Toward open, reproducible and transparent seismic hazard and risk models
EFEHR vision and guidelines

Fabrice Cotton, Graeme Weatherill, Helen Crowley, Laurentiu Danciu, Celine Beauval, Olga Ktenidou, Päivi Mäntyniemi, António Correia

Warning. This draft is is a proposal which needs to be fairly challenged and debated before becoming EFEHR policy. It is important to raise objections if you think that the general philosophy and proposed license don't satisfy your needs, and to explain why this is the case. After this discussion there may be several options which could be presented during the General Assembly.

1. Principles and vision

The seismic hazard and risk models (and associated datasets or outputs) developed by the EFEHR community should follow the following principles.

- Reproducibility. Reproducibility of an experiment – or of a complex model – is one basic and unavoidable principle of modern science (Popper 2002). This is also one of the most challenging topics in the scientific debate because of the vast amount of information produced

- Transparency. Transparency ensures that all aspects of scientific methods and results are available for critique, compliment, or reuse. This not only meets a social imperative, it also allows others to test new questions with existing data, makes it easier to identify and correct errors, and helps unmask academic fraud.

- FAIR principles. The data and models used or produced by the EFEHR community should be Findable, Accessible, Interoperable and Resusable. Such FAIR principles (European commission, 2016) and an open circulation of models and corresponding results increases their acceptability (Pagani et al., 2015)

- Respect of the intellectual property and clear scientific ownership. A proper recognition (and citation) of data support or scientific contribution is indispensable. A clear scientific ownership and proper credit of technical or scientifc efforts foster faster and more efficient research progress, and provide the means to share data with future researchers.

For the purpose of disseminating, curating and replication by third parties of a European Seismic Hazard and Risk Model, all data sets used for the construction hazard and risks models should then be released without restriction of access.. The terms of use of the data should not impose restrictions prohibiting commercial usage but should ensure that further data sets derived directly from the data are made available under the same conditions as those pertaining to the original data.

It is our belief that the ESHM and ESRM should be intended for use by the widest possible audience. This includes the possibility to reproduce the construction of the model, to utilize some or all of its datasets for the purpose of constructing new models, and to utilize the outputs of the model to the broadest possible set of applications including commercial usage.

Who is the intended audience?
- i) Scientists and researchers from public, governmental and research institutions.
- ii) Scientists and researchers in privately funded educational and/or research institutions.
- iii) Engineering sector (including private consultants)
- iv) Insurance/Risk management sector (including private consultants)
- v) General public

2. EFEHR licensing guidelines

a. Definitions:

"EFEHR participating institutions" refers to institutions that are direct signatories of the EFEHR agreement. It excludes organisations who have volunteered data or expertise for the construction of the ESHM and ESRM but who have not signed the EFEHR agreement.

"third party" refers both to organisations (public and private) that are neither participants to EFEHR, as well as to organisations participating to EFEHR that have not been directly involved in the construction of the European Seismic Hazard and Risk Model

"output" refers to all hazard and risk results listed in the appendix below

"commercial usage": In this case would refer to *all* activities for which revenue can be generated by an organization or individual.

b. Proposed license: Creative Commons 4.0 CC:BY

The use of the Creative Commons 4.0 Share-Alike (CC:BY) would fulfill EFEHR principles, ensuring that the authors of each individual data set are given attribution and retain copyright. This does not prohibit commercial usage of the data but does require that derived data sets are released under the same Share-Alike conditions.

This license is compatible with the recommendations of the EPOS consortium (EPOS data policy, 2018). Indeed two licenses will be adopted by the EPOS consortium, CC:BY and CC:BY:NC.

The general rationale for adopting a creative commons share alike license is that no data compiler can withhold their data for the purpose of the EFEHR activities, nor seek to release their contribution publicly with different terms that might, for example, give that organization

advantages with respect to others involved in EFEHR. Restricting commercial access via the "Non-commercial (-NC)" addition to the standard CC licenses may prohibit usage by any for-profit organisations whose activities remain in the public interest (including private universities and engineering companies developing public works). This is likely not the intention of the compilers and may not conform with policies of other organisations in EFEHR. Permitting the data compiler the rights to grant permission of their data set for commercial use may leave EFEHR open to conflicts of interest, or de facto prohibition of usage of part or all of the model based on the different interpretations of commercial activity within each given country.

3. What should be made available?

In this case "data sets" refers to databases that are compiled by the EFEHR participating institutions for the purposes of the model construction. It does not include original "raw" data or intellectual contributions from third parties (i.e. information provided by local experts). "Data sets" may therefore include, for example, the compiled EMEC catalogue, the parametric historical catalogue, the final active fault database, the OpenQuake input files (source model, site model, GMMs, exposure and vulnerability input files).

It does not *necessarily* include the original "raw" data sets that have been used for construction of the main data sets. Excluded would be for example: macroseismic data points, contributing earthquake bulletins, contributing source model files from other projects, input data for the compilation of the active fault databases (e.g. geological maps, seismic reflection profiles etc.), input building or population census data, strong motion time-histories used in creating vulnerability functions etc. Where participating organisations wish to make such data available they may do so at their own discretion providing they themselves have the rights and permissions to do so.

With the ESHM20 and ESRM20, the boundary between original and derived data may be unclear. This includes: i) the ESM flatfile, which is currently released under a CC By-NC 4.0 license, ii) contributions to the EMEC catalogue, which means representations of some events are retained in the final EMEC catalogue without modification from their original source, iii) the contributing seismic source models that have been integrated into the "consensus" source model. Further discussion is needed on these conditions.

5. Collective Ownership

The proposed data policy intends that for those identified "data sets", no organization, participating to EFEHR or otherwise, is prohibited from access nor are they restricted in the use of the data for commercial purposes. As EFEHR currently stands, a joint commitment to the use of, for example, a CC By-SA 4.0 license would ensure that no individual contributor could withhold access or release their own data under different terms, whilst retaining copyright over their own data set. A joint agreement on such a license would also set a clear and unambiguous standard that can be communicated to outside parties wishing to contribute data for the development of the ESHM/ESRM.

However, each compiler or institution can grant additional rights to potential licensees on their own terms without necessarily seeking agreement from the rest of the EFEHR participants. To avoid this circumstance, it would be necessary that copyright is transferred to EFEHR itself. This assumes that EFEHR is a single entity that can be identified as a copyright holder, and would further require transfer of copyright from the compilers of the data sets (e.g. active faults, catalogues etc.). The ability of a data compiler to grant additional rights, be it for commercial or non-commercial purposes, would be lost, which may not be acceptable for all parties. It would require a necessary organizational structure that would require decisions to be taken at the governance level and ensure that the database holders are sufficiently represented at this level. Allowing the compiler of the data set a power of veto over management decisions to grant additional rights, or simply requiring unanimity for such a decision, may give the data compiler some re-assurance that decisions regarding usage cannot be taken without their consent.

6. Restrictions and disclaimer

It is unclear at this point to whom data access or usage may be restricted. CC By-SA would prohibit re-packaging and redistribution for profit by third party companies, but would not prohibit companies using the data for projects/activities for which there is commercial revenue. This would permit practicing engineers, including those in the nuclear sector, to use the data, models and results, so efforts are needed to ensure that liability for use (and misuse) rests with the user rather than the data compiler.

Following the example of the GEM foundation (which has been working with lawyers) the following disclaimer could be adopted by EFEHR:

"Legal statements. This map was created for dissemination purposes. The information included in this map must not be used for the design of earthquake-resistant structures or to support any important decision involving human life, capital and movable and immovable properties. The values of seismic hazard in this map do not constitute an alternative nor do they replace building actions defined in national building codes. Readers seeking this information should consult national databases. These models represent the best information publicly accessible, and the EFEHR consortium recognises their credibility and authoritativeness. This hazard map results from an integration process that is solely the responsibility of the EFEHR consortium"

7. Policy Regarding Publications

Release of data sets independent of scientific publications is generally compatible with the proposed approach. Where the compiler wishes to submit a scientific publication on the database to a journal that does not support SHERPA Romeo Green or Gold Access this should not be restricted per se. However, in such cases it is important that the data itself are made available independently of the journal, with a "sufficient" degree of independent documentation and its own DOI. The user can be encouraged to cite the journal publication.

8. Software Policy

To be determined. For the calculation, OpenQuake uses GNU Affero General Public License v 3.0, which generally conforms to open source policies and ensures that the source code of any modifications needed provide a service over a network must be made available. For other code used in the development of the model, if such software is necessary for complete reproduction of the model it should be encouraged that the code is released under an open source license.

References

Pagani, M., Weatherill, G., & Garcia, J. (2015, March). Seismic hazard models: a view on reproducibility, coherence and quality assurance. In Proceedings of the international workshop on ground motion prediction equation and seismic hazard assessment.

Popper, K. (2002). The Logic of Scientific Discovery. Rutledge classic, 545 pp.

European Commission. Guidelines on FAIR Data Management in Horizon 2020, July 2016. Available from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

EPOS data policy (2018). https://www.epos-ip.org/sites/default/files/repository/PDFFiles/EPOS%20DATA%20POLICY_July2018.pdf.

Appendix 1: Data Sets

Risk:
- Residential, commercial and industrial dwelling/building distribution data.
- Residential, commercial and industrial mapping schemes, dwellings per building, areas per dwelling or building.
- European Exposure Model Sources.
- European Exposure Model (input files for the OpenQuake Engine)
- Capacity curves for European vulnerability classes.
- Fragility functions for European vulnerability classes.
- Vulnerability Model (input files for OpenQuake Engine)
- Mapping table to map the building classes in the exposure model to the vulnerability classes in vulnerability model
- Risk calculation configuration files (i.e. the job.ini)

Appendix 2: Definition of Hazard and Risk "Outputs"

- Hazard curves and maps at all calculation target sites for a complete range of spectral periods defined by the modelers
- Uniform hazard spectra at all calculation target sites for a reasonable range of return periods
- Disaggregation results for selected sites, spectral periods and return periods
- Additional parameters derived from the hazard curves/UHS, for example Eurocode 8 short and long period coefficients, design spectrum corner periods etc.
- National loss exceedance curve (and national statistics such as AAL, PML)
- Event loss table
- Maps of AAL and PML (i.e. losses with a 200 year return period) at admin level 1

Excluded from this list would be: stochastic event sets (at the discretion of the modeler) and related ground motion fields, OpenQuake log files, intermediate calculation steps.

Appendix 3: Creative Commons 4.0 Definitions

More information on Creative Commons licenses can be found here: https://creativecommons.org/licenses/

A brief definition of each license term is provided below:
- **Attribution** (BY)— You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

- **ShareAlike** (SA) — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

- **NonCommercial** (NC) — You may not use the material for commercial purposes.
- **NoDerivatives** (ND)— If you remix, transform, or build upon the material, you may not distribute the modified material.