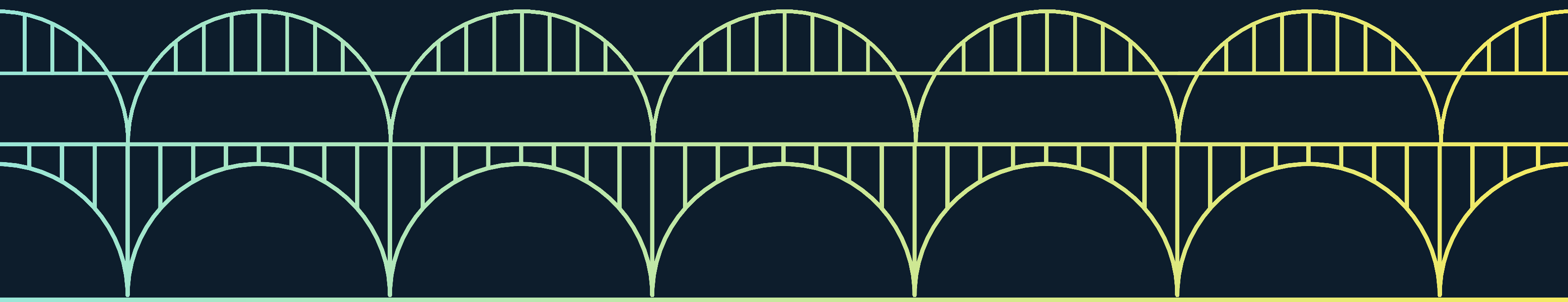
MDENET RESOURCES

The home of model-driven engineering

MDE for Open Science



Kirsty Wallis, Fiona Polack, and Steffen Zschaler

The MDENet "MDE for ..." series introduces ways in which MDE can be applied to address challenges in a wide range of different domains. Intended to help you identify how MDE might help you, each report identifies some key challenges in the domain, describes potential MDE applications through concrete case studies, and clarifies some of the key MDE concepts that will help address these challenges.

This document introduces you to some challenges in Open Science with pointers to how Model-Driven Engineering can help address these challenges. We begin with some case studies to give a concrete feel for what MDE might bring to the Open Science table.

1. CASE STUDIES

Citizen Scientists – access to local air-quality data

The Challenges

Air quality is a heavily researched subject with data being collected by research groups at many universities. Making this data available for collaborative research across different communities creates several challenges:

Challenge 1: Complexity

The problems addressed in air quality research are complex, with the need for monitoring data (which may represent many separate pollutants) to be analysed in the context of other data sources such as local weather readings and traffic monitoring.

Challenge 2: Accessibility

Accessing the various data requires significant technical skills and understanding – dealing with different data formats, sources, and query protocols. This makes the research all but inaccessible to people outside a focused community of experts.

Challenge 3: Non-expert demand

Air-quality data is of increasing interest to citizens, especially in large cities: people want to understand the air quality and how it is affected in their area, how changes to traffic patterns or proposed building work might affect them. In the terminology of open science, it would be useful to provide access to air-quality data to citizen scientists so they can use it to understand their own local environment.

The MDE Approach

Using MDE, we would design a domain-specific query language, which enables citizen scientists to express their information needs about air quality in a way that is close to their conceptual understanding of their environment, and where queries could be executed over the data sources available without the citizen scientist needing to understand the technical details of these data sources.

The Benefits

Discoverability

Making recommendations about potentially relevant data sources or filters to citizen scientists as they are entering a query.

Validation

Checking for semantic consistency (for example, using data-source meta-data to establish whether two data items can be meaningfully integrated).

Reuse and sharing

Enabling citizen scientists to share their queries (in an executable/editable form, not just results obtained) with other stakeholders, creating a community of data users.

Planning research data management for repeatability

The Challenges

Research data management is a key concern for open science. Data management plans are a commonly used tool to support planning and execution. Good research data management comes with several challenges:

Challenge 1: Structured Planning

Careful planning is required before a project starts. Writing good plans is difficult, especially as data management is a different competency to research skills.

Challenge 2: Systematic execution and audit

Once a project starts, it is important that the data management plan is faithfully executed and can be audited. This is difficult with plans written in text documents.

Challenge 3: Lack of reusability

Significant portions of data management plans are reusable across projects – enabling the potential for cross-project knowledge sharing and continuous improvement. To date, data management plans are typically written from scratch every time, using advice provided, for example, by the UK Data Service (for example, their Research Data Management resources).

The MDE Approach

Using MDE, data management plans would be considered as structured models of the data-management process written in a domain-specific modelling language for data-management planning.

The Benefits

Discoverability and learning

Researchers new to data management can be guided in what questions to consider but also what potential answers may be available, where there are stock answers to choose from, and where there is a need for more detailed and project-specific planning.

Structured planning

A data-management model is a structured artefact, which can drive the planning process. Linear text can then be generated automatically, where it is needed for review or documentation purposes.

Findability

Research-data management plans captured in a structured way can be linked through open data technology, making them easier to find, query, and analyse for other researchers and stakeholders.

Data documentation

A data-management plan captured as structured data can form the basis for a (semi-)automated generation of readme files etc documenting the final datasets published by a research project.

Execution support

Because the information is structured using meaningful data-management concepts, it will be possible to automatically generate technology support for data management, such as semi-automated workflows, data-management dashboards, etc. These could significantly simplify the day-to-day data-management activities ensuring data is managed in accordance with the original plan or that changes in the plan are connected to its implementation.

Open methods in computational science

The Challenges

Computational science develops computer models of complex scenarios and typically uses simulation to study emerging effects in these scenarios.

Computational science is faced with several challenges:

Challenge 1: Implicit translation of concepts

Models developed in computational science are usually implemented in standard, general-purpose programming languages such as C++ or Java. This requires scientific concepts to be translated into computing-oriented concepts. The link between the original domain concepts and their computational realisation is not explicitly represented in the final software code.

Challenge 2: Fitness-for-purpose argumentation

Code becomes a scientific instrument and we need to provide an argument of why the computational experiment is valid and robust. Such an argument needs to link back code-level concerns to concerns in the original scientific domain.

Challenge 3: Reuse and continual development

Simulations require significant effort in development. This effort can be recouped if code can be reused as new experiments are envisioned. However, this requires consistent and reliably maintained documentation as developers will likely change over time.

The MDE Approach

Using MDE, we would construct a modelling language that can directly capture the relevant scientific concepts. Model transformations and code generators would be used to explicitly define how these concepts are translated into software code.

The Benefits

This provides distinct benefits, contributing to the “open methods” agenda in open science by making the methodological aspects of the computational model explicit and transparently accessible:

Explicit translations

The way in which scientific concepts are implemented is explicitly documented in translations and these are always correct – as they are what is actually executed. Because they are explicitly described, they can be inspected, critiqued, refined, and experimented on.

Referenceable artefacts

The models and translations become referencable artefacts that can be tracked from publications, referred to in fitness-for-purpose arguments, and version managed to document changes over time.

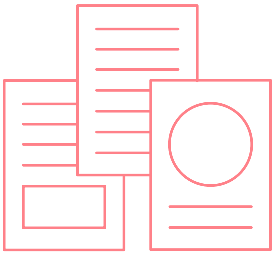
What key challenges are there in Open Science that could benefit from MDE

The definition of Open Science is broad and continually evolving. In general, Open Science can be considered to maximise the potential of research and education outputs by allowing the reuse and repurposing of outputs in novel and innovative ways. It also gives greater accessibility to and transparency of the research process, allowing replication and verification of findings. This opens the academic environment to members of the public, both by allowing access to academic research, but also in the form of citizen science projects which open up the academic process to wider participation.

Open science is a policy priority for the European Commission and the standard method of working under its research and innovation funding programmes. It improves the quality, efficiency and responsiveness of research.

Many countries, funders and institutions have policies on aspects of Open Science, such as Open Access and Open Data. The first truly international framework on open science was adopted by 193 countries attending UNESCO's General Conference in 2021. By making science more transparent and more accessible, the UNESCO Recommendation on Open Science aims to make science more equitable and inclusive worldwide.

There are many challenges to embracing Open Science, but in the context of this document we have chosen to focus on just the headline items where using MDE techniques can make a difference:



193

The international framework on open science was adopted by 193 countries in 2021.

Reproducibility & Transparency

Transparency ensures that all research findings on a topic can be accessed by researchers, policymakers and the public. This should provide a comprehensive picture of the current state of knowledge. Reproducibility requires that when independently repeating a study using the same methods and data, the same results are obtained. This is a stamp of credibility, supporting the trustworthiness of research findings.

FAIR principles - Findable, Accessible, Interoperable, Reusable

These principles aim to encourage researchers to share the outputs of their research, including their data, openly and in a way that can be easily found and responsibly reused.

Research Data Management (RDM)

RDM can be considered to cover all of the decisions and actions to manage data, from the research planning stage through to the long-term preservation and access to data. Data management is a requirement of many funding bodies, and is essential to ensure data quality, minimise risks relating to errors or misuse of data, and to be able to comply with legal, ethical, institutional and funder requirements.

Citizen Science/ Public engagement

Citizen science encapsulates a wide range of activities and practices that involves members of the public in research. Citizen science facilitates large-scale analysis, and may result in new insights. Public engagement has at its core the idea of two way dialogue connecting higher education research and teaching with the world outside.

The leading challenges of transparency, reproducibility, and FAIR are, or stem from, metadata. In order to make data available and efficiently curate or use data, there are key characteristics that need to be stored. These range from the data types or units of measurement, to how the data were collected, limitations placed on the data usage in collection/ curation agreements (e.g. data protection usage limitations), and information about the quality of data (e.g. sampling frame, treatment of missing or inaccurate data). An additional challenge in managing such metadata is that different stakeholders use different terminology and language to refer to similar and related concepts, creating a need for well defined translations between representations.

A further challenge of aspects such as reproducibility is to document and curate the programs and analysis methods applied to data in the original work; this challenge is also key to supporting citizen science in taking further or extending the scale of analysis. Traditional science publishes methods and equipment -- though these are not always sufficient to reproduce experiments. Data science needs to publish code, but also platforms and tools used, and needs to do this better (more systematically, more accessibly) than traditional science to support the spirit and practice of open science.

A key challenge in research-data management is about standardising procedures and platforms for curation, structures and languages used for sharing, and so on, especially across international boundaries, and ensuring that this is done within specific national or funder restrictions.

The key challenges in open science are centred around meta-data, documentation, and standardisation.

3. HOW CAN MDE HELP?

How can MDE concepts help address these challenges?

MDE contributions to open science focus on the facility to define languages, and to transform between languages.

1. Languages.

At the core of MDE is the idea of defining domain-specific modelling languages (DSML). A DSML is a formally defined (that is, computer-interpretable and computer-checkable) language that encapsulates the key concepts of a particular problem domain. This could be a language capturing the key concepts required for documenting research data in biology – for example, SBML is a language for describing knowledge about chemical pathways in biological systems.

Because DSMLs are formally defined, their definitions can be shared and agreed by a community. They can also be supported by tooling, such as editors, validations that can highlight inconsistencies and errors, or automated completions.

In the case studies, we have seen some examples of how DSMLs could support citizen science by providing language that makes science data accessible to non-scientist users, or research data management by capturing the concepts that need to be covered in a data-management plan.

2. Transformations.

A second key idea in MDE is that models expressed in a DSML can be automatically transformed into a representation in another. DSML or any other format. Such transformations are defined in transformation specifications—small programs that capture the relationship between concepts in the different DSMLs or formats, often expressed in dedicated transformation languages.

In the context of open science, transformations can address several challenges: For transparency, transformations can help translate between different terminology used by different communities. For reproducibility, or FAIR data, as described in the open methods case study, transformations can be a way of encoding the methodology, especially in computational science: rather than the translation from the scientists brain to the computational implementation remaining in a software developer's head, a transformation allows it to be expressed as an explicit artefact that can be used for reproduction as well as opening it to inspection and challenge by the scientific community.

4. NEXT STEPS

If you have found this interesting and would like to explore MDE further, there is help to hand.

Join us in MDENet (www.mde-network.org) to engage with an international network of experts in model-driven engineering, access learning resources, events, and funding.

ALSO SEE

Kirsty Wallis is Head of Research Liaison in Library, Culture, Collections and Open Science at UCL where she also leads the daily operation of the UCL Office for Open Science & Scholarship. orcid.org/0000-0002-9570-6174

Fiona A.C. Polack is Professor and Head of Department at the University of Hull, UK. Her research interests are in Model-Driven Engineering and the engineering of complex systems simulation, with specific contributions in relation to the expression and maintenance of fitness-for-purpose of scientific simulations, using argumentation approaches. orcid.org/0000-0001-7954-6433

Steffen Zschaler is a Reader in Software Engineering at King's College London and the director of MDENet, the expert network on model-driven engineering. His research interests include model-driven engineering, graph transformations, and principled simulation engineering. More information can be found at www.steffen-zschaler.de and he can be contacted at szschaler@acm.org. orcid.org/0000-0001-9062-6637

ABOUT MDE NET

The EPSRC network MDENet brings together research and practice in Model-Driven Engineering (MDE).
We will do this by:

Driving Future Research

We will establishing a clear understanding, shared by the community, of the current state of the art in research and the challenges at the forefront of academic research and industrial use. We also aim to identify and create opportunities for cross-disciplinary MDE research.

Training the IT Industry

We will curate and produce training material, lowering the barriers to entry for potential new users of MDE technology. These, in turn, will bring new demand and challenges for future research and development in MDE, creating demand and support for new and improved training materials.

Building the MDE Brand

Focusing on activities to increase the visibility of MDE research, and on community building. MDENet will be the home of the MDE community and the authority on MDE, opening this space to individuals and teams not previously involved in MDE

This network is a broad church. If you are interested in software development, automation, or (computational) modelling in other domains (biology, AI, robotics, finance, ...), this network will likely have something for you.

Join our community platform