Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

OPEN SCIENCE: APPROACHES AND BENEFITS FOR MODELING & SIMULATION

Simon J. E. Taylor
Anastasia Anagnostou
Adedeji Fabiyi

Modelling & Simulation Group
Department of Computer Science
Brunel University London
Kingston Lane
Uxbridge, UB8 3PH, UNITED KINGDOM

Christine Currie
Thomas Monks

University of Southampton
Highfield
Southampton, SO17 1BJ, UNITED KINGDOM

Roberto Barbera

Department of Physics and Astronomy
University of Catania
Catania, ITALY

Bruce Becker

C.S.I.R. Meraka Institute
1, Meiring Naude Road
0001 Pretoria, SOUTH AFRICA

ABSTRACT

Open Science is the practice of making scientific research accessible to all. It promotes open access to the artefacts of research, the software, data, results and the scientific articles in which they appear, so that others can validate, use and collaborate. Open Science is also being mandated by many funding bodies. The concept of Open Science is new to many Modelling & Simulation (M&S) researchers. To introduce Open Science to our field, this paper unpacks Open Science to understand some of its approaches and benefits. Good practice in the reporting of simulation studies is discussed and the Strengthening the Reporting of Empirical Simulation Studies (STRESS) standardized checklist approach is presented. A case study shows how Digital Object Identifiers, Researcher Registries, Open Access Data Repositories and Scientific Gateways can support Open Science practices for M&S research. The article concludes with a set of guidelines for adopting Open Science for M&S.

1 INTRODUCTION

The principles of "Open Science" encapsulate practices that aim to make scientific research accessible to all, typically in some digital format. For most this involves open access publishing where a scientific article is made accessible through a journal or some institutional open access repository. However, Open Science principles go further and aim to make all the artefacts of scientific research openly accessible. This would make reuse and independent collaboration possible as others build on the outputs of research. Importantly, this would also address a perceived crisis in publishing where the results of many papers cannot be reproduced or validated (Baker 2016).

There is a world-wide revolution happening in Open Science, fueled partly by this reproducibility crisis and partly by demands made by funding bodies to make the outputs of publically-funded research openly accessible to society. How could Open Science benefit Modeling & Simulation (M&S)? To investigate this, this paper unpacks Open Science and presents approaches and benefits to M&S. The paper is structured as follows. In section 2 we present contemporary views on Open Science. Section 3 reviews several key issues in openness. Section 4 discusses good practices in the reporting of simulation

studies and introduces one approach, the Strengthening the Reporting of Empirical Simulation Studies (STRESS) standardized checklist. Section 5 presents a case study that shows how Open Science technologies including Digital Object Identifiers, Researcher Registries, Open Access Data Repositories and Scientific Gateways can support Open Science practices for M&S research. Section 6 presents suggestions of approaches for adopting Open Science in M&S. Section 7 summarizes the paper.

2 WHAT IS OPEN SCIENCE?

The FOSTER project (Facilitate Open Science Training for European Research) (www.fosteropenscience.eu) defines Open Science as "… the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods." Open Science therefore refers to efforts to make the output of research more widely accessible to scientific communities, business sectors, and society more generally (OECD 2015). The area consists of strategies that address a wide range of associated topics: open access, open research data, open research protocols and notebooks, open access to research materials, open source software, citizen science, open peer review and open collaboration. It is often facilitated by digital technology and open access repositories. Sometimes technologies such as Science Gateways and e-Infrastructures (cyberinfrastructures) are used.

There are many benefits that arise from openness in science and research (OECD 2015, p.18):

- Improving efficiency in science by reducing duplication and the costs of creating, transferring and reusing data; allowing more research from the same data; and multiplying opportunities for domestic and global participation in the research process.
- Increasing transparency and quality in the research validation process by allowing a greater extent of replication and validation of scientific results.
- Speeding the transfer of knowledge from research to innovation.
- Increasing knowledge spillovers to the economy and increasing awareness and conscious choices among consumers.
- Addressing global challenges more effectively by globally coordinated international actions.
- Promoting citizens' engagement in science and research – Open Science and open data initiatives may promote awareness and trust in science among citizens. In some cases, greater citizen engagement may lead to active participation in scientific experiments and data collection.

Funding agencies across the world are reflecting the need to be more open with respect to the outcomes of publically-funded research programs. For example the European Commission is promoting open (free of charge) access to scientific publications and research data as a core strategy for H2020-funded research projects (ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access). The National Science Foundation (NSF) has developed an outline framework for activities to increase public access to scientific publications and digital scientific data resulting from funded research (www.nsf.gov/news/special_reports/public_access). In the UK, the Concordat on Open Research Data (www.rcuk.ac.uk/media/news/120621) has been produced by HEFCE, Research Councils UK, Universities UK and the Welcome Trust to guide the development of Open Science. All emphasize the impact of Open Science on scientific progress.

Figure 1 shows an Open Science taxonomy produced by the FOSTER project. As it can be seen there is a wide range of "open" concepts ranging from access to scientific artefacts to the measurement of the impact. In this paper we focus on a subset of these that will be of initial interest to M&S researchers and practitioners (although arguably the whole taxonomy is applicable to M&S).

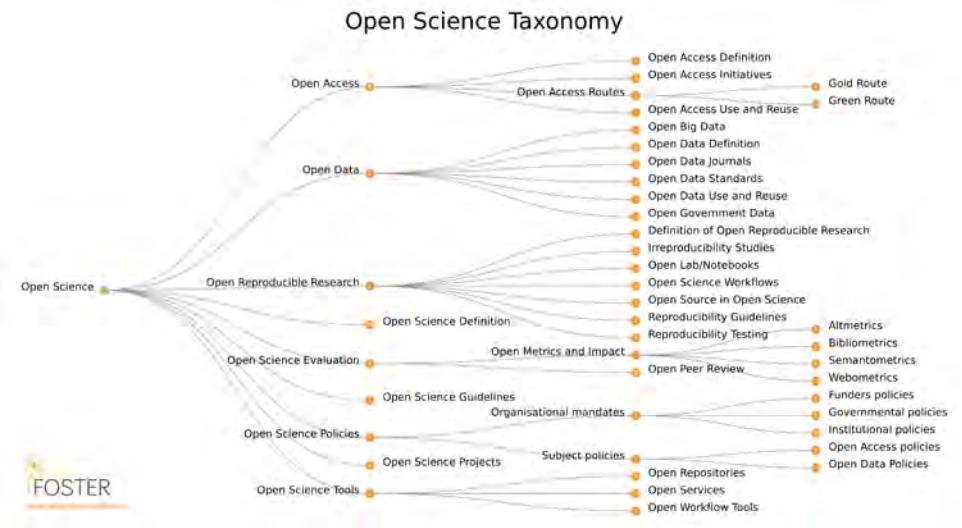

Figure 1: Open Science Taxonomy.

2.1 Open Access Publishing

This is perhaps the most well-known aspect of Open Science. The general concept of open access publishing is that research material (the publication, the data, etc.) is available free (and downloadable) without restriction. There are three different classifications of open access: gold, green and hybrid. Under *Gold* open access policies the publisher of a journal provides free open access to the articles of that journal. This is funded by the author (or institution) paying the journal an Article Publishing Charge (APC) (charges vary considerably). Some journals only charge for printed versions and offer free Gold open access online. Others do not charge. A hybrid model exists where a journal offers Gold open access for specific articles rather than the entire journal. Articles are normally licensed under a Creative Commons license (see later). *Green* open access involves an author self-archiving an article, typically in some Open Access Repository (OAR). This might be the final article, a post-print or a pre-print depending on the agreement with a publisher. The self-archived article might also be subject to an embargo period set by the journal publishers. Around 80% of publishers allow self-archiving (www.sherpa.ac.uk/romeo/index.php). Many funding agencies accept Green open access as an acceptable mechanism of openly publishing research results. A major issue in Green open access, however, is discoverability (i.e. when searching for the self-archived version of a paper). This situation is changing rapidly with new search tool plug-ins such as UnPayWall (paywall.org) that help to locate open access versions of articles.

2.2 Open Data and Reproducibility

Research requires data and can produce data. M&S is a key element of scientific discovery and consumes data and generates data (simulation results). Vast scientific projects such as The Large Hadron Collider and the new Square Kilometer Array radio telescope are extreme examples of research where huge amounts of data are generated (and more through simulation). In the latter it is expected that 1 petabyte of data will be collected every 20 seconds (EC 2010). Big data has emerged as a discipline to address the data needs of contemporary research. World spanning e-Infrastructures have been developed to process and analyze huge volumes of data and associated simulations. Many data issues are being discussed by new data communities focusing on, for example, international standards for data preservation and curation, common storage protocols and metadata, data integrity, access rights and the interoperability of

data sets and data infrastructures to store and process data (OECD 2015). All these present major challenges to Open Data policies in the pursuit of Open Science. For example, the OECD have produced *The OECD Quality Framework and Guidelines for OECD Statistics Activities* (OECD 2011). This is arguably relevant in terms of Open Data in M&S. These are:

1. *Relevance* – "is characterized by the degree to which the data serves to address the purposes for which they are sought by users."
2. *Accuracy* – is "the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure".
3. *Credibility* – "The credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer."
4. *Timeliness* – "reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon."
5. *Accessibility* – "reflects how readily the data can be located and accessed".
6. *Interpretability* – "reflects the ease with which the user may understand and properly use and analyze the data."
7. *Coherence* – "reflects the degree to which they are logically connected and mutually consistent."

We build on these general concepts within the context of a simulation study. Work also continues towards standards for data quality. As part of the 5-star Open Data Initiative (5stardata.info), Tim Berners-Lee has suggested a 5-star deployment scheme for Open Data. This is:

1-star: Make your data available on the Web under an open license (regardless of format);
2-star: Make your data available as structured data (e.g. in a recognized package such as Excel);
3-star: Make your data available in a non-proprietary open format (e.g. CSV);
4-star: Use URIs to specify the format and location of your data; and
5-star: Link your data to other data to provide context.

Other initiatives, such as the Open Data Institute (theodi.org) propose benchmarks for open data to assess how organizations and projects manage their data. They propose Common Assessment Methods for Open Data (CAF) based on context/environment, data use and impact and automated (certifiable) assessment. Open Data also poses complex challenges to national legal and regulatory frameworks.

Researchers embracing Open Data face additional challenges in terms of the time and effort required to follow Open Data guidelines to curate research data. One emerging framework is data citation and impact tracking. This suggests that data should be cited with an identifier and should be listed in the reference/bibliography to enable tracking and development of citation metrics (Kotarski et al. 2012).

2.3 Licensing

The idea of "openness" does not imply that openly published artefacts can be used without citing the owner. The idea of *Creative Commons* (CC) licensing models is meant to give clear limitations on the use, reuse and citation of an artefact. For example, an author might publish data that s/he is happy to use but would like an acknowledgement each time the data is used. A CC license would clearly identify this to others. CC licenses can be drafted by anyone. The Creative Commons non-profit organization (www.creativecommons.org) offers four license types:

- Attribution (by): "All CC licenses require that others who use your work in any way must give you credit the way you request, but not in a way that suggests you endorse them or their use. If

they want to use your work without giving you credit or for endorsement purposes, they must get your permission first."

- ShareAlike (sa): "You let others copy, distribute, display, perform, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first."
- NonCommercial (nc): "You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work for any purpose other than commercially unless they get your permission first."
- NoDerivatives (nd): You let others copy, distribute, display and perform only original copies of your work. If they want to modify your work, they must get your permission first.

For example, if a researcher wants to allow adaptations of his/her work to be shared, they expect those that do to share as well and they would be happy for others to commercially exploit their work, the researcher would choose Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). This license means that others can copy and redistribute the material in any medium or format, and to remix, transform and build upon the material for any purpose. Users must give appropriate credit (and indicate any changes) and distribute their work under the same license as the original. The CC website contains the formal license for all its licenses.

All licenses have extended legal text and are periodically updated. Licenses are therefore followed by the current version number (e.g. CC BY-SA 4.0). Attempts have been made to ensure that CC licenses reflect national copyright laws (e.g. some have rules on mediation and arbitration for dispute resolution). Creative Commons also support a tool (under "share your work") that helps to decide what license to use.

2.4 Uniqueness and Citation Tracking

There is also a growing movement that suggests that data (and all research artefacts) should be uniquely identifiable and should be listed in the reference/bibliography section of a paper to enable to track and develop citation metrics. Both DataCite (www.datacite.org) and ORCID (www.orcid.org) support this. DataCite has made possible the widespread use of Digital Object Identifiers (DOIs), an alphanumeric string assigned to uniquely identify an object. It is tied to a metadata description of the object as well as to a digital location, such as a URL, where all the details about the object are accessible. A DOI link in a paper is resolved into an open access repository that hosts the object, linking published articles and data sets (and other open resources) (i.e. a document stored in a repository has a uniquely discoverable web address). ORCID has created a researcher registry of persistent digital identifiers that uniquely identifies researchers (i.e. a researcher has a uniquely discoverable web address). Research workflows, such as manuscript and grant submission, support automated linkages between identified researchers and their professional activities. This ensures that researchers' works are correctly identified. Indeed many journals and grant submission systems require authors to have this. Researcher identifiers can be linked not only to scientific articles but also to other forms of research outputs, including equipment, experiments, patents and data sets. This clear tracking of work (DOIs) with researchers (ORCIDs) means that citation indices can be more easily maintained, especially in terms of scientometrics (e.g. h-index) and altmetrics (social media based impact tracking) (i.e. you get credit for your work through citations or "mentions"). To briefly demonstrate this, the ORCID of Taylor is http://orcid.org/0000-0001-8252-0189. One article published there is Taylor et al. (2016) that has the DOI http://dx.doi.org/10.1109/DS-RT.2016.35. The Green open access version of the article is at http://bura.brunel.ac.uk/handle/2438/14459. The data and software contained within the paper are accessible at DOIs presented later in this article as an example of Open Science.

3 A REPRODUCIBILITY CRISIS IN M&S?

The issues and concerns of reproducibility introduced above apply equally to M&S. For example:

- Rahmandad and Sterman (2012) sampled one year of articles from System Dynamics Review and found that out of 27 models 16 (59%) included no equations at all while 2 (7%) reported 'some' equations.
- Kurkowski, Camp, and Colagrosso (2005) reviewed 114 discrete-event simulation models of Mobile Ad Hoc Networks (MANETS) and found that 58% of the studies did not specify if a model was terminating or steady state; 0% of studies detailed the pseudo random number generator; 93% of studies did not include any comment on the need to deal with initialization bias and the 7% that did failed to provide any documentation about the analysis procedure used to select a warm-up period; finally 25% of studies did not state the simulation software in which the model was implemented.
- Grimm et al. (2006) focused on agent-based simulation models in ecology and drew similar conclusions that as the modelling becomes more complex, the potential and flexibility increases but the reproducibility decreases. This means that the results of agent-based simulation models are rarely reproducible. This was updated in 2010 (Grimm, et al. 2010) and extended for human decision making (Muller, et al. 2013).
- Janssen (2017) investigated the reproducibility of 2367 agent-based models returned from a search of ISI Web of Science. The study found that 50% of publications report complete or 'some' equations. Source code for the models was only available for 10% of the publications; there was a general lack of transparency in how models work.

Levent Yilmaz in Yilmaz et al. (2014) notes the critical role of reproducibility in M&S as well as automated provenance tracking, discoverability across the artifacts of M&S research and the appropriate use of CC licenses. Indeed he argues that reproducibility is key to credibility in research. The benefits of good reproducibility practice might also include:

- The advancement of operational knowledge (through reusing a published model to further investigate a system);
- To enable reuse of knowledge (models are expensive to develop; reusing models (or model components) can save time and money in M&S projects that could be devoted to a wider ranging study or analysis forms);
- To further conceptual modelling knowledge (a published model will argue how a conceptualization of a system has led to a given model, simulation, results and analysis; accurately reporting this conceptualization will help other researchers tackling similar problems in deciding what to model and what not to model;
- To reuse data where none exists (in many M&S projects data cannot be collected or is limited. In this case expert opinion is captured and modelled and/or missing data is approximated; capturing these assumptions in systematic manner will help to understand the validity of the study and help others to understand and build on the techniques used); and
- Testing of novel simulation methods (the validation of new analysis methods, algorithms, experimentation techniques require careful specification so that they can be assessed and reused elsewhere).

How might we address this reproducibility crisis? The ACM Transactions on Modeling and Computer Simulation (TOMACS), for example, has adopted a Replicating Computational Results (RCR) policy. If an author wishes to take part in this a RCR reviewer is assigned. S/he then works with the

author to replicate the results presented in the paper. Successful papers are given a seal of approval and a short report from the RCR reviewer is published with the paper. Open Science approaches could go further by allowing the model/simulation reported in a paper (and its data, results, etc.) to be easily accessed and used to reproduce and validate results described in that paper (i.e. through DOIs and ORCID). Initiatives such as CodeMeta (https://github.com/codemeta/codemeta) (Jones et al. 2016) are attempting to create a minimal metadata schema for science software and code expressed in JSON and XML. Combining these efforts with standardized checklists (see below) has the exciting possibility of automating reproducibility.

However, what information is actually needed for reproducibility? In the next section we present an approach to structuring this information through a standardized checklist and how this could be used as a basis for Open Science in M&S.

4 GUIDELINES FOR REPRODUCIBILITY IN M&S

The *Strengthening The Reporting of Empirical Simulation Studies* (STRESS) checklist attempts to provide authors with a framework to capture relevant details of a simulation study in such a way to enable others to validate and to reuse and extend the work of others. The STRESS guidelines were developed from (1) a literature review of good practice reporting approaches within ORMS, scientific model-based/empirical disciplines and software engineering; (2) M&S community engagement; and (3) expert review.

Table 1 shows the general STRESS checklist. These are split into five sections: objectives, model logic, data, experimentation and implementation. There are three specific instances of STRESS reflecting different M&S paradigms (agent-based simulation, discrete-event simulation and system dynamics): STRESS-ABS, STRESS-DES and STRESS-SD, respectively. Hybrid and/or distributed simulations can use these guidelines by combining STRESS guidelines to reflect the different paradigms used. Full STRESS definitions are accessible at (Monks 2017). We briefly discuss each section in turn.

4.1 Objectives

Objectives contain three items that define what the study aims to achieve. These are:

- *purpose* and rationale for the project and includes the model's intended use or experimental frame to aid other researchers and modelers understand the choices made in conceptualizing the model;
- *model outputs* that the model will predict; and
- *aims of experimentation*, specific information about how the model is being used to achieve the stated purpose.

4.2 Logic

Logic specifies model logic and logic used in scenarios (if applicable) described in terms of five items. Given the wide range of M&S approaches, STRESS recommends the use of a recognized diagramming approach that is meaningful to the community of practice in which simulation is applied as an aid to communicate model design. Within the main text authors should limit diagrams to conceptual or simplified overviews. Complex diagrams used to communicate complete model design should be included as supplementary appendix material. Components refer to the basic conceptual building blocks of the model and reflects the type of M&S paradigm: STRESS-DES focusses on entities, activities, resources and queues; STRESS-ABS focuses on the environment, agents, topology and interaction; and STRESS-SD focuses on stocks, flows and feedback loops.

4.3 Data

What data is used in the simulation? There are many different forms of data. For example, data sources (spreadsheets, databases, sensors, etc.), input parameters for base runs of the model and scenario experiments, derived distributions as well as associated data pre-processing and assumptions. Recommendations for reporting model data are common across the three modelling disciplines. There may be instances of modelling research where data are confidential or there are commercial reasons why data cannot be published. Ideally in these cases, descriptions should include hypothetical non-proprietary data so that researchers can still verify that a model has been reproduced accurately. Ethical considerations may also apply and should be mentioned with any sensitive data (especially with respect to health systems).

Table 1: General format of a STRESS checklist.

Section	Item No.	Checklist item
1. Objectives	1.1	Purpose of the model
	1.2	Model Outputs
	1.3	Experimentation Aims
2. Logic	2.1	Base model overview diagram
	2.2	Base model logic
	2.3	Scenario logic
	2.4	Algorithms
	2.5	Components
3. Data	3.1	Data sources
	3.2	Input parameters
	3.3	Pre-processing
	3.4	Assumptions
4. Experimentation	4.1	Initialization
	4.2	Run length
	4.3	Estimation approach
5. Implementation	5.1	Software or programming language
	5.2	Random sampling
	5.3	Model execution
	5.4	System specification

4.4 Experimentation

Experimentation deals with how the model was initialized, its run length and the output estimation approach used. In discrete-event simulation, initialization might capture warm-up periods, warm-up analysis procedures and procedures for setting initial conditions for queues and activities are reported. In system dynamics, the initial values of stocks might be considered. In agent-based simulation, the initial agent population size and attribute values and environment setup might be captured. Output estimation approach would depend on if the model was deterministic or stochastic. (e.g. the number of replications; use of variance reduction techniques such as common random numbers or antithetic variates, etc.). Results could be presented here (through a link) if not clear in the accompanying paper.

4.5 Implementation

This captures the implementation of the model/simulation. Software refers to the commercial or open source software, simulation or general purpose programming language or any other form of technology used to implement the model/simulation covered by the previous items (with the version numbers and any

additional information needed to install/execute the software). Random sampling details should be captured if the model is stochastic. The implementation of variance reduction techniques should also be considered. For example, in the case of common random numbers authors should describe how streams or seeds are distributed across components within the model. Model execution refers to how simulated time progresses within the model (which varies across the three approaches). Hardware and runtime information are important to capture the environment in which the (potentially distributed) model/simulation runs (especially if cloud, grid or high performance computing is used).

4.6 Example

For example, in modelling a simple queueing system such as a small shop, the purpose of the model may be to find the optimal number of servers to ensure good service; the model outputs might be average waiting time for service, the average utilization of the servers and the cost of the system; while the aims of the experimentation would be to provide details of the input parameters that can be changed such as the number of servers or the structure of the queues and the objectives. In this case, there may be more than one objective, with the experiment finding a good trade-off between customer satisfaction (i.e. time in the queue) and the cost of the system. Tables 2-4 show examples of how different elements of the checklist could be used.

Table 2: Example reporting for stochastic parameters.

Activity	Distribution	Distribution Parameters	Data source (sample size)
Service time a	Gamma	$\alpha = 4.5$; $\beta = 16.5$; min = 15	Observation (n = 125)
Service time b	Log Normal	$\mu = 7$ $\sigma = 4$	Blogs et al. 2004 (n = 2000)
Service time c	Triangular	Min = 3, Mode = 8, Max = 15	Expert Opinion (n = 3)

Table 3: Example reporting of experimentation setup.

- *The model had a run length of 180 weeks. Based on a MSER-5 analysis, a warm-up period of 60 weeks was used. No initial conditions were included. All point estimates are based on the average of 50 replications of a model run.*

- *The model had a run length of 180 weeks. No warm-up period was included; however, initial conditions for each queue were incorporated based on a discrete empirical distribution of queue length observations. Distributions for the initial conditions are reported in the online appendix. All point estimates are based on the average of 30 batched means (batch size is 6 weeks).*

- *The model had a run length of 30 days. The environment was initialized with a fixed size agent population (n = 10,000). All agents were in the potential adopter state initially and are connected to a random number of agents. All results are based on an average of 1000 replications.*

5 OPEN SCIENCE APPROACHES FOR M&S: A CASE STUDY

The previous section introduced an approach to capturing information to enable reproducibility in M&S. How can we build on this to create Open Science approaches for M&S?

What are the artefacts of M&S research? These might be the published research paper; the model/simulation program and the execution environment in which it runs; ancillary software and environments used to process data or present results; the experimentation schema, data, parameters, distributions used in experimentation and the results produced; associated documentation such as the

STRESS record for the study. An Open Science approach would suggest that all these artefacts would be available openly and in a discoverable form. The research paper would be available via some form of open access. Green open access would enable access to the pre-publication from an open access repository. Gold open access would mean that the research paper would be available directly from the journal or conference in which it is published. The other research artefacts would be published in an open access repository ideally associated with a DOI so that the artefacts are discoverable. Software (such as a model and/or a simulation) is a special case as there are several options. The code could be made available on a repository and linked to a DOI. The code could be published in its execution environment in a container running on a virtual machine, again linked to a DOI. Through Zenodo, for example, the code and all coding artefacts could be conveniently linked to a github deployment and linked to a DOI (https://guides.github.com/activities/citable-code/). Alternatively, to make accessibility as simple as possible, the code could be deployed on a virtualized infrastructure (e.g. a virtual machine running on a cloud or e-Infrastructure) and linked to a Science Gateway web-based front-end. Additionally, the documentation of the simulation study, in this case the STRESS record, could be stored in a repository and linked to a DOI (and the details required by STRESS referenced in turn via DOI links).

Table 4: Example reporting for implementation specifics.

• *The DES model reported was implemented in the commercial software Anylogic 7.5.3 Researcher edition and made use of its Process Modelling Library version x.2. The pseudo-random number generator was provided by the Java class Random version x.y. The model was run on a Microsoft Surface Pro 4, with a 2.2GHz Intel Core i7 processor and 16GB of memory under Windows 10 (build 14393). Model run time was 5 minutes per replication.*
• *The SD model reported was implemented in iThink 10.0.3. Integration method was set to Euler's Method. The model had a run length of 180 months with a DT of 1 month. The model was run on an Apple Macbook Air, with a 1.7GHz Intel Core i7 processor and 8GB of memory under OS X El Capitan version 10.11.16. Model run time was under 1 minute.*

To demonstrate Open Science for M&S, we present a short case study first published in Taylor et al. (2016). The case study has been extended to show how STRESS guidelines can be used to document the M&S study in an Open Science manner.

5.1 Case study

To demonstrate Open Science we have created an agent-based simulation in REPAST (repast.github.io) to study the spread of infection in a city after an outbreak. Agents can be infected, susceptible or recovered (Figure 2). When an infected agent approaches a susceptible agent, the latter becomes infected and if there are more than one susceptible agent in the cell, only one, randomly selected agent, is infected. Infected agents recover after a period and become recovered with a level of immunity. Recovered agents immunity decreases every time they are approached by an infected agent and when immunity becomes zero, the recovered agent becomes susceptible and can be infected again, thereby, forming a host of infection networks. The input parameters for the model include:

- simulation period (specifies how many years the simulation will run);
- recovered count (specifies the initial recovered population);
- infected count (specifies the initial infected population); and
- susceptible count (specifies the initial susceptible population).

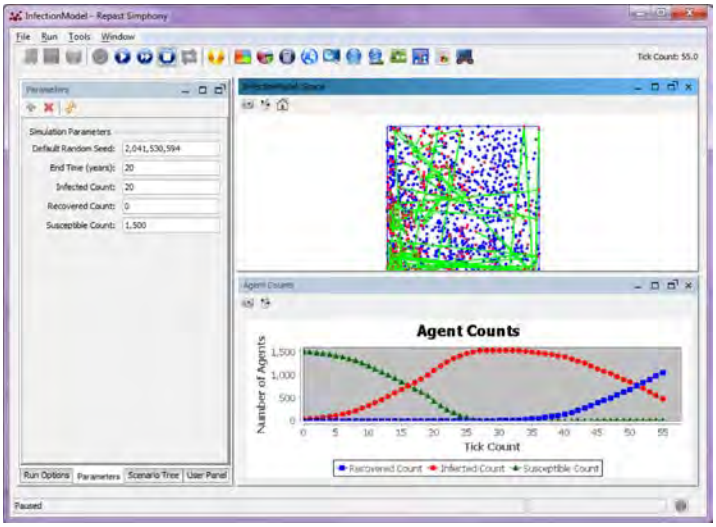

Figure 2: REPAST Infection model.

The outputs of simulation research can all be considered digital objects. To obtain a DOI for each of these, each output must be deposited in an open access repository hosted by a body that has permission to assign a DOI. As an object is deposited various metadata can be added that identify the authors, the URL, the sharing/use agreement, etc. It is also possible to create a DOI Collection that collects all the relevant DOIs together in a single reference. We ran five experiments to produce five sets of results. We also created a simple visualization tool that allows easy analysis of infected/non-infected population trends. We deposited all these research outputs in an Open Access Repository (oar.sci-gaia.eu). The following is the list of outputs and their DOIs .

REPAST Infection Model Example DOI Collection https://dx.doi.org/10.15169/sci-gaia:1457690398.43
REPAST Infection Model Virtual Appliance https://dx.doi.org/10.15169/sci-gaia:1455182324.71
Graphical Visualisation Tool for REPAST Infection Model https://dx.doi.org/10.15169/sci-gaia:1457432416.29
REPAST Infection Model Experiment 1 Results https://dx.doi.org/10.15169/sci-gaia:1457431676.23
REPAST Infection Model Experiment 2 Results https://dx.doi.org/10.15169/sci-gaia:1457431835.0
REPAST Infection Model Experiment 3 Results https://dx.doi.org/10.15169/sci-gaia:1457432005.33
REPAST Infection Model Experiment 4 Results https://dx.doi.org/10.15169/sci-gaia:1457432129.78
REPAST Infection Model Experiment 5 Results https://dx.doi.org/10.15169/sci-gaia:1457432242.73

The STRESS record has been deposited in oar.sci-gaia.eu at http://dx.doi.org/10.15169/sci-gaia:1494421530.94. This is shown below to illustrate how the above DOI links are used within the record. All these are associated with the researchers' ORCIDs (e.g. orcid.org/0000-0001-8252-0189).

To facilitate open use of the simulation we developed a virtual appliance (machine) version that runs on an e-Infrastructure accessed via a Science Gateway (in this case the FutureGateway from github.com/futuregateway) (Fabiyi et al. 2016). Many scientists do not have the IT expertise to install and run simulation software or have access to a simulation package. An alternative is to put the simulation online for people to use. Creating web-based simulations can be quite difficult to implement, especially if high performance computing is required to process a simulation quickly. Science Gateways have been developed to allow easy access and deployment of web-based software. This enables federated single-sign-on access to a range of resources (software, computers, data, sensors, etc.) To demonstrate this we have created the Africa Grid Science Gateway (AGSG) (http://sgw.africa-grid.org/) that hosts a range of applications developed for African scientific communities of practice. We have deployed the REPAST Infection Model on the AGSG. To access this users must first login via an Identify Federation. First time

users will be asked to register (use the catch-all GrIDP Identify Federation and IDPOPEN GARR Identify Provide if your own regional provider is not listed). After registration a user can access the Infection Model via the drop down application list. In this demonstration users can select any of the five experiments with parameters to run via a form (we restricted experiments due to this being a demonstration). Results can be accessed via the AGCG workspace and visualized by uploading the results file to the visualization tool (also accessible via the application list). Figure 3 shows the simple front end that runs the REPAST Infection model and Figure 4 shows the screen to download the results (My Jobs). The full implementation can be found at www.sci-gaia.eu along with many tutorials on Science Gateway development.

Figure 3: Science Gateway for the REPAST Infection Model.

Figure 4: My Jobs page on the Africa Grid Science Gateway.

6 OPEN SCIENCE GUIDELINES FOR M&S

There are many benefits of Open Science for M&S ranging from increased transparency and collaboration to the transfer of knowledge from research to innovation and the increased impact of research on society. Funding agencies are encouraging and mandating the adoption of Open Science. However, there is a balance to be made when intellectual property rights or confidentiality are at issue.

As introduced in this article, there are many approaches to Open Science that might be used for M&S. An initial set is summarized below.

- Publish openly using Gold or Green open access.
- Adopt good Open Data and Reusability practices that encourage independent verification and/or standardized reporting checklists such as STRESS.
- Consider making your data, results, software, etc. openly accessible (and trackable) by submitting your works to Open Access Repositories that support the use of Digital Object Identifiers (DOIs).
- Use Creative Commons licenses to specify how your work should be shared and used.
- Use a Researcher Registry such as ORCID to uniquely identify yourself and link this to your works via DOIs.
- Ensure that you use both DOIs and ORCIDs when publishing or in social media to correctly identify yourself and your works so that these can be tracked through scientometrics and altmetrics.
- Consider deploying your simulations via a Science Gateway or similar portal-based approach to enable the widest possible access to your work.

7 CONCLUSIONS

This paper has presented Open Science and an approach to using Open Science in M&S. It has introduced the STRESS guidelines to structure the information needed to support reproducibility in M&S. A short case study has shown how Open Science can be realized in M&S. A set of guidelines for Open Science in M&S has been presented.

We hope that this article will encourage debate on how Open Science can be widely adopted in M&S and how this can be recognized as a major area for the study of openness. For example, the FOSTER Taxonomy could be extended with the concept of Open Simulation that, while having some overlap with Open Data, has specific needs in its own right.

In conclusion, the adoption of Open Science in M&S can significantly benefit the discipline as a whole.

ACKNOWLEDGMENTS

This work is partially funded by the Sci-GaIA – Energising Scientific Endeavour through Science Gateways and e-Infrastructures in Africa project No. 654237 (H2020-INFRASUPP-2014-2: INFRASUPP-7-2014). See www.sci-gaia.eu for many materials on Open Science and supporting technologies such as the Open Science Platform.

REFERENCES

Baker, M. 2016. “1,500 Scientists Lift the Lid on Reproducibility”. *Nature* 533:452–454. dx.doi.org/10.1038/533452a.

EC 2010. “Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data”. Final report by the High-level Expert Group on Scientific Data, European Commission, October, ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204.

Fabiyi, A., S.J.E. Taylor, A. Anagnostou, M. Torrisi, and R. Barbera. 2016. "Investigating a Science Gateway for an Agent-Based Simulation Application Using REPAST". In *Proceedings of the 2016 IEEE Symposium on Distributed Simulation-Real Time Applications (DS-RT 2016)*, 29-36. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc., dx.doi.org/10.1109/DS-RT.2016.20.

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen, C. Jørgensen, W. M. Mooij, B. Müller, G. Pe'er, C. Piou, S. F. Railsback, A. M. Robbins, M. M. Robbins, E. Rossmanith, N. Rüger, E. Strand, S. Souissi, R. A. Stillman, R. Vabø, U. Visser, and D. L. DeAngelis. 2006. "A Standard Protocol for Describing Individual-Based and Agent-Based Models". *Ecological Modelling* 198(1–2): 115-126. DOI: dx.doi.org/10.1016/j.ecolmodel.2006.04.023.

Grimm, V., U. Berger, D.L. DeAngelis, J.G. Polhill, J. Giske, and S.F. Railsback. 2010. "The ODD Protocol: a Review and First Update". *Ecological Modelling* 221(23): 2760-2768. DOI: dx.doi.org/10.1016/j.ecolmodel.2010.08.019.

Janssen, M. A. 2017. "The Practice of Archiving Model Code of Agent-Based Models". *Journal of Artificial Societies and Social Simulation* 20(1):2. DOI: 10.18564/jasss.3317.

Jones, M.B., C. Boettiger, A. Cabunoc Mayes, A. Smith, P. Slaughter, K. Niemeyer, Y. Gil, M. Fenner, K. Nowak, M. Hahnel, L. Coy, A. Allen, M. Crosas, A. Sands, N. Chue Hong, P. Cruse, D. Katz, and C. Goble. 2016. CodeMeta: an Exchange Schema for Software Metadata. KNB Data Repository. DOI: dx.doi.org/10.5063/schema/codemeta-1.0

Kotarski, R., S. Reilly, S. Schrimpf, E. Smit and K. Walshe. Best Practices for Citability of Data and Evolving Roles in Scholarly Communication. Geneva: Opportunities for Data Exchange, 2012. core.ac.uk/download/files/324/30437756.pdf

Kurkowski, S., T. Camp, and M. Colagrosso. 2005. "MANET Simulation Studies: the Incredibles". *ACM SIGMOBILE Mobile Computing and Communications Review* 9(4):50-61. DOI: dx.doi.org/10.1145/1096166.1096174.

Monks, T, C. Currie, S.J.E. Taylor, S. Onggo, M, Kunc, and S. Robinson. 2017. Strengthening The Reporting of Empirical Simulation Studies (STRESS) Checklists. University of Southampton eprints. http://eprints.soton.ac.uk/id/eprint/407453.

Müller, B., F. Bohn, G. Dreßler, J. Groeneveld, C. Klassert, R. Martin, M. Schlüter, J. Schulze, H. Weise, and N. Schwarz. 2013. "Describing Human Decisions in Agent-Based Models – ODD + D, an Extension of the ODD Protocol." *Environmental Modelling & Software*. 48:37-48. DOI: dx.doi.org/10.1016/j.envsoft.2013.06.003.

OECD 2015. "Making Open Science a Reality". OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. DOI: dx.doi.org/10.1787/23074957.

OECD 2011. "Quality Framework and Guidelines for OECD Statistical Activities". 17 January, search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291.

Rahmandad, H., and J. D. Sterman. 2012. "Reporting Guidelines for Simulation-Based Research in Social Sciences". *System Dynamics Review* 28(4):396-411. DOI: dx.doi.org/10.1002/sdr.1481.

Taylor, S.J.E. A. Fabiyi, A. Anagnostou, R. Barbera, M. Torrisi, R. Ricceri, and B. Becker. 2016. "Demonstrating Open Science for Modeling & Simulation Research". In *Proceedings of the 2016 IEEE Symposium on Distributed Simulation-Real Time Applications (DS-RT 2016)*, 191-192. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. DOI: dx.doi.org/10.1109/DS-RT.2016.35

Yilmaz, L., S.J.E. Taylor, R. Fujimoto, and F. Darema, 2014. "Panel: The Future of Research in Modeling and Simulation". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S.Y. Diallo, I.O. Ryzhov, L. Yilmaz, S. Buckley, and J.A. Miller, 2797-2811. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. DOI: dx.doi.org/10.1109/WSC.2014.7020122.

AUTHOR BIOGRAPHIES

SIMON J. E. TAYLOR is the leader of the Modelling & Simulation Research Group in the Department of Computer Science, Brunel University London (https://tinyurl.com/ya5zjh8z). He leads major projects in industry and Africa. He a member of the ACM SIGSIM Steering Committee and founder of the Journal of Simulation. He has chaired several major conferences and his published over 150 articles. His email address is simon.taylor@brunel.ac.uk and his ORCID is orcid.org/0000-0001-8252-0189.

ANASTASIA ANAGNOSTOU is Research Fellow at the Department of Computer Science, Brunel University London and a member of the Modelling & Simulation Research Group. She holds a PhD in Hybrid Distributed Simulation, a MSc in Telemedicine and e-Health Systems and a BSc in Electronics Engineering. Her research interests are related to the application of modeling and simulation techniques in the Healthcare and Industry. Her email address is anastasia.anagnostou@brunel.ac.uk and her ORCID is orcid.org/0000-0003-3397-8307.

ADEDEJI FABIYI is PhD candidate in the Modelling & Simulation Group in the Department of Computer Science, Brunel University London. His research interests are in Cloud Computing, Network Security, Grid Computing, and IT Security. His email address is adedeji.fabiyi@brunel.ac.uk and his ORCID is orcid.org/0000-0002-7797-8272.

CHRISTINE CURRIE is Associate Professor of Operational Research in Mathematical Sciences at the University of Southampton, UK, where she also obtained her Ph.D. She is Editor-in-Chief for the Journal of Simulation. Her research interests include mathematical modelling of epidemics, Bayesian statistics, revenue management, variance reduction methods and optimization of simulation models. Her email address is christine.currie@soton.ac.uk and her ORDID is orcid.org/0000-0002-7016-3652.

THOMAS MONKS is funded by NIHR CLAHRC Wessex where he is Director of the Data Science Hub. He holds a BSc in Computer Science and Mathematics, MSc in Operational Research and PhD in Simulation Modelling. His research interest is applied simulation modelling and optimization in healthcare. His views do not necessarily reflect those of the NHS, NIHR, or Department of Health. His e-mail address is thomas.monks@soton.ac.uk and his ORCID is orcid.org/0000-0003-2631-4481

ROBERTO BARBERA is Professor of Experimental Physics of Fundamental Interactions at the Department of Physics and Astronomy of the Catania University, Italy. He has been involved in CERN experiments and he is one of the physicists involved in the ALICE Experiment at LHC. He oversees the design and the development of the Catania Science Gateway Framework and the Future Gateway. He is strongly involved in the establishment of Certificate Authorities, Identity Federations and Open Access Digital Repositories in various regions of the world, including Africa and the Middle East. His email address is roberto.barbera@ct.infn.it and his ORCID is orcid.org/0000-0001-5971-6415.

BRUCE BECKER is Senior Researcher at the Council for Scientific and Industrial Research, Meraka Institute, South Africa. He has worked on the ALICE experiment at the LHC and has kick-started the South African National Grid, a federation of institutes, national laboratories and research groups providing an integrated computational and data infrastructure. He works closely with SANReN in the area of identity federations, network-intensive applications and other advanced services. His email address is bbecker@csir.co.za and his ORCID is orcid.org/0000-0002-6607-7145.