Kia kaha! Stay strong!

Can non-open data standards legally taint conforming codebases and databanks?

Some thoughts on how proprietary intellectual property contained in semantic standards might transfer across to aligned software and conforming datasets through the inadvertent creation of derivative works

Free Software Foundation Europe (FSFE) Legal and Licensing Workshop (LLW) 2023
Gothenburg, Sweden ▪ 19–21 April 2023 ▪ Thursday 20 April 12:00–12:30 session

Robbie Morrison
robbie.morrison@posteo.de
Schillerstraße 85, 10627 Berlin, Germany

Release 03 ▪ 23 May 2023 ▪ Generated file: 2023-morrison-fsfe-llw-2023-taint-slidedeck.03.pdf ▪ Role: minor post-presentation improvements, CC0-1.0 and patents
Git information: c924eb0 ▪ 2023-05-23 15:40:17 +0200 ▪ ~/synk/fsfe-llw/gothenburg-2023/fsfe-llw-2023-presentation/slides

Abstract

Data standards are used to define and regularize data semantics, technical interchange, and, in some cases, legal status. Data standards are high-level specifications that can substantially influence both aligned software and conforming databanks.

This briefing concentrates on simulation software which attempts to represent selected aspects of the real word to some degree of fidelity. From a software perspective, data standards and API specifications overlap but API functionality may be ad-hoc whereas data standards aspire to be coherent and complete for their given contexts. From a database perspective, data semantics may be represented variously by data models, ontologies, vocabularies, reference architectures, and database schemas.

The selected domain for this briefing is energy systems modeling. These are computer simulations of future energy systems used to test and compare conjectured scenarios. As such, they are extremely data intensive and critically dependent on both data availability and the legal right to use and reuse that information.

Data standards, like all technical standards, may be subject to a range of statutory applications, intellectual property rights, and private and public licensing arrangements. They may contain patented material, are possibly protected as European 96/9/EC databases, and will certainly attract copyright. They may be offered under FRAND (fair, reasonable, and non-discriminatory) terms, under public but non-open licenses, and under Open (Knowledge) Definition-compliant licenses. Proprietary standards are normally subject to sales contracts and may have four-figure cover prices — with the issuing parties often quasi-public industry bodies yet also dependent on this revenue.

This briefing will look at how the intellectual property in such standards can transfer to both aligned software and conforming datasets. The material presented is speculative. But these are vital questions to resolve for system simulations that cover matters of public interest. Such analysis needs to be genuinely open, independently reproducible, and available for extension and experiment. Moreover, for many applications in the global south, including Africa, even modest fees are prohibitive.

The key legal issue is whether works that rely on data standards for both semantics and exchange structures might class as derivative works of those standards. Use of non-open standards may therefore legally compromise duly informed software or databanks issued under otherwise suitable open licenses. If the answer is potentially yes, then only strictly open standards should be used to inform this amalgam of code and data when applied to investigations of public interest.

This briefing will highlight several examples in the domain of energy systems analysis that have either required reimplementation, are of questionable open provenance, or are clearly problematic. And as usual in this context, legal risk aversion is a key consideration.

THIS MATERIAL IS SPECULATIVE

I also waive the Chatham House Rule and allow attribution

Preamble

Some background

Energy system modeling

- 1990 : began campaigning on global warming in a personal capacity
- 1995 : modeling national energy systems: multi-sector, high-resolution, contiguous time
- 2003 : added a GPL-2.0 license to *deeco* and attempted to build an online community
- 2016 : joined the Open Energy Modelling Initiative (openmod)
- 2017 : began advocating for genuinely open data of public interest with a focus on Europe

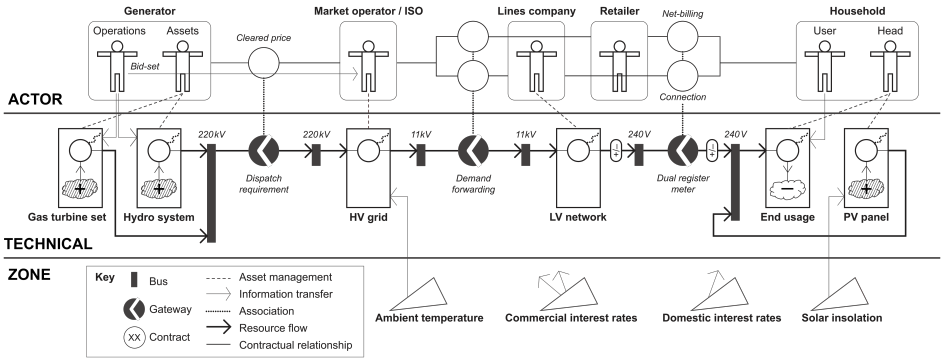
Legal experiences

- 1993 : arrested and detained but released without charge while hanging a climate protection banner on an electricity corporation headquarters
- 1995 : assisted pro-bono barrister Don Anderson in PlaNet Wellington Trust *v* Alan Marsden before a District Court (PlaNet was an early non-profit social media provider)
- 1998 : called forward to represent the community at an Environment Court hearing

Energy systems analysis

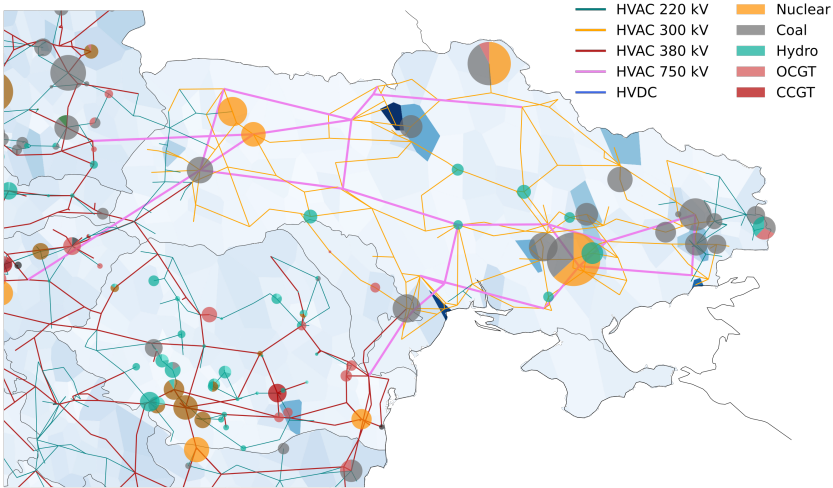
To provide some context for these discussions

A quick schematic ..



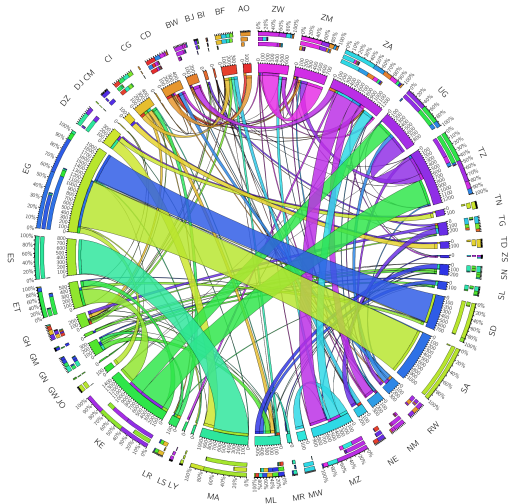
Hybrid agent-based/operational energy system optimization model

a topical example ..



Part of the open science PyPSA model of Europe being developed at TU Berlin (with Voronoi cells)

and some results



Cumulative electricity trade (2015–2065) among African countries for the reference scenario (TWh) (Pappis *et al* 2022:8)

Model-based energy systems analysis

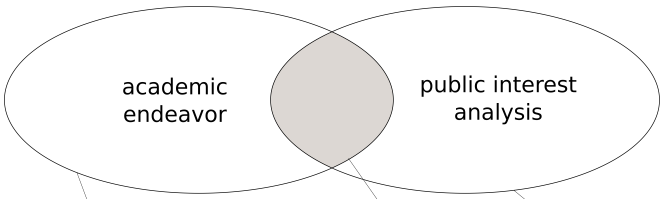
- open source is *arguably* becoming the norm for frameworks (see TIMES and US EIA efforts)
- open analysis more generally is *inching* along behind
- major efforts within the open modeling community to develop, adopt, and formalize:
  - **semantic standards** (see Open Energy Ontology)
  - **metadata standards** and **data cataloging systems** (perhaps based on DBpedia)
  - **technical standards** (see netCDF containers)
  - **agreed sets of scenarios** and **aligned interpretation methods** (see pyam)
  - **model comparison protocols** and **related assessments** (see EMF, MODEX, ECEMF)
  - **repeatable workflows** and **archiving practices** (currently problematic)
  - **open data licensing** to support extraction, reuse, and legal interoperability
  - **research data management** policies and infrastructures (see OPSD, OEP)
- uptake in the global south
- increasing overlap with IPCC-style integrated assessment modeling (IAM), demand simulation tooling, and water take and land usage (nexus) considerations

Model-based energy systems analysis

- open source is *arguably* becoming the norm for frameworks (see TIMES and US EIA efforts)
- open analysis more generally is *inching* along behind
- major efforts within the open modeling community to develop, adopt, and formalize:
  - **semantic standards** (see Open Energy Ontology)
  - **metadata standards** and **data cataloging systems** (perhaps based on DBpedia)
  - **technical standards** (see netCDF containers)
  - **agreed sets of scenarios** and **aligned interpretation methods** (see pyam)
  - **model comparison protocols** and **related assessments** (see EMF, MODEX, ECEMF)
  - **repeatable workflows** and **archiving practices** (currently problematic)
  - **open data licensing** to support extraction, reuse, and legal interoperability
  - **research data management** policies and infrastructures (see OPSD, OEP)
- uptake in the global south
- increasing overlap with IPCC-style integrated assessment modeling (IAM), demand simulation tooling, and water take and land usage (nexus) considerations

Open for what reason?

Public interest analysis and open science both need **legally unencumbered** frameworks and datasets but the motivations and objectives behind each endeavor are somewhat different



industry operations

industry coordination and smart systems software not included in this diagram

numerical reproducibility

- model outputs should align
- framework comparisons important

increasingly valued:

- open science ethos
- research data management (RDM) policy
- contribution and novelty in this context

numerical repeatability

- numbers should match
- workflow and archiving important

driven by:

- transparency
- engagement
- contestability

openmod foundation

The Open Energy Modelling Initiative (openmod) was committed to a fully open processing chain from its formation in September 2014

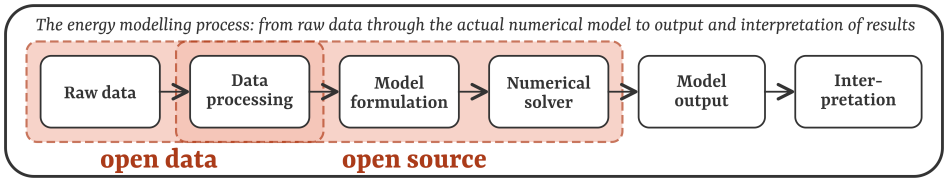
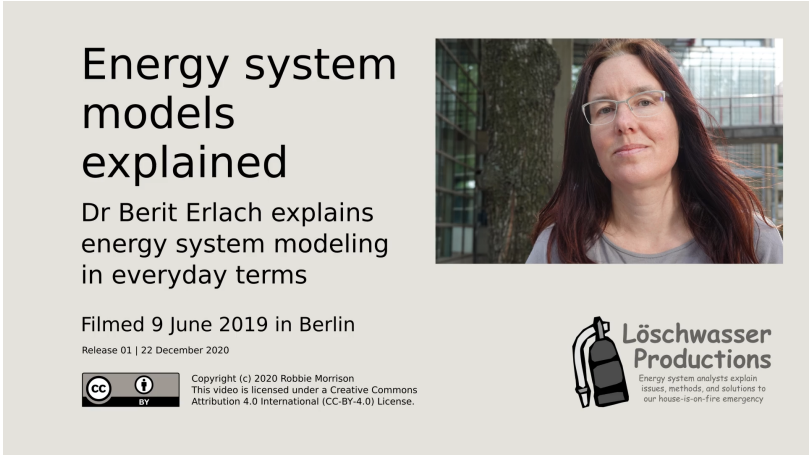


Diagram drawn by Stefen Pfenninger in 2014

Original source: https://openmod-initiative.org/img/open-model-process.png
Redrawn source: https://commons.wikimedia.org/wiki/File:2014-pfenninger-openmod-workflow-diagram.svg

A descriptive YouTube ▪ 13 minutes



Erlach, Berit (22 December 2020). *Energy system models explained: Dr Berit Erlach explains energy system modeling in everyday terms*. Berlin, Germany: Löschwasser Productions. Video 00:13:17. Filmed 9 June 2019 in Berlin, Germany by Robbie Morrison. Reference LP-001-01. CC-BY-4.0 license.

Semantic standards

Species nomenclature as a semantic standard

Common redstart ♀ near Berlin
Photograph: Robbie Morrison ▪ 10 October 2021

Common Redstart

Male

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Aves
Order:	Passeriformes
Family:	Muscicapidae
Genus:	*Phoenicurus*
Species:	***P. phoenicurus***

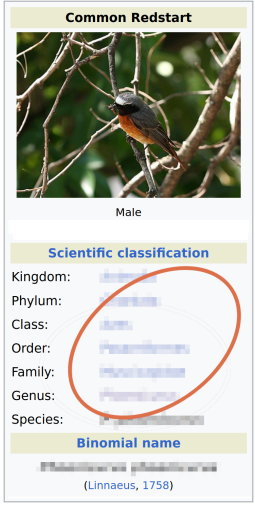
Binomial name

Phoenicurus phoenicurus
(Linnaeus, 1758)

Wikipedia article utilizing
Linnaean taxonomy

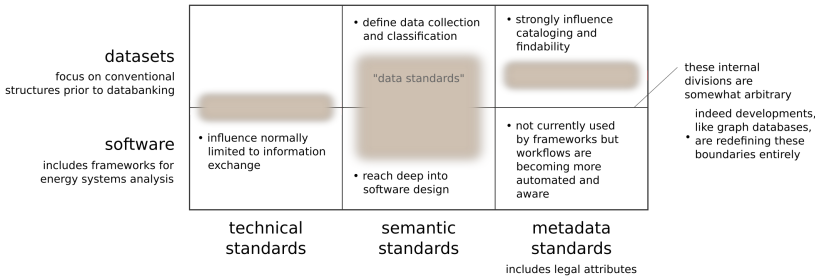
What if the underlying scheme was proprietary ..

Common redstart ♀ near Berlin
Photograph: Robbie Morrison ▪ 10 October 2021



Wikipedia article *without*
Linnaean taxonomy

Semantic standards



Semantic standards are more usually termed **data standards** but apply as much to software development as to data collection and classification

Proprietary semantic standards are widely used in engineering, including in the energy domain. Issuing bodies include the International Electrotechnical Commission (IEC). Terms are typically FRAND and cover prices can be USD $2000.

Some types of semantic standards

General:

- controlled vocabularies
- structured taxonomies (Linnaean taxonomy mentioned earlier)

Database-oriented:

- database schemas — usually specific to a particular database architecture
- entity-relationship models

Software-oriented:

- UML class diagrams

Mixed role:

- information exchange specifications
- reference architectures
- formal ontologies

Open Energy Ontology ▪ overview

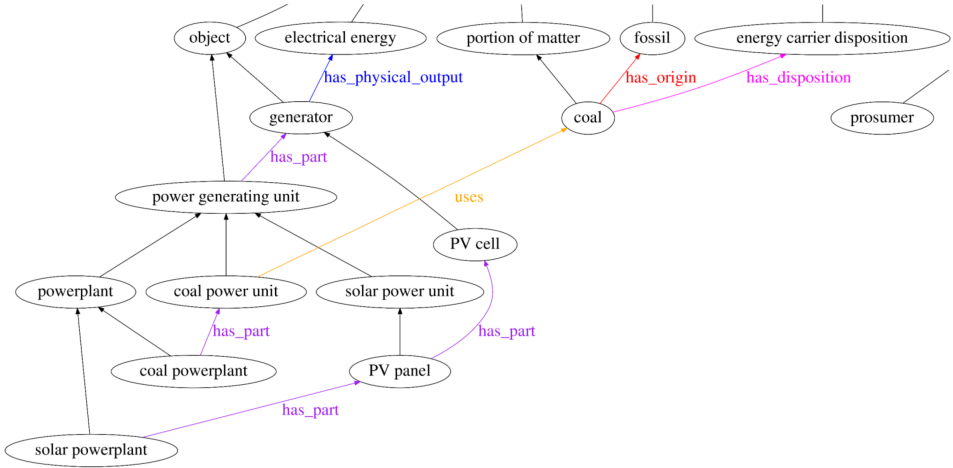
A formal ontology for the energy systems domain is being developed within open energy modeling community (Booshehri *et al* 2021)

Ontologies are concept driven, support knowledge graphs, and are important for AI technologies

- this exercise began as an informal glossary maintained on a wiki
- previous EnArgus ontology was legally encumbered
- upper structure is the Basic Formal Ontology (BFO) (ISO/IEC 21838-2)
- licensed CC0-1.0 — and one question is would dual-licensing with MIT would be useful?
- currently v1.13.0
- some stats:
  - monthly releases
  - 31 contributors
  - 1430 classes — expectation is 5–10k classes for domain coverage
  - 15180 axioms (including relationships)

- YouTube link (00:03:40): https://www.youtube.com/watch?v=rsxK3Y189ko&t=1184s

Open Energy Ontology ▪ relationships example



Open Energy Ontology ▪ class example

generating capacity

- **Name**: generating_capacity
- **Description**: Generation capacity for one unit of the technology
- **Type**: real
- **Is_about**:

  declared net capacity

  - **Name**: declared net capacity
  - **Path**: https://openenergy-platform.org/ontology/oeo/OEO_00230002

- **Unit**: MW

  - **Definition**:

    Declared net capacity is a power capacity stating the maximum power a power generating unit or a power plant can deliver to the electrical grid. It equals the sum of the rated powers of all plant generators minus all power used internally within the plant.

  - **Sub classes**: block size
  - **Super classes**: power capacity

Legal issues

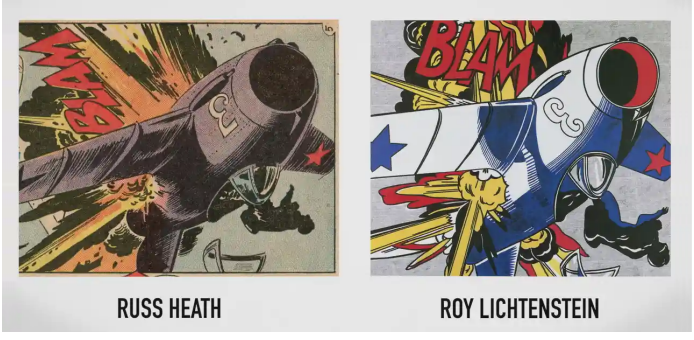
Most relate to derivative works under copyright law

EU legislation is presumed to apply in the first instance

Derivative works in other domains

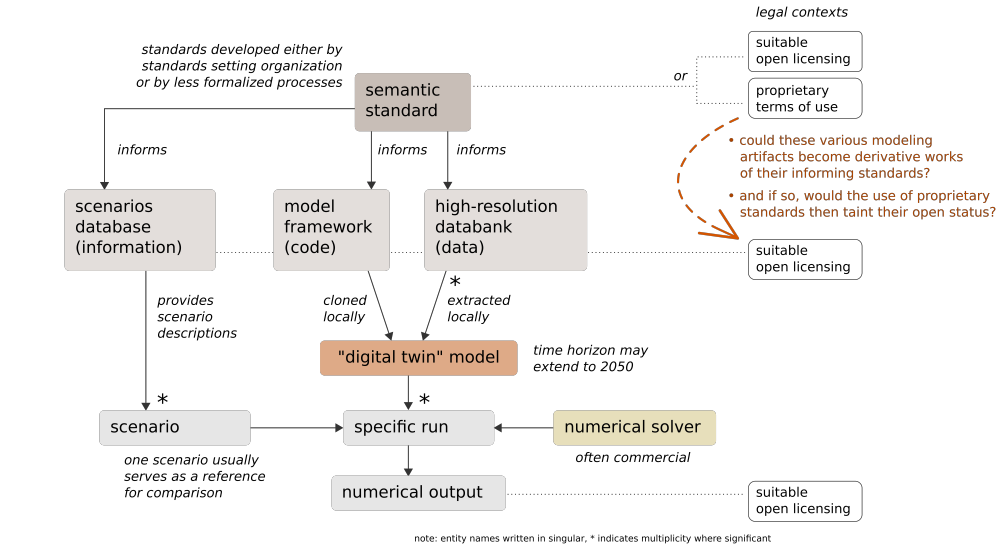
A still from a 2023 documentary on Roy Lichtenstein showing two images from 1962 (Alberge 2023). Russ Heath was a notable comic illustrator employed by a US publishing house and Roy Lichtenstein was a leading pop artist.

The use of **affirmative defenses** for copyright infringement as provided for in some jurisdictions, such as transformative use, are not under consideration here — rather **upfront legal certainty** is sought.



Source: Alberge, Dalya (9 April 2023). "It's called stealing': new allegations of plagiarism against Roy Lichtenstein". *The Guardian*. London, United Kingdom. ISSN 0029-7712.

The issue here



A chain of propositions (1/2)

1. the semantic standard — being the **underlying work** — would doubtless attract copyright (as discussed in slide 2/2)
2. any complying model framework or databank would doubtless qualify as a **derivative work** — albeit with added original contributions present (that rationale also covered in slide 2/2)
3. that **added originality** will likely attract copyright in its own right (more later on the question of separating works)
4. the alleged derivative work is clearly being offered for **distribution, reproduction, and use**
5. the **proprietary terms** — or even the **simple absence of a public license** — in the underlying work will therefore **collide** with the open licenses being applied to the complying framework (often MIT or Apache-2.0) or complying databank (typically CC-BY-4.0 for main data and CC0-1.0 for metadata)
6. any prospect of **separating underlying and additional contributions** in this context is effectively nil

More context (2/2)

1. a semantic standard certainly represents a **tangeable expression of concepts** for the purposes of copyright
2. the legal threshold is to establish that **major elements** of that underlying work remain
3. naming conventions, entity attributes, and structural relationships would all normally leave **clear signatures**
4. the derived work would be **impossible to create** in the absence of the underlying semantic standard
5. changes to the underlying semantic standard would **normally require** changes to the derived work in order to maintain a wider interoperability — that being the key point of the semantic standard
6. in addition, **patent encumberance** in the underlying work cannot be ruled out
7. similarly, **96/9/EC database protection** in Europe in the underlying work cannot be ruled out

Follow-up questions

1. as CC0-1.0 explicitly excludes patent grants (§4.a), dual licensing with MIT could be advisable?
2. can semantic standards attract 96/9/EC database protection as well?
3. not traversed here but where does machine learning sit relative to semantic standards?

Some conclusions

1 In general, most parties involved are **risk-averse** and litigation is to be avoided at all costs:

- affirmative defenses are of little interest
- statutory exemptions in Europe are bound to research and education and thus too limited

2 For **public-interest energy system analysis** — including digital twin-style modeling:

- overarching **semantic standards** should probably be **open licensed** for both content and code
- thus **CC0-1.0 OR MIT** might be best (using the SPDX disjunctive operator)
- conversely, **proprietary standards** should not be used, no matter how widely adopted they are

3 As a secondary issue, the **cover price** for proprietary semantic standards is often problematic, particularly for open source and open science projects and for projects operating in the global south

4 the **open development** of semantic standards should provide a much better model in any case

End matter

Readings

- Alberge, Dalya (9 April 2023). "'It's called stealing': new allegations of plagiarism against Roy Lichtenstein". *The Guardian*. London, United Kingdom. ISSN 0029-7712. Covers new documentary titled "WHAAM! BLAM! Roy Lichtenstein and the art of appropriation"
- Booshehri, Meisam, Lukas Emele, Simon Flügel, Hannah Förster, Johannes Frey, Ulrich Frey, Martin Glauer, Janna Hastings, Christian Hofmann, Carsten Hoyer-Klick, Ludwig Hülk, Anna Kleinau, Kevin Knosala, Leander Kotzur, Patrick Kuckertz, Till Mossakowski, Christoph Muschner, Fabian Neuhaus, Michaja Pehl, Martin Robinius, Vera Sehn, and Mirjam Stappel (1 September 2021). "Introducing the Open Energy Ontology: enhancing data interpretation and interfacing in energy systems analysis". *Energy and AI*. **5**: 100074. ISSN 2666-5468. doi:10.1016/j.egyai.2021.100074. CC-BY-4.0.
- Hirth, Lion, Ingmar Schlecht, and Jonathan Mühlenpfordt (6 November 2018). *Open data for electricity modeling: an assessment of input data for modeling the European electricity system regarding legal and technical usability — White paper*. Berlin, Germany: Neon Neue Energieökonomik. A report for the Federal Ministry for Economic Affairs and Energy, Germany. CC-BY-4.0.
- Marshall, Michael (17 June 2019). What is a species?. *New Scientist*. Web version.
- Morrison, Robbie (20 May 1990). *"Global Warming" pamphlet dated 20 May 1990*. London, United Kingdom. doi:10.5281/zenodo.4823717. With transcript. CC-BY-4.0.
- Morrison, Robbie (19 January 2022). *Open energy data and open data standards for improved public policy — Presentation*. doi:10.5281/zenodo.5879031. Berlin, Germany. Energy Systems Catapult webinar. CC-BY-4.0.

- Pappis, Ioannis, Vignesh Sridharan, Mark Howells, Hrvoje Medarac, Ioannis Kougias, Rocío González Sánchez, Abhishek Shivakumar, and Will Usher (March 2022). "The effects of climate change mitigation strategies on the energy system of Africa and its associated water footprint". *Environmental Research Letters*. **17** (4): 044048. ISSN 1748-9326. doi:10.1088/1748-9326/ac5ede. CC-BY-4.0.
- Parzen, Maximilian, Hazem Abdel-Khalek, Ekaterina Fedorova, Matin Mahmood, Martha Maria Frysztacki, Johannes Hampp, Lukas Franken, Leon Schumm, Fabian Neumann, Davide Poli, Aristides Kiprakis, and Davide Fioriti (13 September 2022). *PyPSA-Earth. A new global open energy system optimization model demonstrated in Africa — Preprint*. doi:10.48550/arXiv.2209.04663. Accepted for publication. CC-BY-NC-ND-4.0.

Abbreviations

- ECEMF : European Climate and Energy Modelling Forum
- EMF : energy modeling forum
- FRAND : fair, reasonable, and non-discriminatory
- IAM : integrated assessment model
- IEA : Energy Information Administration — a part of the US Department of Energy
- IEC : International Electrotechnical Commission
- MODEX : Model Experiments for the Energiewende — a cross-model comparison exercise
- OEO : Open Energy Ontology
- OEP : Open Energy Platform — a knowledge portal to support open energy system modeling
- openmod : Open Energy Modelling Initiative — an informal organization of modelers
- OPSD : Open Power System Data — a power systems data portal based in Berlin
- SPDX : Software Package Data Exchange
- UML : unified modeling language

Additional material

Problematic defaults for public interest data within Europe

The following normally apply within the European Economic Area:

- database protection attaches automatically under **directive 96/9/EC**
- **copyright for collections** attaches to datasets exhibiting nontrivial selection and arrangement
- **public sector information** is subject only to use (and thus not reuse) under **directive 2019/1024**
- material under **statutory reporting** is normally legally encumbered (due to above and more)
- the proposed **data producers right** (DPR) remains a legislative possibility under the **Data Act**
- scientific organizations with substantial EU funding continue to use **non-open bespoke licenses**
- the notion of **public interest data** has been partially displaced by the notion of "high value" data

The only realistic solution:

- to push for the use of **CC-BY-4.0 licensing** on main data and **CC0-1.0 waivers** on metadata

slidedeck ends