | **C3: Reusability** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **C1.1** | **C1.2** | **C2.1** | **C2.2** | **C2.3** | **C3.1** | **C3.2** | **C3.3** | **C3.4** |
| Kepler | White- | Full | Extendible | Actor-oriented | Java | Extendible | Good, data | Generic but | Source and binary code available from |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | box. | Integration + Encapsulation. | base interface (Actor, Parameter). | programing interfaces.<br><br>Support for Web services, OGC services, Cloud. | | base interface (Port, Channel)<br><br>Directors (control flow).<br><br>Reuse by composite actors or sub-workflows | provenance as part of the workflow. | strong focus on molecular processing and biology. | dedicated web site.<br><br>Front-end tools available.<br><br>Documentation, tutorial, mailing list, etc. available online.<br><br>http://kepler-project.org |
| Taverna | White-box. | Full Integration + Encapsulation. | A kind of "service client" or "service type" (Processors). | WSDL-based services or services at BioCatalogue<br><br>Processors + XML-based document description (SCUFL)<br><br>Support for Web services, OGC services, Cloud, Grid. | Java | Data Links between processors.<br><br>XML-based document description + Coordination constraints (control flow).<br><br>Reuse by nested workflow type processors. | Very good, well connected to my Experiment<br><br>Focus on workflows, services, inputs, outputs. Data must be handled outside the system. | Generic but strong focus on life sciences and bioinformatics. | Source and binary code available from dedicated web site.<br><br>Front-end tools available.<br><br>Documentation, tutorial, mailing list, etc. available online.<br><br>http://www.taverna.org.uk |
| Triana | White-box. | Full Integration + Encapsulation. | A kind of "service client" or "service type" (Units). | WSDL-based services (UDDI repositories) or P2P services<br><br>Support for Web services, P2P, Grid, WS-RF | Java | Data links between units.<br><br>Control flow by built-in toolbox.<br><br>Reused by publishing workflows as a WS in UDDI | Good, data provenance as part of the workflow. | Data mining, statistics. | Source code available from SVN server. Binary code form dedicated web site.<br><br>Front-end tools available.<br><br>Poor documentation and tutorials.<br><br>http://www.trianacode.org |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | repositories. | | | |
| VisTrails | White-box. | Full Integration. | Extendible base interfaces (Modules). | Module-oriented programing interfaces.<br><br>Initial support for OGC services (via extension). | Python | Data links (ports, connections) between modules.<br><br>Control flow by built-in modules.<br><br>Reused by aggregating related workflows as a *vistrail.* | Very good, focus on visual products, versioning, history of collection of workflows, etc.<br><br>See also Crowdlabs. | Generic but strong focus on visual-centric apps such as medical imaging, simulation, geosciences, etc. | Source and binary code available from dedicated web site.<br><br>Front-end tools available.<br><br>Documentation, tutorial, mailing list, etc. available online.<br><br>http://www.vistrails.org |

Table 4. Comparison matrix for scientific workflow systems based on the criteria in Table 2

| **IM tool** | **C1: Integration strategies** | | **C2: Integration process** | | **C3: Reusability** | | | |
|---|---|---|---|---|---|---|---|---|
| | **C1.1** | **C1.2** | **C2.1** | **C2.4** | **C3.1** | **C3.2** | **C3.3** | **C3.4** |
| myExperiment | White-box. | Mediation. | It mainly supports Taverna workflows, but other workflow systems may be integrated | *Research objects:* Aggregated and individual resources such as workflows, metadata, results, etc. | Delegated to associated systems (e.g. Taverna)<br><br>Aggregation of research objects | Users can comment and tag workflows.<br><br>Workflow execution produces provenance graph that describes the datasets and processes involved in generating an output | Bioinformatics, Life sciences | Very complete. Lots of tutorials and information<br><br>http://www.myexperiment.org/ |
| Crowdlabs | White-box. | Encapsulation. | It supports VisTrails workflows | *Vistrails workflows:* Aggregated and individual resources. | Delegated to VisTrails system | Yes | Medical imaging, simulation, geosciences | Poor<br><br>http://www.crowdlabs.org/ |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SysMO-DB | White-box. | Mediation. | It provides unified model (JERM) to support any source data model | *Assets:* workflows, datasets, standard operating procedures, publications, users. | Closely related to MyExperiment | Yes. Ontology annotation in data files and JERM data model | Molecular Biology | Complete<br><br>http://www.sysmo-db.org/ |
| Galaxy | White-box. | Full Integration. | Local components inherited from "Galaxy" interfaces. | *Galaxy objects*: Workflows, datasets, histories, and Pages<br><br>Histories can be seen as abstract workflows. | Aggregation of Galaxy objects | Yes. Users can comment and tag Galaxy objects.<br><br>Galaxy Pages & Histories (documentation). | Genomics | Very Complete. Lots of tutorials, demos and related user info<br><br>http://galaxyproject.org/ |
| DataONE | White-box. | Full Integration. (nodes as a System of Systems approach). | New datasets are fully integrated in the federation of nodes. | Datasets**:** main focus on long-term preservation of data. | Delegated to an Investigator Toolkit | Users provide metadata records on datasets | Earth Sciences | Very complete<br><br>http://www.dataone.org/ |

Table 5. Comparison matrix for virtual research environments based on the criteria in Table 2