

International Journal of Spatial Data Infrastructures Research, 2015, Vol.10, 27-54

Describing models in context – A step towards enhanced transparency of scientific processes underpinning policy making*

Nicole Ostlaender¹, Tom Bailly-Salins², Matthew Hardy³, Andrea Perego⁴, Anders Friis-Christensen⁵, Silvia dalla Costa⁶

^{1,4,5,6}European Commission, Joint Research Centre {nicole.ostlaender; andrea.perego; anders.friis; silvia.dalla-costa}@jrc.ec.europa.eu

²European Commission, Joint Research Centreⁱⁱ Tom Bailly-Salins
tbaillys@gmail.com

³European Commission, Joint Research Centre
matthew.hardy@ext.jrc.ec.europa.eu

Abstract

The transparency and reproducibility of scientific evidence underpinning policy is crucial to build and retain trust. This paper describes an application that takes a significant step towards enhanced transparency of scientific models used for policy making: The *Modelling Inventory Database and Access Services (MIDAS)* developed by the Joint Research Centre (JRC) describes models in use by the JRC in their scientific context by linking them to other models, to related data, to supported policies and to domain experts. To effectively share the resulting knowledge across different domains and with policy makers within the institution

*This work is licensed under the Creative Commons Attribution-Non commercial Works 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

DOI: 10.2902/1725-0463.2015.10.art2

ⁱⁱ Tom Bailly-Salins worked at the DG Joint Research Centre at the time of the completion of the described work.

MIDAS represents the resulting complex network of relations and entities through visual aids based on visual analytics and data narratives. This paper describes not just the application in order to contribute to emerging dialogue on best practice for model documentation, it describes the process and main challenges we met with, and the approach taken to overcome them.

Keywords: Transparency, reproducibility, modelling, knowledge representation, open data, open science, visualisation

1. INTRODUCTION

In 2014 Nature and Science dedicated a common editorial to reproducible science, highlighting that “open, transparent and reproducible research is a cornerstone of the scientific method” (Science 2014, Nature 2014). However, especially in the digital age and the use of computer simulations and digital data this proves a challenging task.

The movements for Open Data (OD), Public Sector Information (PSI) and Open Research Data (ORD) form major contributions towards the goal of reproducible science. The key policy drivers of Open Data, starting with the launch of open data government initiatives such as Data.gov and Data.gov.uk, have been “economic growth and business innovation”. Since then the movement has gained tremendous speed, and the list of open data portals for public sector information has been growing ever since. With the 2013 *G8 Open Data Charter* (UK Cabinet Office, 2013) Open Data was globally recognised for its importance and potential impact. Key policy commitments and relevant regulatory basis within the European Union include the EU *Directive on the re-use of public sector information* (Directive 2013/37/EU), the EU implementation of the G8 Open Data Charter as part of the *Digital Agenda for Europe* (COM/2010/245), and the *INSPIRE Directive* (Directive 2007/2/EC).

Open Research Data, describing the concept of open access to scientific data, is also rapidly gaining traction within the scientific community and recognition on the policy level. Here, again, growth and business innovation play an important role: for example the *Europe 2020 strategy* (COM/2010/2020) underlines the central role of knowledge and innovation in generating growth. The related *Commission communication on better access to scientific information* (COM/2012/401) states that “research results, including both publications and data collections, need to be circulated rapidly and widely, using digital media to accelerate scientific

discovery, to enable new forms of data-intensive research and to allow research findings to be systematically taken up by European business and industry”ⁱⁱⁱ. At a concrete level, Horizon 2020 introduces a general principle for open access to scientific publications, which means that all projects receiving Horizon 2020 funding will be obliged to make sure any peer reviewed journal article they publish is openly accessible, free of charge. Furthermore, Horizon 2020 includes a limited scoped pilot for open research data, which supports the goal of improving access to scientific results from projects, in order to support increased transparency, innovation and quality.

The desired full transparency and reproducibility of scientific results, however, go beyond OD and ORD: They require Open Science i.e. open access to and efficient communication of *knowledge* to understand how to frame the scientific question and know what to do with the data. Therefore Open Science also requires opening up and sharing of information on workflows, related entities, processes such as models and algorithms, methods and the context involved in the execution of experiments (Grazzini and Pantisano, 2015). Ongoing activities, such as those of the Research Data Alliance (RDA) Working Group on *Data Citation*^{iv} and RDA Interest Groups on *Data in Context*^v and *Reproducibility*^{vi}, Nature’s special issue on *Challenges in irreproducible research* and the Editorial of Nature and Science in November 2014 (Nature, 2014, Science 2014), all illustrate the current drive to address the challenges ahead.

Sharing this knowledge will not only foster reproducibility and transparency, but enable collaborative, inter-disciplinary research, by enabling scientists from other disciplines to reproduce how a certain problem was solved in one domain, to re-use the results, or to re-produce or re-purpose a process in their own domain. These innovations will help in addressing the Grand Challenges of our century.

In this paper we describe a real world application that takes a significant step towards enhanced transparency of the scientific processes that underpin policy making. Modelling is one of the key expertise areas of the European Joint Research Centre (JRC) as the in-house science service of the European Commission (EC). It is used in various aspects of the policy cycle, from policy

ⁱⁱⁱ Other relevant policy documents on EU level are: Commission communication on a reinforced European research area partnership for excellence and growth (COM/2012/392) and Commission Recommendation of 17 July 2012 on access to and preservation of scientific information (2012/417/EU)

^{iv} <https://www.rd-alliance.org/groups/data-citation-wg.html>, last access: 27th March 2015

^v <https://www.rd-alliance.org/groups/data-context-ig.html>, last access: 27th March 2015

^{vi} <https://www.rd-alliance.org/groups/reproducibility-ig.html>, last access: 27th March 2015

anticipation, formulation, implementation and ad-hoc support to policy evaluation, providing evidence-based scientific support to the main political priorities of the Commission. The application we intend to describe is the *Modelling Inventory Database and Access Services (MIDAS)*, an inventory that contains descriptions of over 160 models in multiple disciplines that are directly or indirectly used by JRC to inform policies of the European Union. Independent of their domain or type, MIDAS describes the models in their scientific context by linking them to other models, to related data, to supported policies and to domain experts, and, wherever possible, providing access to Open Data and other resources, using Persistent Identifiers (PIs). To effectively communicate the resulting complex network of information across different domains, and to a non-technical audience, MIDAS represents the resulting knowledge through visual aids.

The application has been developed by the JRC. We see our approach as an essential first step to ensure the transparency, traceability and accountability of policy supported by models, and MIDAS is, to our knowledge, a unique example in its field. For this reason, the experience described in this paper, including some of the technical and organisational challenges met, is an important contribution to the international effort to increase the reproducibility of science.

The scope of this paper is the application's main principles, the front end, and the process of collecting the information related to the inventory. The more technical aspects like underlying metadata profile, and detailed system architecture will be dealt with in future publications.

The rest of the paper is structured as follows: section 2 addresses related work and in section 3 we summarise the background and scope of the MIDAS system. Section 4 describes our approach by summarising the development process, the major challenges and how we met them, with a focus on the cultural challenges of our work. Section 5 describes the MIDAS system, putting particular emphasis on the visualisation of the complex content it provides. In section 6 we discuss the approach taken, in respect to the challenges we met. We close with a brief conclusion of the challenges and impact of our work.

2. RELATED WORK

The developments we present in this paper touch upon various aspects of open science, including the use of PIs, the nature of model inventories, transparency and reproducibility, and communication and mining of knowledge through visual aids. In this section we briefly elaborate on each of these aspects and some related work.

PIs are identifiers which reliably refer to resources regardless of those resources' physical location or current ownership (Tonkin 2008). Two characteristics are

essential for PIs: Persistence and unambiguity (GBIF 2011). Persistence ensures that an identifier is permanently assigned to a resource. Unambiguity ensures that a resource can be uniquely identified. A PI only becomes useful if information about the resource it represents can be easily retrieved. For this, resolvable PIs can be used. In these cases, a resolver provides the mapping from the PI to the location of the corresponding digital object or related metadata.

Perhaps the best-known example of a PI is the Digital Object Identifier (DOI), which is assigned for identification of traditional publications using a well-established system. The DOI consists of a prefix which identifies the DOI registry and a suffix, a local identifier. In this way DOIs are easily resolvable by the DOI Proxy Server System. Based on DOIs, the CrossRef association (<http://www.crossref.org/>) enables the effective tracking and linking of citation, for enhanced access and transparency in scientific publications.

Recently DOIs have also been used to identify scientific data, and the DataCite (<https://www.datacite.org/>) organisation has been established to enable citation and link tracking of data in a similar way as the CrossRef association, with potentially similar benefits. DataCite DOIs use specific metadata elements tailored towards data (e.g. *location*), using these DOIs will allow not only to create an inventory of data, but enabling possible links with publications, which is a significant step towards transparency and reproducibility of scientific evidence.

The idea of creating an inventory of models stems from the wish to share knowledge and foster re-use of models within their domains. As a result, many existing model inventories are domain-specific, listing models within one or more closely-related domains. Examples are the model repository hosted by the *Community Surface Dynamics Modeling System* CSDMS for models relating to the Earth's surface (CSDMS Facility), the *Model Documentation System* MDS for the air quality domain (ETC/ACM), the *BioModels Database* for biological processes (Juty et al 2015) and Nexus Tools Platform (alpha) (United Nations University 2015) for water, soil and waste, to name but a few.

An example of a *domain-independent* model inventory on the other hand is the LIAISE^{vii} Toolbox (Rennings, 2013). Like the application described in this paper, the LIAISE toolbox has the aim of supporting the policy process, in particular in

^{vii} The LIAISE *Community of Practice on Impact Assessment Research for Sustainable Development* is an active multidisciplinary network of research organisations and researchers contributing to policy impact assessment and evidence based policy-making. More information on the LIAISE toolbox and its scope can be found in (Jacob et al, 2013).

the field of ex-ante and ex-post Impact Assessments which are also of major interest for the MIDAS community.

The above-named inventories provide the user with a classical view on descriptive elements about the models, with some options to create linkages between models, but usually only through common keywords. All the listed inventories cover descriptions of inputs, in the sense of the parameters required, but only the Nexus Tools platform uses code lists for parameters, effectively allowing to compare or potentially link models based on their input and output parameters. Apart from these implicit links, explicit relationships between models, or between models and data, have not been identified in the investigated inventories.

By contrast, a prominent and successful example of representation, visualisation and *reproduction* of scientific workflows is the *myExperiment* (De Roure et al., 2009) platform, jointly developed by the universities of Southampton, Manchester and Oxford in the UK. *myExperiment* allows scientists to describe, execute and share scientific workflows. The focus of *myExperiment* lies in creating and sharing reproducible scientific workflows. The majority of the published workflows are written using Apache Taverna, which allows design and execution of workflows using web services, though *myExperiment* is not limited to Taverna (De Roure et al., 2009).

A drawback of *myExperiment* in comparison to the MIDAS platform is that the knowledge contained is accessible only to experts, not necessarily in a particular domain, but at least in the representation of scientific workflows. For a non-technical or non-scientific audience the knowledge remains unaccessible. Addressing this issue, in 2010 TED fellow Eric Berlow gave an inspiring talk on TEDGlobal on “Simplifying complexity”. He suggested embracing complexity by relying on “the simple power of good visualization tools to help untangle complexity and just encourage you to ask questions you didn't think of before” (Berlow, 2010). Using Visual Analytic (VA) techniques, in his talk he applied network representations of entities to a food network and to the U.S. counterinsurgency strategy in Afghanistan.

Kehrer and Hauser (2013) made a survey of more than 200 scientific papers dealing with VA and give a detailed overview of the field and related techniques and areas of application. Putting a particular emphasis on the visual representation of relationships and dependencies, as an important aspect of what we try to achieve, we would like to highlight the work of the visualisation tool *Circos* (Krzywinski et al, 2009) and its use of circular diagrams to represent knowledge and facilitate the identification and analysis of similarities and differences arising from comparisons of genomes. The use of circular diagrams to visualise knowledge and help understanding complex relationships in other

domains is illustrated by the example of the application of a circular chord diagrams to *Violence and guns in best-selling video games* in The Guardian (Guardian, 2013).

This builds our bridge to data narratives, since, besides being a chord graph, the example of *Violence and guns in best-selling video games* is also a data narrative in form of an “infographic”, a class of visualisation where narratives are combined with interactive graphics, allowing the data to tell a story. Segel and Heer, (2010) provide a good overview of the area of data narratives, including many examples and a categorisation of narrative components. Data narratives and infographics are widely adopted in journalism and are closely related to of Open Data because this initiative has increased the availability of accessible source material considerably.

3. BACKGROUND AND SCOPE OF MIDAS - THE MODELLING INVENTORY DATABASE AND ACCESS SERVICES

Documenting the models and model combinations in use by the JRC, in an understandable manner, is a major challenge: Models in JRC are applied within various institutes, organised by discipline, with hundreds of scientific and administrative users across Commission sites in various countries. The list of domains ranges from greenhouse gas emissions, land use change and ecosystem services, energy consumption and economy to structural integrity assessment, to name but a few. In addition, the types of models vary widely, including, for example, stochastic, deterministic, general equilibrium, partial equilibrium, and recursive models. The majority of these models can be run in combination with other models. For this purpose some of the models are already integrated in modelling platforms. Thus, they form networks of interaction, further complicated by the related input datasets, scenarios, methods etc.

Documenting these model combinations and networks of interaction and making them more transparent becomes particularly important if they contribute to an ex-ante Impact Assessment (ex-ante IA). Ex-ante Impact Assessments have been established as an EC instrument in 2002 in (COM/2002/276). They have to be carried out for all EC policies and initiatives. Ex-ante IAs contain comprehensive assessments of the potential economic, social and environmental impacts of a new policy, in comparison with the so-called *baseline scenario*, which describes the current situation without the policy in place. Their role has been re-inforced in the EC Better regulation package (COM/2015/215) and related guidelines (SWD/2015/111).

In 2012 the JRC therefore initiated the development of *MIDAS - the Modelling Inventory Database and Access Services* with the following scope:

- i. To describe all models that are in use within JRC to directly or indirectly support the policy cycle of EU policies. This covers models either developed or co-developed by the European Commission, as well as third party models used in the JRC for policy support activities.
- ii. To facilitate sharing and preservation of knowledge and to facilitate understanding of “what’s going on in a particular modelling field in JRC” for all its users, independent of their domain and expertise.
- iii. To enhance understanding of models by describing their context: how they are made, what they support, how they can be run and who has the expertise.
- iv. To enhance transparency of models; partly in order to ensure consistency in the use of model combinations, input data, parameterisation and underlying assumptions, but also to enable understanding and usage across domains and disciplines.
- v. To maintain the history of models no longer in use. The mandate of MIDAS is an up-to-date inventory. Since a first model inventory done in 2010, there was a change rate of up to 30% within the various JRC institutes. Developing an inventory with a *memory* reflecting the period in which a model had been in use was therefore an important aspect.

To our knowledge, MIDAS goes beyond the scope of existing inventories by covering many different domains, unlocking the represented knowledge for scientists from various domains and policy makers, and thus enabling inter-domain re-use. By linking the data sources, it greatly enhances transparency of models, model networks and model results.

4. PROCESS & APPROACH

When formulating the MIDAS vision and scope in 2012, we were aware that this was not just about designing a system. Instead, great institutional, cultural and conceptual challenges laid ahead of us: we had to convince scientists to invest time and resources to *share their knowledge*, in an *understandable manner* and *across different domains*, with us, with their peers, and with their policy making colleagues, and to maintain the information as long as the models are in use.

The institutional challenges were tackled right from the start of the process: A solid governance structure was put in place that involves a board for taking strategic decisions, a coordination and federated communication structure for implementing these decisions in each of the involved institutes, and an entity for quality control and user testing, to facilitate the sharing of information.

To work on the cultural challenge of sharing, the top-down governance approach was complemented by a bottom-up approach with a focus on continuous involvement of the modelling community to: (i) understand barriers to and incentives for creating and sharing knowledge in the given setting, and (ii) collect and refine the requirements towards such an inventory and corresponding information system to maximise the relevance for the user.

A specific Working Group was initiated at the end of 2012 with aim of “touching base” with both modelers and policy makers. The group started with interviews focusing on models of corporate importance to get a better understanding of the needs of modelers and model users within the organisation.

During this exercise the team identified the first set of core entity types and corresponding relationship types for describing the models in context. In particular this covered other related models, related datasets, peer-reviewed publications describing models and their quality, reports on model application, people within the organisation who have the expertise to run them and policy documents which the models supported. The ability to identify that two models related to the same entity, uniquely identified through a PI independent of the local storage or access mechanism, was seen as a major benefit. Especially problems and opportunities concerning the consistency of underlying datasets were of interest for model users. Whenever possible, direct access to the entities, in particular to data in question was requested following wherever possible Open Data principles, to provide a step towards the reproducibility of results.

These findings were translated into a preliminary version of MIDAS, hosting only a few model descriptions, to explain the concept. Using this first version as a showcase, the MIDAS development team initiated the first data collection campaign contacting more than 100 modeling experts in JRC to obtain descriptions of the models they are working with. The team dedicated considerable time to face-to-face meetings and single or group training activities. Though these activities were time-intensive, they proved to have the highest acceptance and best user feedback especially in the early phases of the campaign.

During this exercise the team focussed on identifying the following challenges to *knowledge sharing*: 1) barriers and incentives for creating and sharing knowledge, 2) issues and solutions concerning implicit and tacit knowledge and 3) ways to effectively communicate the resulting complex network of information across different domains and to a non-technical audience.

Concerning sharing, one of the main barriers identified was the absence of feedback for the given input and recognition for work undertaken. To overcome this and stimulate the sharing of knowledge, the development team implemented

game design elements in a non-game context (gamification, as defined by (Deterding et al, 2011)) in the MIDAS system. Gamification is an increasingly popular topic in system design, and (Hamari et al 2014) carried out a literature analysis of empirical studies in the field of gamification, concluding that “according to a majority of the reviewed studies, gamification does produce positive effects and benefits”. MIDAS implements scoreboard-like features to i) provides users with immediate feedback on what they entered, ii) provide them with good visibility of their own research^{viii} and iii) allow them to see their performance in relation to others.

Another identified barrier for sharing was when it was perceived as a burden due to duplication, i.e. an environment where users have to fill in their information several times into several different applications. MIDAS was designed to be placed in Service Oriented Architecture (SOA), to make the most of existing infrastructure and to retrieve information that is already stored and maintained elsewhere. More information on the architecture is given in section 5.1.

Concerning implicit and tacit knowledge the use of domain-specific terminology and abbreviations used by experts when talking about processes and involved entities were hindering effective communication of knowledge across domains. To meet this challenge we decided to visualise the descriptions and connections by drawing on experience acquired in the EuroGeoss project, where visual representation was used to document models and scientific workflows (Vaccari et al, 2012): Processes, focussing on inputs and outputs, were formalised in easy-to-read flow charts. Domain-specific terminology hindering the identification of two models relating to the same entity was already addressed through the use of PIs.

As so often, the whole turned out to be greater than the sum of its parts, and the knowledge of all experts combined revealed both issues and synergies. This information, however, was partly hidden in second or third order relationships, e.g. the links of connected models to underlying datasets. As a response, VA visualisation techniques were implemented, showing all entities in context and allowing users to capture “the big picture”.

^{viii}A similar challenge has been described by the creators of ^{my}Experiment (de Roure et al 2009). As a response their strategy is focusing on credit, attribution and licensing.

Further functionality to allow users to browse the resulting information was inspired by data narratives (Segel and Heer, 2010) (e.g. details-on-demand, drill-down-story and visual highlighting and multi-messaging). Narratives and storytelling are in fact one of the inbuilt benefits underlying our main concept of “models in context”, as it is the context that creates the story. We further detail the visualisation principles in Section 5.

After the data collection campaign was complete, the activity and tool were first officially presented to its user community in October 2013. Since then regular data collection campaigns have been carried out to further improve and extend the content. Today MIDAS allows modelers and policy makers in JRC and other Commission Services to find more than 160 models - and 1000+ entities linked to these models through more than 1600 connections - and to assess their use for various aspects of the policy making cycle.

5. THE MIDAS SYSTEM

In this section we provide an overview of the MIDAS system starting with the MIDAS architecture, focussing on the interaction with other services. This is followed by a brief overview of the MIDAS front end, and the representation of the model descriptions in so-called fact sheets. The last section describes the different types of visualisation available within MIDAS.

5.1 The Architecture

MIDAS is situated within the Commission Intranet. It is accessible for all staff within the Commission Services. All tools provided in MIDAS are arranged around the MIDAS database, which forms the core of the application.

MIDAS hosts the descriptions of the models and related entities. For models the DB hosts a complete metadata profile, based on previous work carried out by the JRC Institute for Prospective Technological Studies (IPTS). The profile has been enhanced with elements coming from the existing inventory CSDMS, and modified and extended during the data collection campaigns described in section 4 to reflect user needs. The profile itself still undergoes extensions, as new user needs are identified, however ensuring backwards compatibility with previous versions.

For related entities the DB is situated within a SOA and only hosts minimal information: Title, if applicable acronym, short description, and, wherever possible, persistent (or at least unique) identifier PI. In case of a PI the resolving engine is related to the type of PI and the type of entity and provider. If a PI and resolving engine are not available, we use URLs to resolve into the location of the

entity itself or a metadata entry of the entity. Figure 1 summarises the MIDAS main entities and main PI and resource providers.

Figure 1: MIDAS Entities and Related PI and Resource Providers



In case of EU policies we connect to the EUR-Lex service of the EU Publications Office (<http://eur-lex.europa.eu/>). EUR-Lex provides free access, in the 24 official EU languages, to EU laws (EU treaties, directives, regulations, decisions, consolidated legislation, etc.), preparatory acts (legislative proposals, reports, green and white papers, etc.), EU case-laws (judgements, orders, etc.), and many more. EUR-Lex provides the *CELEX number*, a unique identifier for each stored document, regardless of language. To resolve the numbers they can be concatenated with the URL pattern EUR-Lex provides. The additional information about EU policies that MIDAS stores, such as the full title but also related legal documents (e.g. the ex-ante impact assessment that accompanies each policy proposal) is retrieved through the dedicated web services offered by EUR-Lex.

For documents we use a mixed approach: in the case of publications with JRC participation we connect to the in-house repository service, the JRC Scientific Knowledge Portal (SKP) that provides access to all documents written by or with participation of Commission staff, using the identifier assigned by the JRC to all its' publications. MIDAS stores the identifier which is then resolved providing access to the location of the corresponding metadata record, and to the document itself. In case of non-JRC publications we make use of DOIs and the corresponding resolver (see section 2).

For experts within the Commission we connect to the European Commission Authentication Service (ECAS). ECAS is not strictly persistent, as the account is deactivated when people leave the service, and the account id can no longer be resolved. We are currently investigating the complementary use of ORCID

(<http://orcid.org/>), though we recognise that ORCID's are specific to the research domain.

For data in particular, a common sustainable approach for the unambiguous identification of entities is still missing. MIDAS therefore links to various repositories and services providing direct access to data sources, such as the European Environment Agency (EEA), the Statistical Office of the European Union (Eurostat), the European Central Bank (ECB), the Organisation for Economic Co-operation and Development (OECD), and the Food and Agriculture Organisation of the United Nations (FAO). Information on identifiers, and related information like titles and abstracts have been harvested from services made available by the providers. We are in the process of linking to the EU Open Data Portal (<https://open-data.europa.eu/>), from which we will obtain, in the future, references to data published by EU and other institutions, rather than linking to the institutions individually.

Models also suffer from the absence of a common entity identification approach. MIDAS represents the connections with other models in use by the JRC, and therefore uses identifiers provided by the MIDAS system itself. External models can be registered using a URL which resolves to the location of a metadata entry for the model, however, at the moment no additional metadata for these external models is retrieved or stored.

Besides these main entities, MIDAS also makes use of authority code lists, in particular the Named Authority Lists (NAL's) codes and the associated labels used by the Publications Office of the European Commission in order to identify, for example, the various Directorates-General and other Commission Services. These NAL's are also known as (Common) Authority Tables (CAT's), controlled vocabularies or value lists (EU Publications Office, 2015).

The relationships between entities are organised in a structure inspired by a triplestore, listing all relationships into subject (identifier) predicate (relationship) and object (identifier). Inspired by Common European Research Information Format -CERIF (<http://www.eurocris.org/Index.php?page=CERIFreleases&t=1>) each relationship is equipped with a start and an end date, to capture their temporal extent.

5.2 The MIDAS Front End

The MIDAS front end (Figure 2) is a web application, from where all the main tools of MIDAS can be accessed.

Figure 2: MIDAS 4.2



So far the main tools are:

- **Inventory:** The MIDAS Inventory lists all the models and the specified related entities in the MIDAS database. Here users can get a quick and structured overview of the MIDAS DB.
- **Editor:** The MIDAS Editor allows users to create / modify model descriptions. It gives access to all code lists and the related inventories MIDAS makes use of.
- **Search:** A classical “google like” search engine where users can search for specific model descriptions based on free text search or predefined filters. It gives access to the metadata per model, the so-called **model fact sheet**.
- **Browse:** MIDAS browse tools are meant to visualise the big picture: allowing users to view all models in their context, to browse links between entities, to

explore them, and to mine information which is otherwise hidden in flat descriptions.

- **Statistics:** An alternative way to view the MIDAS database. Based on the relationships between entities users can find statistics on most popular data sets and data providers, total number of model-to-model relationships etc.

The most relevant aspects of MIDAS for this paper are the model fact sheets and the MIDAS Browse tools, which will be described in more detail in the following sections.

5.3 The Model Fact Sheet

The model fact sheet is the main representation of each model described in MIDAS. It summarises the model metadata and each model's related entities. Figure 3 shows a sample fact sheet of one of the entries in MIDAS: *LUISA - Land-use-based Integrated Sustainability Assessment Modelling* (Baranzelli et al 2014).

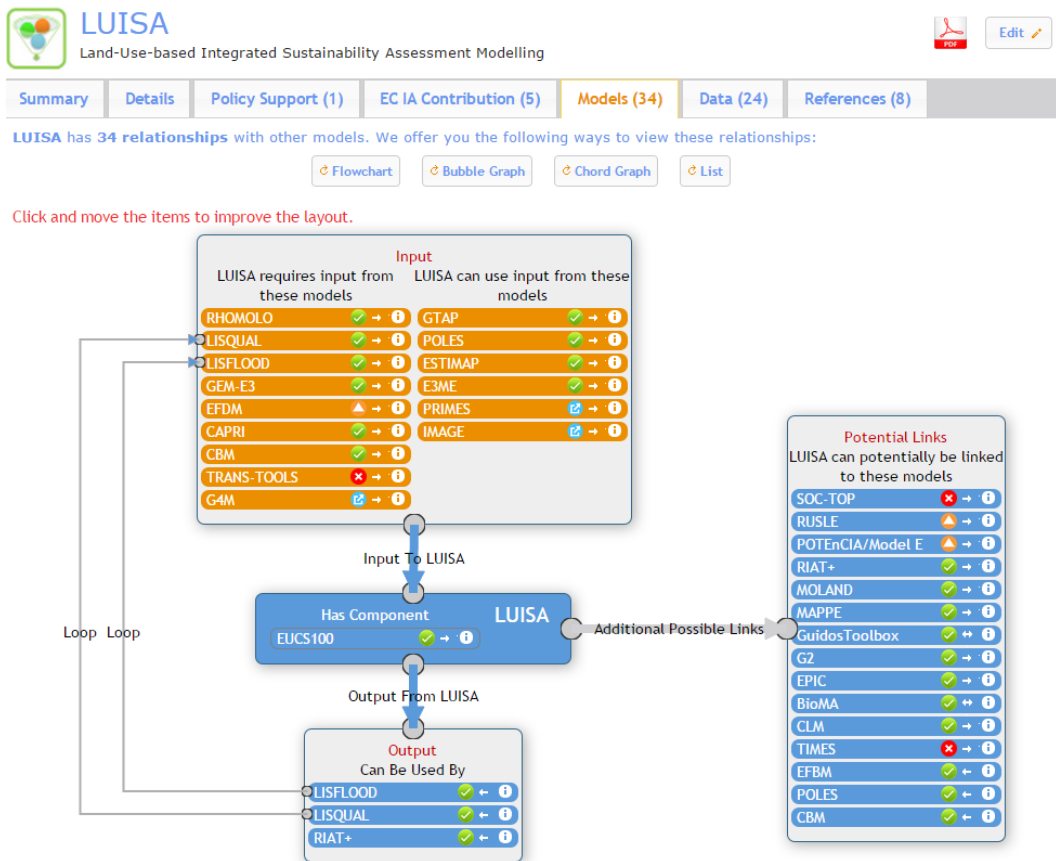
Figure 3: MIDAS Models Fact Sheet, Summary Tab – LUISA (Baranzelli et al. 2014)

The screenshot displays the MIDAS Portal interface for the LUISA model fact sheet. At the top, the European Commission logo and 'MIDAS PORTAL Modelling Inventory Database & Access Services' are visible. A navigation bar below the header shows various icons and a search bar. The main content area is titled 'LUISA Land-Use-based Integrated Sustainability Assessment Modelling'. A green checkmark icon indicates that the model is operational. The text describes the model's purpose: to assess environmental consequences of policies by modeling land use changes. It mentions that LUISA is based on 'Dynamic Land Functions' and is used for policy design. The right sidebar contains sections for 'Model Links', 'Point Of Contact' (Carlo Lavalle), 'Model Status' (Operational / in use in JRC), 'Policy Areas' (Agriculture, Fisheries and Food, etc.), and 'Client Institutions' (Directorate-General for Agriculture and Rural Development, etc.).

The fact sheet is organised into several tabs. The metadata elements represented in the *Summary* tab (Figure 3) and the *Details* tab are classical metadata elements, including aspects like a contact point, short description, information on model parameterisation and output types, supported policy areas, client institutions within the EC, spatial and temporal extent and resolution (if applicable), property rights and the conditions for access and use of the model. The remaining tabs *Policy Support*, *EC IA Contribution* (as a specific category of policy support of particular interest for the users), *Models*, *Data* and *References* describe specific related resources and the numbers in brackets show how many relationships have been established in each category.

To provide an example Figure 4 shows the models tab of LUISA: a flow chart representation formalises the links that have been established between LUISA and other models.

Figure 4: MIDAS Model Fact Sheet, Models Tab – LUISA



The flow chart aims to show the particular flows of inputs and outputs, and the potential links which describe a potential collaboration or link between two models. The information represented is knowledge collected from various modelers. Once a relationship is described that affects two models (e.g. one model using the output from another model) it is shown in both model descriptions, informing the user which of the involved parties established the link. Possible feedback loops between two models are automatically identified and marked in the flowchart. Users can attach additional notes to each relationship to describe for example which output parameter of a specific model they use as input to their own model.

5.4 The Browse Tools










While the fact sheet offers the user a view on one particular model, we have developed a range of tools that embrace the complexity of the bigger picture. These tools allow users to visualise and browse all relationships, getting a global overview of all entities as well as more detailed information through drill-down views if required.

So far we provide three different visual representation techniques: Bubble Graphs, Word Clouds and Chord Graphs, all described in more detail in the following sections. All tools are built using the javascript Library D3 Data Driven Documents (<http://d3js.org>), an extensive library which produces SVG output graphs, allowing for a full web integration with customisable mouse events, css styling, and javascript scripting.

The choice of tool was made based on the initial scoping of MIDAS and the implemented gamification strategy, to allow users to see which models are connected and whether they use the same data, but also how models perform in terms of number of supporting publications and number of supported policies (score-board). While D3 is widely adopted for browser-based visualisations, we believe that its use in the context of a model inventory and for the specific purpose of model transparency is novel.

The different tools can be parameterized to provide specific information. The MIDAS Browse tool offers predefined entry points to the tools, all summarized on the MIDAS Browse page in Figure 5.

Figure 5: Summary of All Browse Tools Currently Existing in MIDAS 4.1

BROWSE




Explore **relationships** between models, data, policies, people

NOTE: This is an expert tool. If you are using MIDAS for the first time, we suggest the [Search](#)

'A picture is worth a thousand words'

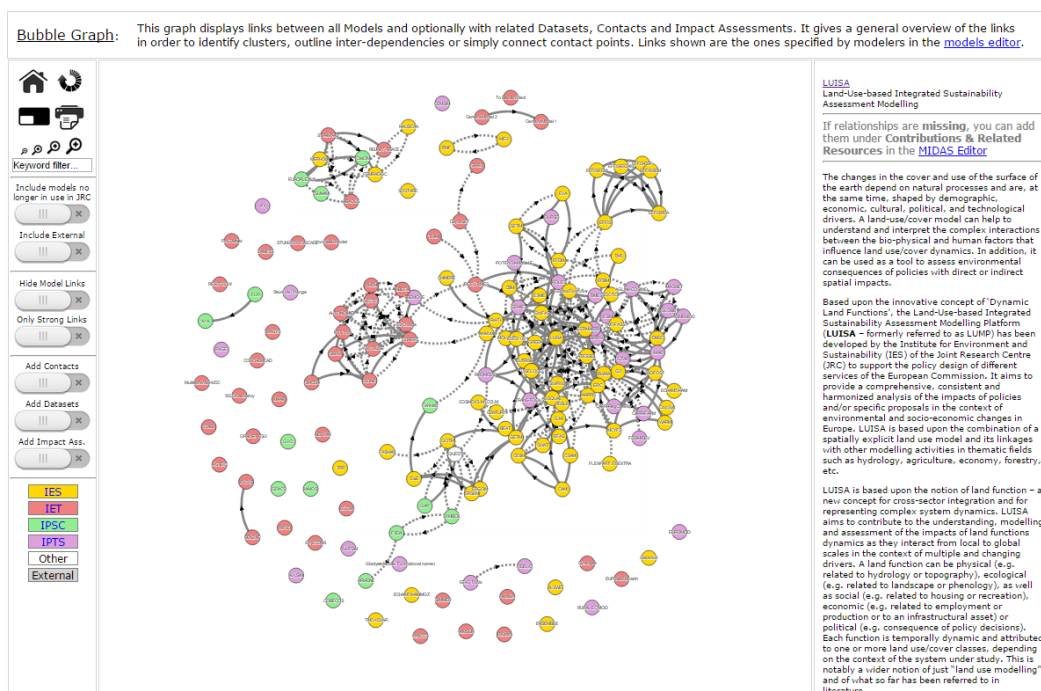
One of the main goals of **visualization** is to allow users to absorb large amounts of data quickly.

MIDAS browse tools allow users to visualize the context of models: You can see **which models are connected**, whether they **use the same data**, and how they support policies. You can explore all these links and mine information which are otherwise hidden in flat descriptions.

<p>Bubble Graphs </p> <p>Bubble Graphs allow users to browse all relationships between entities</p>	<p>Models</p> <p>How models inter-relate. Visualise 'centres of gravity'</p>	<p>Models & Data</p> <p>Which models use the same input data?</p>	<p>Models & Policy</p> <p>Which models support the same impact assessment?</p>	<p>Models & People</p> <p>Visualise model contact points</p>
<p>Word Clouds </p> <p>Look at models based on their characteristics i.e. support of impact assessment</p>	<p>Impact Assessment</p> <p>Models most used for impact assessment</p>	<p>Themes</p> <p>Thematic areas that models support</p>	<p>JRC Publications</p> <p>Models best documented through publications with JRC</p>	<p>Connectivity</p> <p>Models most linked to other models</p>
<p>Chord Graphs </p> <p>Chord Graphs focus on inter-model relationships</p>	<p>Strong Links</p> <p>Models that depend on each other for input or output</p>	<p>All Links</p> <p>Includes also potential links or collaboration</p>	<p>Policies</p> <p>Models supporting policies</p>	

Bubble Graphs are graph representations showing entities as circles (“bubbles”) and connections between these entities as lines. The lines show the direction of the connection, and the line representation gives information on the type of relationship. A global overview of all the models is initially clustered on the basis of the number of inter-connections (gravity type of layout), and colored along the lines of JRC Institutes. This allows a quick identification of highly interconnected models, shows the intensity of the cooperation between Institutes or can be used as a simple measure of the total number of JRC models.

Figure 6: Bubble Graphs Representing Model-to-Model Relationships in Midas



Bubble graphs also allow user interaction: Filters and other customizations allow the user to build customised graphs, e.g. showing only strong model links, or including datasets/people/policies in the global picture. This allows for different perspectives, like a “model perspective”, a “data perspective”, a “policy perspective” or a “people perspective”.

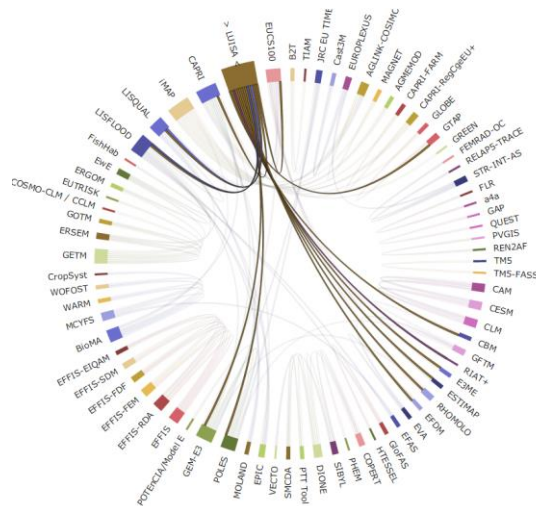
These views are particularly interesting for information mining and for understanding dependencies hidden in second order relationships, such as whether two models that directly or indirectly depend on each other are actually also operating on the same data.

have repeatedly served the purpose in the past. It also acts as a score board for models that have a particularly high score in terms of supported initiatives.

Chord Graphs are based on graph theory, showing nodes and edges. They represent entities as an outer circle of nodes, and relationship between these entities (edges) are displayed as chords linking the nodes.

Chord graphs provide a quick global overview of all models and how they are interlinked. Users can easily identify models with high connectivity: i.e., those models that either provide input to (or use input from) many other models. Color coding is used to represent the direction of the relationship. The graph is interactive and can focus on a specific model at a time.

Figure 9: Chord Graph Sample: Connections of LUISA



This representation gives visibility to highly interconnected models. Thus, models that might have a lower visibility within the institution due to few direct policy support activities, but that provide important input to many other models, receive through the chord graph representation the visibility they deserve.

However, the more entities the graph contains, the smaller each representation becomes, and the harder it is to identify a single model and its relationships. The MIDAS database is growing, and so is the chord graph showing both active and retired models. Therefore we implemented various settings that allow inclusion or exclusion of certain categories of models (e.g. those no longer in use by the JRC). Furthermore in the fact sheet of each model it is possible to view only those models that are directly or indirectly related to the model of interest, meaning it shows all models the model of interest itself relates to (1st order

relations) as well as *their* 1st order relations (i.e. 2nd order to the model of interest). This enhances the readability of the graph and allows the user to focus on a single model.

6. DISCUSSION

Looking back at the process of implementing MIDAS, we can say that our approach and the majority of strategies to overcome the listed challenges have been successful. Below, we reflect on some of the strategies adopted and on lessons learned.

The governance structure and regular updating campaigns proved an important piece in the puzzle, but to address the issue of sharing, visual tools appear to be a more durable strategy, as the visual appeal of the tool greatly encouraged modellers to check, extend and improve the detailed information about their models: The tool gives them visibility and responsibility over the published information, while the score-board like visualisation motivates them to extend their model descriptions. The feedback we received from users on this type of visualisation was 100% positive, and in 5 out of 8 cases users considerably improved their model description after being confronted with the resulting visualisation. However, users also noted that the effect would be stronger and the tool more useful if access to MIDAS was extended to the scientific community outside the Commission Network.

The visual tool also proved very helpful to identify errors or inconsistencies which would otherwise have gone unnoticed in a flat list-based textual format: Once visualised the errors cause a change of a 'primitive features' like shape. For example: a term that is much shorter (or longer) than the norm, for example due to an error caused by the user when filling in the information, stands out in both the Bubble graphs and Chord graphs, even if the amount of represented terms in the graphs is large. This is caused by the so-called "pop-out effect", which was first described in (Treisman & Gelade (1980)), and discussed in the context of evaluation of information visualisation in (Shovman et al 2009).

However, as highlighted in the Chord graphs example given in section 5.4, the success of a specific type of visualisation very much depends on the content that is visualised. It is therefore necessary to adapt the visualisation alongside the growing MIDAS inventory. The MIDAS team spends part of its time on investigating new libraries and graph types that can be used. In particular the visualisation of the temporal aspect is an area of investigation: the use of data and models evolves in time, and the entities themselves evolve through versions, related institutions are split up, fused or change name over time, and even relationships and policy contributions evolve.

Since the first version of MIDAS was made available, both the modellers and the administration/management were asked to provide feedback on what they would be interested to see to further enhance transparency of both models and model results. A lot of new ideas spawned from these consultations. Of particular interest are scenarios (e.g. baseline scenarios used in ex-ante impact assessments to compare policy options to a baseline situation) together with underlying assumptions, model configurations for a particular run, and model projections. Also the provision of a snapshot archive for data as it was used for a specific policy support activity is of interest for the community, to enhance transparency and reproducibility of the results. We will look into these aspects in the future version of MIDAS.

For data resources, the use of PIs has just been initiated, and unambiguous identification of entities including versioning, persistence of the given identifiers, long term preservation of the related data, clear license conditions and access conditions, are still a long way off. For the moment the maintenance carried out by the MIDAS team to ensure that entities are still available at the referenced location is quite high. However, we believe that with OD and ORD, time is working in our favour, and we monitor the sources we use. Once a provider starts making data available as Open Data (e.g. through the EU Open Data Portal) and/or starts making use of PIs (e.g. the OECD) we will adjust MIDAS to make use these identifiers instead.

By monitoring the sources we aim to slowly but surely reduce the maintenance work and provide a better service for our users. We also monitor whether data available as Open Data remains accessible, as even here the long-term-availability of data is not necessarily given. (Tonkin 2008) even states that "Technology cannot create a persistent identifier, in the digital library community's sense of the term". However, we are not deterred by this as we think that the benefits of the presented results largely outweigh the efforts.

7. CONCLUSION

MIDAS describes models in use by the JRC by documenting their context: their relationships with other models, with data, with documents and supported policies. By doing so, MIDAS reveals and documents implicit and tacit knowledge, fosters use and re-use of existing models and model results, and provides an important first step towards enhancing transparency and reproducibility of scientific evidence underpinning policy.

In the process of our work we faced both technical and organisational/ cultural challenges, with the latter often being the more prominent. In particular, a major issue is avoiding duplication of work and providing incentives for scientists to open up and share their knowledge through visibility for their work. MIDAS

currently makes the resulting knowledge accessible for both experts and policy makers within the Commission. Extending the audience to the wider scientific community, and creating links to other existing initiatives and inventories has to be explored to maximise incentives and usefulness of the tool and the information it contains.

In the future development of MIDAS we will take further steps towards transparency and reproducibility of models and model results underpinning policy making. In particular we will look into the specific support given by models to ex-ante Impact Assessments, where, in preparation for new EU policies, different policy options are compared concerning their potential economic, social and environmental impacts. The MIDAS developments may include documenting the assumptions underpinning the used model runs, to represent the interactions between the different models involved, and to provide access to the data produced to support the ex-ante IA. Also here extending the audience to the wider scientific community and the general public should be explored to reach the intended goals.

Considerable effort will be required to maintain the content of MIDAS also in the future. A significant step to facilitate this maintenance will be the development of a broader knowledge management strategy that promotes professional recognition of scientists for publishing not just scientific papers but also data and models.

ACKNOWLEDGEMENT

We would like to thank Massimo Craglia (DG JRC) for his input on reproducibility, Open Data and Open Research Data. We would like to thank Carlo Lavallo (DG JRC) who provided the example of LUISA (Land-Use-based Integrated Sustainability Assessment Modelling) (Baranzelli et al 2014). We would also like to thank our three anonymous reviewers for providing valuable comments, helping to further improve the paper.

REFERENCES

2012/417/EU: Commission Recommendation of 17 July 2012 on access to and preservation of scientific information; CELEX: 32012H0417

Baranzelli C, Jacobs C, Batista E Silva F, Perpiña Castillo C, Lopes Barbosa A, Arevalo Torres J, Lavallo C. (2014). The Reference Scenario in the LUISA platform – Updated configuration 2014 Towards a Common Baseline Scenario for EC Impact Assessment procedures. EUR 27019. Luxembourg (Luxembourg): Publications Office of the European Union; JRC94069; doi:10.2788/85104

- Berlow, E. (2010). Simplifying Complexity, TEDGlobal 2010, at http://www.ted.com/talks/eric_berlow_how_complexity_leads_to_simplicity [accessed 28 September 2015]
- CSDMS Facility (2015), University of Colorado, Boulder: Community Surface DynamicsModelling System (CSDMS), at <http://csdms.colorado.edu/> [accessed 28 September 2015]
- COM/2002/0276 Communication from the Commission on impact assessment; CELEX: 52002DC0276
- COM/2010/2020 EUROPE 2020 A strategy for smart, sustainable and inclusive growth; CELEX: 52010DC2020
- COM/2010/245 COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A Digital Agenda for Europe; CELEX: 52010DC0245
- COM/2012/0401 COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Towards better access to scientific information: Boosting the benefits of public investments in research; CELEX: 52012DC0401
- COM/2012/392 COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A Reinforced European Research Area Partnership for Excellence and Growth; CELEX: 52012DC0392
- COM/2015/215 COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Better regulation for better results - An EU agenda; CELEX: 52015DC0215
- De Roure, D., Goble, C. and Stevens, R. (2009). The Design and Realisation of the ^{my}Experiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems* 25, pp. 561-567; doi:10.1016/j.future.2008.06.010
- Deterding, S., Dixon, D., Khaled, R., Nacke, L. (2011). From game design elements to gamefulness: defining "gamification", *Proceedings of the 15th*

International Academic MindTrek Conference: Envisioning Future Media Environments; 01/2011

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE); CELEX: 32007L0002

Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information; CELEX: 32013L0037

ETC/ACM (2015), MDS - Model Documentation System, at <http://acm.eionet.europa.eu/databases/MDS> [accessed 28 September 2015]

EU Publications Office (2015). Named Authority Lists, Metadata Registry (MDR), at <http://publications.europa.eu/mdr/authority/index.html> [accessed 28 September 2015]

GBIF (2011). A Beginner's Guide to Persistent Identifiers, version 1.0. Released on 9 February 2011. Authors Kevin Richards, Richard White, Nicola Nicolson, Richard Pyle, Copenhagen: Global Biodiversity Information Facility, 33 pp, at http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf [accessed 28 September 2015]

Guardian (2013). Violence and guns in best-selling video games. The Guardian, at <http://www.theguardian.com/world/interactive/2013/apr/30/violence-guns-best-selling-video-games> [accessed 28 September 2015]

Grazzini, J., Pantisano, F. (2015). Collaborative research-grade software for crowd-sourced data exploration: from context to practice - Part I: Guidelines for scientific evidence provision for policy support based on Big Data and open technologies, JRC Technical Report EUR 27094; doi:10.2788/329540

Hamari, J., Koivisto, J., Sarsa, H. (2014). Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification, 47th Hawaii International Conference on System Sciences (HICSS) 6-9 Jan. 2014, pp. 3025-3034; doi:10.1109/HICSS.2014.377

Jacob, K., Arampatzis, S., Manos, B., Bournaris, T. (2013). A Toolbox for Impact Assessment and Sustainability, Procedia Technology, Volume 8, 2013, Pages 355-359, ISSN 2212-0173; doi:10.1016/j.protcy.2013.11.047

- Juty, N., Ali, R., Glont, M., Keating, S., Rodriguez, N., Swat, M.J., Wimalaratne, S.M., Hermjakob, H., Le Novère, N., Laibe, C. and Chelliah, V. (2015). BioModels: Content, Features, Functionality, and Use. *CPT Pharmacometrics Syst. Pharmacol.* 4; doi:10.1002/psp4.3
- Kehrer, J. and Hauser, H. (2013). Visualisation and Visual Analysis of Multifaceted Scientific Data: A Survey. *Visualisation and Computer Graphics. IEEE Transactions on*, vol.19, no.3, pp.495-513, March 2013; doi:10.1109/TVCG.2012.110
- Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M.A. (2009). Circos: an Information Aesthetic for Comparative Genomics. *Genome Res* (2009) 19:1639-1645; doi:10.1101/gr.092759.109
- Nature Editorial (2014). Journals unite for reproducibility. *Nature News*, Publisher: Nature Publishing Group, Date: Nov 5, 2014; doi:10.1038/515007a
- Rennings K. (2013). Modelling. *Economic Modelling. LIAISE Toolbox*, at <http://www.liaise-kit.eu/ia-method/modelling> [accessed 28 September 2015]
- Science Editorial (2014). Journals unite for reproducibility. Published Online November 5 2014; *Science* 7 November 2014: Vol. 346 no. 6210 p. 679; doi:10.1126/science.aaa1724
- Segel, E., Heer, J. (2010). Narrative Visualization: Telling stories with Data, in: *IEEE Transactions on Visualization and Computer Graphics*, Volume 16 Issue 6, November 2010, Pages 1139-1148; doi:10.1109/TVCG.2010.179
- Shovman, M. M., Szymkowiak, A., Bown, J. L., Scott-Brown, K. C. (2009). Changing the View: Towards the Theory of Visualisation Comprehension *Information Visualisation*, 13th International Conference In Information Visualisation (2009), pp. 135-138; doi:10.1109/iv.2009.78
- SWD/2015/111 COMMISSION STAFF WORKING DOCUMENT Better Regulation Guidelines; CELEX: 52015SC0111
- Tonkin, E. (2008). Persistent Identifiers: Considering the Options July 2008, *Ariadne Issue 56* <http://www.ariadne.ac.uk/issue56/tonkin/> [accessed 28 September 2015]
- Treisman, A.M., Gelade, G. (1980). A feature-integration theory of attention *Cognitive Psychology*, Vol. 12, No. 1. (January 1980), pp. 97-136; doi:10.1016/0010-0285(80)90005-5

- UK Cabinet Office (2013). Open Data Charter. Charter on open data signed by G8 leaders to promote transparency, innovation and accountability, from 18 June 2015, at: <https://www.gov.uk/government/publications/open-data-charter> [accessed 28 September 2015]
- United Nations University (2015). UNU FLORES Nexus YTools Platform (alpha). Interactive comparison of Nexus related modelling tools, at [https://data.flores.unu.edu/projects/ntp/#/dashboard/elasticsearch/Nexus%2520Tools%2520Platform%2520\(alpha\)](https://data.flores.unu.edu/projects/ntp/#/dashboard/elasticsearch/Nexus%2520Tools%2520Platform%2520(alpha)) [accessed 28 September 2015]
- Vaccari, L., Craglia, M., Fugazza, C., Nativi, S., and Santoro, M. (2012). Integrative Research: The EuroGEOSSE Experience, in: IEEE Journal of selected topics in applied earth observations and remote sensing, Vol. 5, No. 6, December 2012: 1603 – 1611