Energy system modeling: Public transparency, scientific reproducibility, and open development

Robbie Morrison

Schillerstraße 85, 10627 Berlin, Germany

ARTICLE INFO

ABSTRACT

A mid-2017 survey shows that 28 energy system modeling projects have made public their source code, up from six in 2010, and none in 2000. Another six web-based energy sector database projects and nine hybrid projects were established during this same period, some explicitly to service open modeling.

Three distinct yet overlapping drivers can explain this shift in paradigm towards open methods: a desire for improved public transparency, the need for genuine scientific reproducibility, and a nascent experiment to see whether open source development methods can improve academic productivity and quality and perhaps also public trust.

The associated source code, datasets, and documentation need suitable open licenses to enable their use, modification, and republication. The choice for software is polarized: teams need only consider maximally permissive (ISC, MIT) or strongly protective (GPLv3) licenses. Selection is influenced by whether code adoption or freedom from capture is uppermost and by the implementation language, distribution architecture, and use of third-party components. Permissive data licenses (CC BY 4.0) are generally favored for datasets to facilitate their recombination and reuse. Official and semi-official energy sector data providers should also prefer permissive licensing for copyrightable material.

© 2018 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Calls to "open up" energy system models are growing, particularly for those models used to inform public policy development [1–10]. Simultaneously, a number of energy system projects are releasing their source code under open software licenses and starting to build user and developer communities. In parallel, several open energy sector database projects have been established to collect, curate, and republish the datasets needed by these models. This seismic change in practice is reviewed, together with the legal issues, mostly due to copyright, that enable and constrain these activities.

There are three distinct yet overlapping motivations for making energy system models open: improved *public transparency* as a reaction to sustained criticism over policy opaqueness, *scientific reproducibility* as a response to concerns over minimum scientific standards, and *open development* as an attempt to leverage the

benefits that open source software development methods can offer. These three motivations can be seen as a continuum, with public transparency as the least ambitious and open development the most.

While this article is aimed at energy policy models, much of what is discussed is likely to be applicable to other computational domains, such as the numerical modeling of urban air quality, economic systems, and climate protection strategies.

The legal examples provided reference either US or German law, primarily because these two jurisdictions are responsible for most of the litigation on open licensing and consequently most of the analysis.

Some recent appeals for greater openness in energy system modeling [1,3,6,7,10] remain silent on the issue of software licensing, presuming perhaps that source code can be lawfully used once published. This is correct *if and only if* open software licenses are provided. Otherwise, standard copyright prevails and precludes all usage beyond simple inspection. In contrast, datasets under standard copyright can *probably not* be legally machine processed, although the legal analysis on this matter remains extremely limited (discussed later).

The situation concerning the republication of source code and

datasets under standard copyright is clear. It is a breach of copyright to *publicly* distribute code and data if open licenses are absent. This means that encumbered code and data *cannot* be republished to support public transparency, scientific reproducibility, or open development. Hence, DeCarolis et al. [4:1849] erroneously conclude that "models with open source code and data but with no license are assumed to allow redistribution without any restrictions".

Indeed, only open licenses can unequivocally grant the right to study, use, improve, and distribute the associated code, data, and content — known as the four freedoms and first articulated for software by the Free Software Foundation (FSF) in February 1986 [11:121–122].

But open software licensing is as much a development model as it is a legal instrument [12:ii]. Open development implies that projects actively build communities by using code sharing platforms, social media channels, periodic workshops, and other forms of engagement. Open development should be seen as aspirational: it is not a necessary condition for public transparency or scientific reproducibility.

Open data has only really became an issue for energy system research with the advent of open modeling. Prior to that, closed source projects could purchase and use proprietary information under non-disclosure agreements (NDA). Or they could employ publicly available copyrighted data without attracting attention. In contrast, fundamental research domains like climate modeling have long shared unencumbered code and data. But energy system models need information from official and semi-official sources, including system and market operators, with much of it privately held. These operators and their umbrella organizations have, thus far in Europe at least, been reluctant to open license their public datasets or release key engineering information, leading to the current impasse and giving rise to crowdsourced projects to circumvent at least some of these limitations. It is presumed in this article that such information meets the legal threshold for copyrightability.

Selecting code and data licenses for an open energy system project can be daunting. This article discusses the issues involved and provides some guidance. The computer language used to implement a model can have a significant influence on the choice of software license, distinguished thus: compiled languages (C++, Java, C), interpreted languages (Python, R), and translated languages (MathProg, GAMS).[1] So too can the selected distribution architecture and the third-party libraries and source code that the project intends to either utilize or make available to other projects. These various considerations are strongly coupled.

This article proceeds thus. First, public transparency, scientific reproducibility, and open development are reviewed. A short audit of open energy system projects follows. Next, standard copyright and open licensing are examined. Attention then turns to the specifics of code and data in relation to open modeling, including the selection of suitable licenses.

2. Public transparency

Public transparency is a public policy ideal which requires, at the least, that the model in question be fully documented and that the datasets used be made available for inspection, but neither necessarily under open licenses. Some authors prefer to term the headline concept *comprehensibility* rather than transparency [3:2]. The qualifier *public* is used to exclude other less onerous forms of transparency, such as providing peer reviewers with secondary material.

Public transparency should help discourage what Geden [13:28] terms "policy-based evidence-making" in contrast to evidence-based policymaking. Both the model framework and its underlying design (expressed in code) and the selected scenarios (represented as data) embed limitations and assumptions that merit close scrutiny.

Acatech et al. [1:16–17] suggest that public transparency is best served with layered publishing for different audiences, ranging from short policymaker summaries to technical reports in sufficient detail to enable the results to be replicated. Cao et al. [3:4] consider (grammar corrected) "open source approaches to be an extreme case of transparency that do not automatically facilitate the comprehensibility of studies for policy advice". While that may be true, open development can also enhance transparency. Diligent open source projects produce clean code and good documentation, if only to service their own needs. Wiese et al. [10] argue that the public trust needed to underpin a rapid transition to zero carbon energy systems can only be built through the use of transparent open source energy models. Opaque policy models simply engender distrust. Strachan et al. [14:2] suggest that closed energy models that provide public policy support "fall far short of best practice in software development and are inconsistent with … publicly funded research". The Deep Decarbonization Pathways Project (DDPP) seeks to improve its modeling methodologies, a key motivation being "the intertwined goals of transparency, communicability and policy credibility" [15:27].

The oft heard call that models should publish their equations needs some examination. Mathematical programs, typically linear (LP) or mixed-integer (MILP) and written in an algebraic modeling language (MathProg, GAMS), can list their equations over some few pages because their codebase is essentially the programmatic expression of these equations [16:5835–5836]. But a sophisticated simulation/optimization framework (implemented in say C++ or Python) may need hundreds of pages to adequately record its workings. For instance, the core of deeco is documented in a 145 page PhD report [17] and a 239 page user manual [18], with later enhancements adding proportionately to this material. It is rare (in the author's experience) for software descriptions to be sufficiently complete and correct to enable reimplementation. Rather, the original developers must be contacted to fill in any number of absent details.

Allied to the notion of public transparency is that of market transparency [19:9]. Market transparency measures include the 2013 European electricity market transparency regulation 543/2013, intended to improve market liquidity and system security and also the standing of minor players [20]. The regulation requires that transmission system operators and wholesale market operators collectively gather, aggregate, and publish both electricity market and system reliability information. The machine use of this data, but not its republication, is permitted (discussed later). Moreover, the datasets thus provided need only be made available for five years and can then go dark.[2]

Paywalled research literature presents a significant barrier to public transparency, while unembargoed open access (OA) publishing promotes scientific reproducibility and open development. Free-of-charge provision of material under standard copyright and OA publishing under open content licenses are distinct concepts with differing objectives and attributes [21].

---

[1] MathProg is an open language from the GNU GLPK project that supports a subset of the proprietary AMPL language.

[2] Regulation 543/2013 [20] states (3 § 1): "The data shall be up to date, easily accessible, downloadable and available for at least five years. Data updates shall be time-stamped, archived and made available to the public."