20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Data integration technologies to support integrated modelling

MJR. Knapen[a], O. Roosenschoon[a], R. Lokers[a], S. Janssen[a], Y. van Randen[a] and P. Verweij[a]

[a] *Alterra, Environmental Sciences Group, Wageningen University and Research Centre, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands*
Email: rob.knapen@wur.nl

Abstract: Over the recent years the scientific activities of our organisation in large research projects show a shifting priority from model integration to the integration of data itself. Our work in several large projects on integrated modelling for impact assessment studies has clearly shown the importance of data availability for integrated modelling, but of no less importance is the integration, or alignment, of the required input data itself. Moving from the fairly technical model integration in OpenMI and OpenMI related projects, and moving towards basic semantic integration in the SEAMLESS and SENSOR projects, our focus is now shifting towards researching and applying techniques such as Semantic Web technologies to improve data discoverability, its integration, and in the future on reasoning about the constructed integrated knowledge. This paper will present an overview of the on-going work in our European 7th Framework Programme (FP7) project TREES4FUTURE, focussing on automated harvesting of forestry related data sets and enriching its meta data for search ability; the FP7 LIAISE Network of Excellence on linking impact assessment instruments such as models and data to sustainability expertise; and the FP7 research project SEMAGROW on developing visions on processing and querying large RDF triple-stores of integrated agricultural data. In the end we aim at bringing the results of all these projects together to achieve a next step in integrated modelling and to present ways to use Natural Language Processing based methods to help providing meta data.

Keywords: *Integrated Modelling, Data integration, Semantic, Web of Data, Natural language Processing*

1. INTRODUCTION

Integrated modelling - a set of interdependent science-based components (models, data, and assessment methods) that together form the basis for constructing an appropriate modelling system (Laniak et al., 2012) - from a technological perspective over the years has shown to revolve around two central issues: finding ways to connect scientific models from different domains; and finding the proper input data for these connected scientific models to apply them for new studies, sometimes in new regions. The integrated models can be helpful for decision makers to evaluate ex-post or assess ex-ante the impacts of their choices. The models provide a simplified representation of reality and can simulate potential contrasting pathways into the future, thus improving the understanding of interdisciplinary cause-and-effect relationships.

Input data for scientific models typically are described by some form of meta-data, providing information about the characteristics of the data and its content. Unfortunately the storage format, structure, completeness and accuracy of such meta-data varies enormously, and finding available suitable data followed by required pre-processing to make the data usable as input is very time-consuming.

For long our organisation has been working in the field of environmental impact assessment, and supporting it with Information Technology. We followed a typical path of wrapping scientific models with graphical user interfaces, building more complex decision support systems (DSS) around them, hard connecting two or more models inside such a DSS, developing and working with internal frameworks for model linking, and contributing and applying open standards for it (e.g. OpenMI).

Last year we examined the use of OpenMI (Moore & Tindall, 2005) in several larger EU projects our organisation has contributed to and noted that the technology definitely can help, but also still has problems with adaptation outside of its original domain (hydrology) and in more exploratory scientific settings. Technically it works and certainly provides benefits, but the researchers do not like its restrictiveness and do not like losing control over the data that with OpenMI is automatically exchanged between the scientific models. As it turns out they most of the time prefer a “soft-linking” approach, i.e. manual (perhaps scripted) exchange of data between the models. From one point they clearly like to stay in control of what happens in-between the models and do not trust it when it is too much a black box to them. But also, even though at a technical level the data can be exchanged between the models, conceptually it might make no sense and need more expertise than a computer can currently provide to validate it. In other words, we are at risk of creating “integronsters”, constructs that are perfectly valid as software products but ugly or even useless as models (Voinov & Shugart, 2013).

Interesting enough, and in line with the “soft-linking” approach, the most successful and re-used component of the SEAMLESS project (Van Ittersum et al., 2008) is the integrated data-base that was developed and filled during the course of the initial project and in follow-ups. The developed DSS for researchers and decision makers is mainly used for educational purposes. With OpenMI and similar frameworks the technical concepts of linking models are basically standardized. The process for accepting OpenMI as an OGC (http://www.opengeospatial.org) standard is nearing completion, so it seems to be an appropriate time to take integrated modelling one step further and start looking more at improving *data integration technologies*. While we have most been focussing on supporting integrated modelling with component-based modelling approaches, other approaches like service-based modelling, a hybrid approach, and the resource-oriented approach being research (Granell et al., 2013) all are pointing towards the need for more attention to the use of the technologies that are driving the development of the Web of Data for Integrated Environmental Modelling (IEM). Particularly its open nature provides huge benefits in a field that depends on the use of scientific models and data from different research domains. For this the work on Spatial Data Infrastructures (SDIs) is relevant but not sufficient. A non-SDI expert still has problems accessing and deciphering the data due to the inherent complexity of geospatial data standards (Tamayo et al., 2012); the lack of support for proper connections and linkages between geospatial data and services; and the diversity of interaction paradigms (Granell et al., 2013).

Projects our institute is currently participating in, like TREES4FUTURE (http://www.trees4future.eu), LIAISE (http://www.liaise-noe.eu), and SEMAGROW (http://www.semagrow.eu), all use or research one or more of the building blocks of the Web of Data to improve connecting data to other data, and data to scientific models.

2. WEB OF DATA

In recent years the Web (the Internet) has evolved from a global information space of linked documents to one where both documents and data are linked. Traditionally most data published on the Web has been made

available as raw dumps in formats such as CSV (comma separated values), and the relation between linked documents has been implicit since the used HTML data format lacks the expressiveness to connect the individual entities described in the documents. This is now being addressed with a set of best practices for publishing and connecting structured data on the Web, known as Linked Data. This term basically refers to data published on the Web in such a way that it is machine-readable, its meaning explicitly defined, is linked to other external data sets, and itself can be linked to from external data sets.

Linked Data relies on RDF (Resource Description Format) to describe typed statements that link arbitrary things in the world (Web of Data, or Web of things in the world). Besides the use of RDF there are a few simple principles for publishing Linked Data on the Web: (i) Use URIs as names for things; (ii) Use HTTP URIs so that people can look up those names; (iii) When someone looks up a URI, provide useful information using the standards (RDF, SPARQL); and (iv) Include links to other URIs, so that they can discover more things (Heath & Bizer, 2011).

RDF itself provides a graph-based data model with which to structure and link data that describes things in the world. It encodes data in the form of subject, predicate, object triples. The subject and object of a triple are both URIs that each identify a resource, or a URI and a string literal respectively. The predicate specifies how the subject and object are related, and is also represented as a URI. When the subject and the object URIs are references to namespaces of different datasets this forms a RDF link (Bizer et al., 2009).

The OWL (Web Ontology Language) and RDFS (RDF Vocabulary Definition Language) provide a basis for creating vocabularies (collections of classes and properties) that can be used to describe entities and how they relate. Anyone is free to publish vocabularies to the Web of Data, and RDF triples can be used to link or define the mappings between the classes and properties of these vocabularies.

Linked Data typically is accessible with SPARQL (pronounced "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language). This is an RDF query language, able to retrieve and manipulate data stored in Resource Description Framework format. It was made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is considered as one of the key technologies of the semantic web. SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns (Wikipedia, 2013). Knowledge bases are made accessible, i.e. humans and machines can query them using the SPARQL language, as SPARQL endpoints.

The Linked Open Data project (http://linkeddata.org) gives a good visible example of the adaptation and application of the Linked Data principles. The on-going aim of this project is to identify existing data sets available under open licenses, convert them to RDF according to the Linked Data principles, and publish them on the Web. There is also increasing interest in publishing scientific data as Data Papers or Data Journals, making them citable thus giving proper credits to authors, and promoting access and re-use.

3. PROJECTS

The technologies behind the Web of Data are playing an increasing role in many current projects of our organisation. No matter how it is approached, policy impact assessment has to deal with data from multiple domains and it requires this data to be connected. While component- and service-based modelling help the work of linking the models in a technical sense, they still do not seem to sufficiently address or support connecting the models and the data in a semantic sense. This still requires lots of attention from the researchers involved. Using Web of Data technologies with their inherent open and standardised nature certainly looks like a sensible step to take. The projects described here give an overview of current work that all somehow relates to integrating (or linking) data, and from which at some point in time results hopefully can be merged to leverage the individual progress into a bigger picture.

3.1. LIAISE

The EU FP7 network of excellence LIAISE (www.liaise-noe.eu) is designed to identify the causes for non-use of Impact Assessment (IA) tools and bridge the gaps between researchers with a generally strong orientation towards their (disciplinary) peers and practitioners who tend to focus on their policy domain and policy problems. LIAISE aims at: 1) understanding of the policy process and the resulting needs for IA knowledge and IA tools; 2) description of IA tools and scientific IA expertise in a standardized way; 3) a shared IA toolbox targeted at the needs of both researchers and practitioners; 4) a shared IA research agenda integrating scientific knowledge gaps and the priorities for the development of new IA knowledge that arise from the future policy agenda; and 5) safeguarding the project results beyond the period of project funding,

by developing an institutional setting and a business plan that facilitate the extension of the present consortium towards a broad center of IA expertise with a structural permanence.

The LIAISE Shared Toolbox

The shared toolbox (Roosenschoon et al., 2012) is the LIAISE facility where IA users directly interact with the information and tools that support them in conducting their impact assessments. It contains descriptions of different types of knowledge that can be used in the context of policy Impact Assessment. This includes a library of scientific models, IA methods, good practices, and IA experts. These sources of knowledge are described and can be searched using key phrases from the policy IA domain. For the different categories against which knowledge source are described, taxonomies have been developed. These are both from the sphere of research (e.g. modeling technique) as well as from the sphere of policy making (e.g. impact areas). The taxonomies are applied to describe all objects included in the shared toolbox, thereby making it possible to query it and search for suitable resources from different perspectives depending on the users' needs.

To extend the reach of the shared toolbox (currently stored in a Drupal CMS) and to ensure future accessibility LIAISE finds it desirable to (i) publish the content as Linked Open Data (LOD) as part of the Web of Data, and (ii) consume other LOD as references and new toolbox content. For this the Toolbox content will be made available as RDF, accessible through a SPARQL Endpoint. Since the toolbox content is stored in a Drupal CMS, but without making us of its emerging build-in RDF capabilities, it appears to be most viable to follow a data extraction – RDF-izing – static publishing route. This will use e.g. a script for processing the toolbox content and use a mapping definition to generate one or more static RDF files from it. These files can then be made available for download, and imported into a triple store to make the data available through a standard SPARQL endpoint.

LIAISE and Natural Language Processing

Having taxonomies and ontologies however only is a start. The, mostly not really difficult, but very laborious work of manually providing all the metadata for each simulation model and every data set seems to be a much bigger hurdle. For example, LIAISE Focus Group meetings found that people were hesitant to provide all this requested metadata information. When machine-readable metadata already exists part of the work can be automated. Currently however this is limited to specific types of data sets where still the metadata can be filled in very sparsely. Yet, scientific papers, reports, manuals, web sites, leaflets, and other forms of documentation exist. These unstructured text source however do not present the metadata in a way that can directly be processed by a computer.

LIAISE will investigate whether and how Natural Language Processing (NLP) techniques can be used to automatically derive the required metadata from unstructured text sources and relate it to the LIAISE ontology defined by the project, to support answering search questions for simulation models and data sets. This could lead to a Toolbox which does not only rely on active provision of meta-info by "real" people, but which also gets its content from automated discovery of relevant meta-info. Expected results are an initial evaluation based on a study of existing literature and currently available technologies. A second result is the definition of a few test cases describing how NLP could be applied and to what purpose. Based on that a prototype / proof-of-concept system will be build and tried and evaluated. Figure 1**Error! Reference source not found.** illustrates the global processing steps and flow:

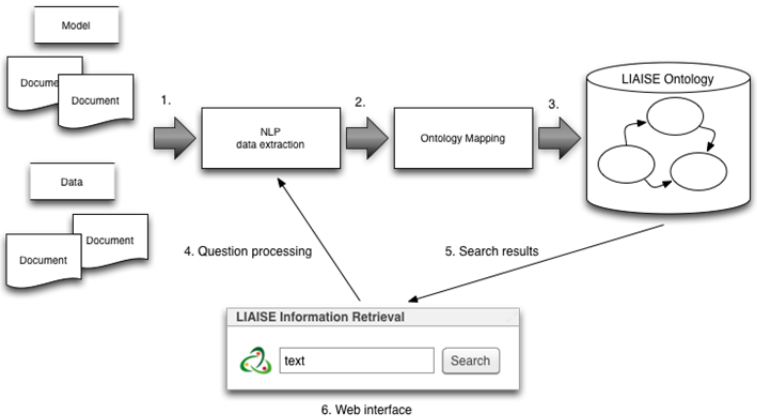

Figure 1: Possible NLP workflow for LIAISE

Documentation for simulations models and data sets will be selected and processed using NLP techniques (step 1). Outcomes will e.g. be relevant key phrases and other information that is relevant based on the LIAISE Reference Model. This data will be mapped to the defined ontology (step 2), and stored (step 3). For retrieval through a web interface (6) or from the Front Office website, questions posed in natural language will be processed (step 4), related to the stored information and used to find (step 5) matching search results (i.e. models and data sets).

3.2. TREES4FUTURE

TREES4FUTURE is an Integrative European Research Infrastructure project that aims to integrate, develop and improve major forest genetics and forestry research infrastructures. It will provide the wider European forestry research community with easy and comprehensive access to currently scattered sources of information (including genetic databanks, forest modeling tools and wood technology labs) and expertise.

This will help forestry researchers and the European forestry sector to respond, in a sustainable manner, to increasing demands for wood products and services in a context of genetic adaptation and changing climatic conditions. It will create a new and better-linked research infrastructure, which will increase our knowledge about the adaptation of forests to climate change, and tree characteristics suited for tailor-made wood supply - thus optimizing the short- and long-term exploitation of forest resources. (www.trees4future.eu).

TREES4FUTURE Clearinghouse

In the forestry domain many datasets are available, however knowledge of the existence and access to these datasets is hampered as they are scattered among different organisations and individuals. As part of the work done in the project a common web enabled access point (a clearinghouse) is build for the data relevant in the forestry domain for Europe. This clearinghouse will enable the searching for- and discovering of datasets using a common reference framework (i.e. an ontology) for the domain.

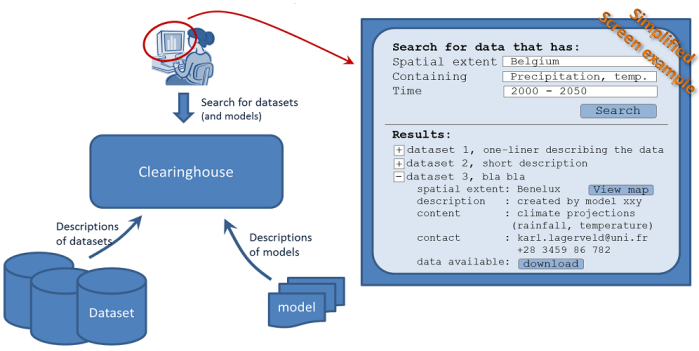

Figure 2: TREES4FUTURE Clearinghouse simplified screen example

Error! Reference source not found. Figure 2 shows an example of possible clearinghouse functionality: the user can enter search criteria (e.g. 'Belgium', 'precipitation, temp.' and '2000-2050'), press the 'Search' button and is then presented a list of datasets fitting the search criteria. The clearinghouse gets this information via a number of potential sources: (*i*) existing dataset descriptions published via a catalogue server; (ii) existing dataset description in vendor specific formats; and (*iii*) existing dataset without descriptive information. Depending on the characteristics of the source the descriptive information can (*i*) be fully automatically incorporated into the clearinghouse, (*ii*) semi-automatically transferred, or (*iii*) requires manual information entry.

TREES4FUTURE Ontology

Although the ontology for TREES4FUTURE started out with terms specific to the forestry domain, based on what the modellers, researchers and domain experts defined, this was considered too restrictive. The current ontology is more flexible with central concepts like 'data file', 'model' and 'attribute. A data file quantifies or qualifies one or more attributes. Data attributes may have temporal and/or spatial dimensions. A model has a number of model attributes. Model attributes can describe input-, output- or model controller/behaviour attributes. Attributes have traits/aspects like unit, spatial reference and precision. The following five core concepts cover the clearinghouse domain:

- Dataset – a collection of data (e.g. 'Soil')

- Model – a computer model producing one, or multiple output attributes and that might require one, or multiple input attributes in order to be executed (e.g. 'EfiScen')
- Attribute – a characteristic of the content of a dataset, and/or of the input- and or outputs from a model (e.g. 'pH', or 'texture').
- Taggable – an abstraction of dataset, model and attribute that can have multiple tags of the form key-value. E.g. an attribute can have a tag 'name=thickness' and another tag 'unit=mm'
- Tag – a key-value pair (e.g. 'unit' = 'mm')

At conceptual level there is a high similarity (almost an equality) between model and dataset. In this ontology Models, Datasets and Attributes are considered as a specialisation of the taggable concept. Taggables can have any number of tags. Each tag describes one aspect (single or complex) of the concept it is assigned to. One tag can be assigned to multiple concepts. Tags can contain simple values (numbers, dates, strings) or can contain a link to an external ontology or vocabulary.

Researching the use of tags as part of this more flexible ontology approach for annotating datasets and models and later being able to find data suitable as model input should give some insights into the usefulness of this method. Tags can be added more "voluntary", e.g. as a community effort. This is similar to the work being done in the EU FP7 TATOO project (http://www.tatoo-fp7.eu) (Pariente et al., 2011).

3.3. SEMAGROW

SEMAGROW (www.semagrow.eu) is a EU 7TH Framework funded ICT project that is researching the technologies needed to keep up with the expanding growth of the Linked Open Data cloud. The trend to open up data and provide them freely on the Internet has intensified in volume as well as quality and value of the data made available. The linked data community is grasping the opportunity to combine, cross-reference, and analyse unprecedented volumes of high-quality data and using it the build innovative applications. It is clear that it needs to be accepted that some schemas might be better suited to a given dataset and application and that there is no consensus about a "universal" schema or vocabulary for any given application, let alone for the Web of Data and related initiatives such as the LOD cloud. Infrastructure will be needed that besides being efficient, real-time responsive and scalable, is also flexible and robust enough to allow data providers to publish in the manner and form that best suits their processes and purposes, and data consumers to query in the manner and for that best suits theirs.

To address these challenges SEMAGROW is carrying out fundamental databases research and developing methods and infrastructure that will be rigorously tested on large-scale current use cases, provided by FAO (http://www.fao.org/), AgroKnow (http://agroknow.gr/) and Alterra, Wageningen UR, as well as on their projected data growth beyond the project's end, laying the foundations for scalable, efficient, and robust data services needed to take advantage of the data-intensive and inter-disciplinary science of the future. In essence it is researching how to scale up the current Web of Data technologies to support even larger heterogeneous datasets. For this all data, including for some of the use cases data currently stored in NetCDF (http://www.unidata.ucar.edu/software/netcdf/) format, has to be turned into RDF triples ("triplified"), aligned, stored, and indexed for efficient retrieval (Figure 3**Error! Reference source not found.**).

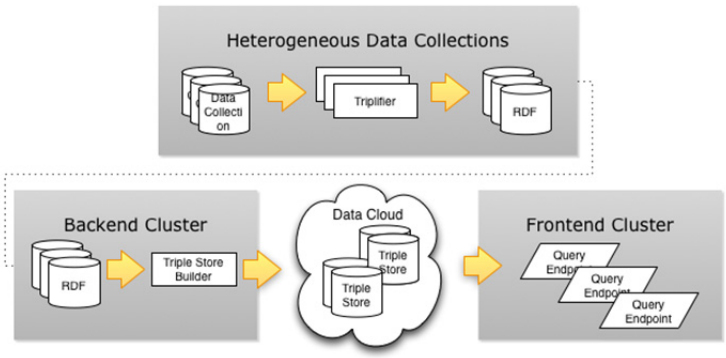

Figure 3: SEMAGROW Heterogeneous data processing

4. DISCUSSION

In the three described projects interesting research is happening that all relates to improving the way data can be stored and searched, tagged with flexible metadata, and published on the Web of Data for integrated

Impact Assessment purposes. Researchers are becoming familiar with using and publishing open data, and application developers with using such data. Meanwhile ICT is already putting promising developments to work (for these projects but also e.g. at Google, or the US National Security Agency).

All these projects are already addressing issues mentioned by (Laniak et al., 2012) that are on the roadmap for IEM. As always with these large EU projects with many international partners involved it is difficult to have a strong overall coordination. And there are many other EU FP7 projects that are related but since they run at the same time it is difficult to cooperate with. Still we are aligning the work we do as part of the mentioned projects, and others like AgMIP (http://www.agmip.org), and combining the efforts.

It will be interesting to see what can be done with the large scale data handling made possible by SEMA-GROW and what new opportunities arise. Data now still kind of hidden in e.g. NetCDF files that is stored as triples and more freely searchable can lead to new kinds of research and perhaps even IA "modelling" or data analysis. With RDF triples based on core IA ontologies defined and maintained by LIAISE, that can align them to other well-known upper ontologies in the environmental domain. While using a flexible tagging system with community support and NLP based assistive tools to help providing the required metadata. Together these technologies, with their development driven by real world Use Cases, can improve integration of the available data and new data becoming available for Integrated Modelling.

REFERENCES

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *5*(3), 1-22.

Granell, C., DíAz, L., Schade, S., OstläNder, N., & Huerta, J. (2013). Enhancing integrated environmental modelling by designing resource-oriented interfaces. *Environmental Modelling & Software*, *39*, 229-246.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, *1*(1), 1-136.

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P. et al. (2012). Integrated environmental modeling: a vision and roadmap for the future. *Environmental Modelling & Software*.

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.

Pariente, T., Fuentes, J. M., Sanguino, M. A., Yurtsever, S., Avellino, G., Rizzoli, A. E. et al. (2011). A model for semantic annotation of environmental resources: the tatoo semantic framework. In *Environmental Software Systems. Frameworks of eEnvironment* (pp. 419-427). Springer.

Roosenschoon, O., Reis, S., Turnpenny, J., Adelle, C., Jacob, K., Wascher, D. et al. (2012). *Bridging the gap between modellers and model users, why does this gap exist and what can we do about it.* In proceedings of the iEMSs Sixth Biennial Meeting: International Congress on Environmental Modelling and Software, 2012.

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.

Van Ittersum, M. K., Ewert, F., Heckelei, T., Wery, J., Alkan Olsson, J., Andersen, E. et al. (2008). Integrated assessment of agricultural systems–A component-based framework for the European Union (SEAMLESS). *Agricultural systems*, *96*(1), 150-165.

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